人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0113

ＡＩチャレンジ研究会 (第１３回)

Proceedings of the 13th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2001 年 6 月 15 日　　場　所　早稲田大学理工学部　55 号館 S 棟 2F 第 3 会議室

Waseda University, June 15, 2001

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-0113-1

ROBITA: グループ会話ロボット

ROBITA: Group Conversation Robot

松坂要佐 小林哲則

Yosuke Matsusaka and Tetsunori Kobayashi

早稲田大学 理工学部

School of Science and Engineering, Waseda University

{yosuke,koba}@tk.elec.waseda.ac.jp

Abstract

This paper describes a conversation system, which can participate and take part in group conversation. Group conversation includes many new problems, which are not cared in conventional one-to-one conversation: recognition of message exchange (recognize who is speaking and to whom he is speaking), expression of them to the users and strategy to take part in the conversation. We solved these problems by utilizing multi-modal interface: face recognition, face direction recognition, sound source estimation, speech recognition and gestural output using real body of the robot. The systematic combination of these functions realized a human-friendly group conversation system.

1 はじめに

複数人と共に話題を共有しながら行なう多対多の会話をグループ会話と呼ぶ。我々がこのようなグループでの情報交換を通じて、意思決定をしたり着想を得たりする機会は多く、グループ会話は人間同士のコミュニケーション形態として非常に重要性が高い。このため、グループ形態での人間の知的活動を支援する情報システムのように、人間同士のコミュニケーションに加わるシステムを実現しようとするならば、システムはグループ会話に参与できる能力を備える必要がある。

しかしながら、人間の行なうグループ会話に関する心理学的立場の研究や、人間のグループ会話を支援するシステムについての研究はいくつかなされているものの[2, 1]、グループ会話に参与する会話システムを実現する立場の研究は皆無であった。従来の対話システム研究は、人と機械が一対一で情報交換することを前提としたものがほとんどである[3, 4]。いくつかのシステムは複数のユーザを相手にすることを目標として設計されているが[6, 5]、それらにおいても個々の局面においては一人のユーザと対話をしているにすぎず、グループ会話の重要な用件となるユーザ間の対話に対し関与できない。人間同士の対話をモニタし、対話内容に応じて適切な情報を提供することを目的とするシステムも存在するが[7, 8]、これらのシステムは対話をモニタする立場に徹しており、システムが複数の人間と対等な立場で会話することを目指すものではない。

グループ会話に参与可能なシステムを実現するためには、従来の対話システムにおいては無視できたさまざまな問題を扱わなければならない。対話システムにおいては、個々の局面において各会話参加者がどのような役 (発話者、聴者など。2 に詳述。) で対話に臨んでいるかといった対話の状況をシステムが理解するとともに、次の局面でそれがどのように変化をするかを適切に予想しながら会話を進行できることが必要である。しかし、システムがただ一人のユーザを相手にする場合や、2 人のユーザの対話をモニタする立場に徹する場合などでは、これらは発話者を手がかりとして一意に決めることができた。これに対し、グループ会話では観察者 (2 参照) なる役が生じることで会話参加者間の関係は複雑化し、対話の状況は自明ではなくなる。円滑に会話を進行させるためには、システムは正確に対話の状況を理解できるとともに、自らも状況の曖昧性を減少させる適切な行動をとれることが必要となる。また、観察者である対話参加者が将来発話者として会話に参加する機会を得るためには、会話に対する協調的な態度を積極的に示すことで、現発話者の注意を喚起する必要がある。

これらの問題は会話システムにおいて本来重要な課題であるべきにも関わらず、従来それが問題として現れない特殊な問題設定において対話研究が進められてきたこと

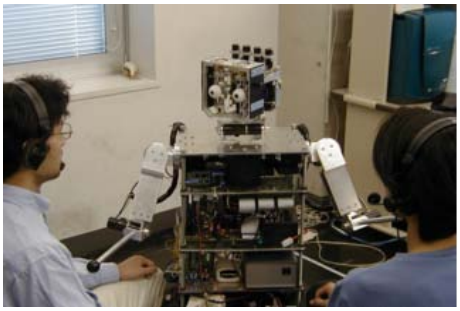


Figure 1: View of the group conversation.

によって、十分な検討がなされて来なかった。

そこで本研究では、グループ会話の特徴分析に基づいて、これに参与するシステムに求められる機能について検討を行ない、これを人間型のロボット ROBITA (Real-world Oriented BImodal Talking Agent) を用いて実現することを試みた。本稿では、分析・検討の結果およびそれに基づいたシステムの設計指針について述べるとともに、試作したシステムの概要とその動作例を紹介する。また、実ロボットを用いてグループ会話を実現することの意義についても検討を行なう。

2 用語の定義：グループ会話と会話参加者の役

本研究で扱うグループ会話とは、一つの話題を複数人で共有しながら発話を交換しあっている状態を指すものとする。このような会話では、通常会話参加者の誰か一人が他の誰かに注意を向けて発話を行い、それを他の会話参加者全員で聞くという形が繰り返されるのが一般的である。この形態で行なわれる会話に対し、本稿では以下の用語を定義する。

現在発話権を持って発話している人を発話者と呼ぶ。「発話権を持って発話する」とは、メッセージの送り手として発話することを指し、相槌などを挟んでいる人はこれに含まれない。発話者の発話にあたり注意を向けられた人を主たる聴者と呼ぶ。また、発話者と主たる聴者の対を当事者と呼ぶ。当事者間の発話の交換を対話と呼ぶ。当事者となることを前提として当事者の対話を周囲から観察している人を観察者あるいは従たる聴者と呼ぶ。当事者、観察者を合わせて会話参加者となる。発話者、主たる聴者、観察者、従たる聴者などを会話参加者に与えられた役と呼ぶ。

ある局面で誰がどの役を担っているかを会話の状況と呼ぶ。これらの用語を用いるとき、グループ会話とは、参加者の中で当事者が入れ替わりながら対話を繰り返す状況ということができる。図 2 に会話参加者の役の関係を模式的に示す。

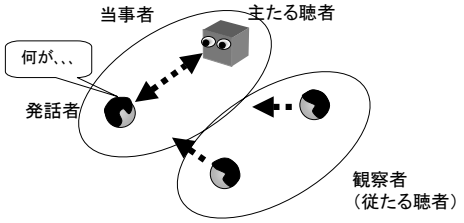


Figure 2: Part of the participant in the group conversation. Participants can divide into 2 groups, one is "parties concerned" and another is "observers". Parties concerned consist of two participants. One is a participant who has a turn ("focusee"). The other is the participant who message of the focusee is directed ("primary receiver").

3 グループ会話の特徴

グループ会話を特徴付ける最も重要な点は、従たる聴者 (あるいは観察者) の存在である。これが基となって以下のような特徴が生じる。

3.1 会話状況の曖昧性とその理解・明確化の必要性

一対一や一対多の対話では、観察者は存在しない。このため、自分自身の役は常に自明であり、対話の状況に曖昧性はなかった。これに対しグループ会話においては、観察者なる役が生じることによって、一つの局面でとりうる会話の状況に曖昧性が生じる。発話者でない場合は、自らが演じるべき役として、主たる聴者と観察者の二つの可能性が生じる。

役に応じて求められる役割が異なるため、会話参加者が役の判断を誤れば円滑な会話の進行は望めない。例えば、観察者として会話を静観すべきことを理解できていなければ、不適切なタイミングで発話を開始してしまうことになるし、主たる聴者として次の局面で発話することが望まれていることを理解できていなければ、会話に不自然な沈黙が生じる。このため、会話参加者は自らの役を適切に判断する能力が求められるとともに、他の会話参加者が戸惑わないよう、相手に期待している役を明確に伝えられることが望まれる。

3.2 発話の広報性と当事者間対話の理解に基づく発話の必要性

一対多の対話では、発話はその局面において何らかの方法で決められた唯一の聴者に向けて行なわれる。従たる聴者は存在せず、当事者以外は発話内容に気を留める必要がない。これに対し、グループ会話では発話者は主たる聴者を一人選定するものの、発話内容自体は従たる聴者を含む参加者全員に伝わることが発話の前提となる。このとき、従たる聴者に直接投げかけられた発話ではなく

ても、次の局面でその内容の理解を前提とした発話が求められることがある。このため会話に参加するためには、従たる聴者も当事者間の対話内容の全体を理解し、会話の流れを把握する必要がある。従たる聴者が会話の流れを把握した上で適切な情報を提供すべく積極的に発話権をとりに行く行動をとることができれば、グループ会話はより活性化されたものとなる。

3.3 アウェアネスの不平衡性とその改善努力の必要性

ここでは会話参加者が他会話参加者を対話相手として意識していること、あるいはその意識の程度を指してアウェアネスという言葉を使う。

グループ会話において、観察者に対する当事者のアウェアネスは、当事者間のアウェアネスに比べ低くなり、不平衡となる可能性を持つ。即ち、観察者の態度いかんによっては、当事者だけで対話が進み観察者が疎外される危険性を持つ。グループ会話においては、観察者の様子は当事者の視界に写るのであるが、観察者の行動が当事者に対し会話に対する興味を感じさせないものであるならば、コミュニケーションのきっかけをつかむ (話しかけられる) ことは難しい。会話参加者全員が対等な関係をもって、一体感のある会話を行なうためには、観察者は当事者の高いアウェアネスを喚起しなければならない。

4 システムの設計

前章では、グループ対話を実現するシステムが持つべき能力や期待される役割について述べた。本章ではそれらをどのように実現するかを述べる。

4.1 実ロボットの身体表現を用いた対話状況の明確化

前章に述べたように対話状況は曖昧なものであるから、参加者はその明確化を行なう必要がある。代表的には、発話者は主たる聴者として誰を選んだかを明確に伝えることが望まれるし、聴者は誰に発話者となってほしいかを明確に意思表示する必要がある。通常、人間同士で行なう対面対話においては、これらの意思伝達には各会話参加者の視線が重要な役割を果たしている[2, 12]。話者は相手を正視することによって、その相手に対して発話を行なっていることを表現する。聴者は発話者を見つめることによって、その相手から情報を期待していることを表現する。通常これらの対話状況を明確化するための意志伝達は無意識のうちに行なわれる。これを人同士が通常用いる方法と異なる手法により伝えることにすると、処理が意識に登り違和感を感じる可能性がある。そこで我々が構築するシステムにおいても、擬人化した顔・身体を持つロボットの身体表現によって、対話状況の明確化を行なうこととした。具体的には以下の方法による。

a. 視線による明確化

システムが発話者となる場合主たる聴者に視線を向け、聴者となる場合発話者を正視することとした。発話が途切れたならば、予想される次発話者に視線を向け、次発話者として期待していることを明確化することとした。

b. 体の向きによる明確化

発話者の場合には、体の向きにも意味を持たせた。視線が実際に発話を向けた人を表現するのに対し、体の向きは対話相手として意識している人を表すものとした。通常両者は一致するが、例えば従たる聴者からの割り込みを制止する場合などでは、体を主たる聴者に向けたまま顔だけを従たる聴者に向け「少し待ってください」と発話することになる。このことによって、発話相手としての意識は主たる聴者に残し後から何か伝えることを表現しながら、「待て」という指示を割込み者に伝えることができる。

4.2 マルチモーダルな情報処理による対話状況の理解

前節に述べたように、会話参加者は状況の明確化を主に身体表現により行なっているため、システムは画像処理によってそれを理解する必要がある。会話の問題は当然音情報処理の問題を含むため、グループ会話システムは必然的にマルチモーダルシステムとなる。

a. 発話者の推定

発話者の推定に関しては 2 つのモードを用意した。最初のモードでは、ロボットに設置したマイクを使う。まず音源定位 (5.2.2 参照) を行ないおおよその到来方向を求める。次に、画像処理によってその方向にいる人物を探し、結果を発話者として定める。

もう一つのモードでは、各対話参加者に個別に持たせたマイクを使う。この場合は、まず誰がどのマイクを使い、どの位置にいるかの対応づけを行なう。各マイクに入力された声を話者認識 (5.2.3 参照) し、各マイクと人の対応付けを行い、さらに全体を見渡しながら各話者の顔認識 (5.2.1 参照) を行うことで、人とその位置を対応付ける。このようにするとき、入力のあったマイクを持っている人を発話者と決めることができ、さらにどちらに視線を向けるべきかも決めることができる。

現状では、音声認識をロボットに搭載したマイクで行なうことは困難であるため、各参加者にマイクを持たさざるを得ない。このため、後者の方法が現実的といえるが、将来音声認識をロボットに搭載したマイクで行なうようになれば、前者の方法が重要となる。

b. 主たる聴者の推定

主たる聴者の推定については、発話者の視線を手がかりとする。視線の抽出は難しいので、実際にはこれを顔向きの認識で代用した。発話者の顔が向いている方の聴者を主たる聴者とし、それ以外を従たる聴者と判断した。発話者の視線は発話の終りに差し掛かったところで高い安定性を示す傾向がある[2, 15]。そこで、主たる聴者の判

断も発話の終了時の顔向き情報を重視した。

c. 次発話者の推定

前節 a に述べたように、従たる聴者は次話者を予測する必要がある。4.5 節 1) の原則からいえば、次発話者は現在の主たる聴者である可能性が高い。よって、発話者の顔向きから決めた現主たる聴者を次発話者とした。

4.3 当事者対話の理解に基づく従たる聴者の割込み発話

3.2 に述べたように、グループ会話では当事者でない場合にも発話を理解する必要があり、その理解内容を踏まえて発話することが求められる。しかしながら、ユーザはシステムとは設定タスク内の対話をするのに対し、ユーザ同士の間ではかなり広範な話題について対話する傾向があり、ユーザ間の対話を含む全発話の理解をすることは極端に高度な要求である。そこで本システムでは、極限られた文章に選択的に反応するようにした。具体的には、「このシステムは 4 歳です」「英語が話せます」など、システムを説明する表現だけをとりあげ、これを選択的に理解することとした。また、仮に事実と反する説明が行なわれた場合、それを訂正するための割り込みを行なうようにした。

4.4 アウェアネスの改善

3.3 に述べたように、各々の参加者の視界からは、今現在注目している相手だけではなく、そのほかの参加者たちの存在と行動が見て取れる。会話に参加できるきっかけを増やす為には、会話の状況に応じて適切な行動をとることで、他の参加者の自らに対するアウェアネスを向上させる必要がある。これらアウェアネスの喚起の問題は、自らの意志の表現という観点から 4.1 で述べた対話状況の明確化と密接に関わっている。本システムにおいては、発話交替毎に会話において主たる役割を負っている発話者を視線で追うと同時にうなずきをすることによって、会話に積極的に参加する姿勢を他の参加者たちに示すこととした。

4.5 システムの振舞い

Sacks らは、グループ会話において、次の原則が成り立つことを指摘している：1) もし現在の話し手が次の話し手を指定するならば、その選ばれた相手は次に話す権利と義務を持つ。2) もし現在の話し手が次の話し手を指定しないならば、最初に話し出した人が発話権を得、話者交代はそこで起きる。3) もし現在の話し手が次の話し手を指定せず、ほかの参加者が話さないならば、現在の話し手は話し続けることができる[16]。これらの原則を満たすよう、発話権の遷移を適切に制御する能力が必要となる。

以上を鑑み、システムには、担うべき役に応じて以下の振舞いをさせることとした。

a. 発話者としての振舞い：

a) 発話内容と主たる聴者を決め、主たる聴者に体と視線を向けて発話を行なう。

b) 発話途中に従たる聴者から割込みがあった場合、現在の当事者間の対話を中止できる状態であれば、割込み者に発話権を委譲する。

c) 中止できなければひとまず対話を中断し、割込みを制止した上で、元の対話に戻る。制止は、体を主たる聴者に向けたまま視線を割込み者に向け、少し待つよう発話することで行なう。

d) 発話の終了によって、発話権を主たる聴者に委譲する。

b. 主たる聴者としての振舞い：

e) 発話者を探し、発話者に視線を向ける。

f) 相槌とうなづきを交えながら発話を受理する。

g) 発話が終了したら、発話者となる。

c. 従たる聴者発話者としての振舞い：

h) 発話者を探し、発話者に視線を送る。

i) うなづきを交えながら発話を受理する。相槌はいれない。

j) 事実に反する説明があったら、対話に割込んでそれを訂正する発言をする。

k) 発話者が発話権を委譲したら、その時点における主たる聴者を次発話者と予想し、そちらに視線を向ける。発話者の発話権の委譲については、「ＸＸですか？」など発話権を委譲を表す典型的な文末表現を含む文が発話された時点、あるいは発話者の発話終了後一定時間誰も発話しなかった時点で、委譲があったものと判断する。

5 システムの概要

本章では、前章最終節に述べた振舞を実現するために用意した、ハードウェア、ソフトウェアモジュールの構成、および各ソフトウェアモジュールの内容について述べる。

5.1 ハードウェア

ロボットは人間を模した顔と身体を持ち、対話状況の明確化のための身体表現を行なう。二つの眼球にそれぞれ 2 自由度、首にピッチとロール軸の 2 自由度、腰に 1 自由度、両腕に各 4 自由度持ち、対話において必要とされる簡単な身体表現が可能となっている (4.1 参照)。ロボットは全方位に移動可能な台車の上に載っており自律移動が可能である。ロボットの眼球部には小型の CCD カメラ 2 台を搭載しており、画像を取り込むことができる。ロボットの顔両脇には 2 系統のマイクが設置されており音源定位に用いられる。ロボットの胴体部には、バッテリ、制御機器の他、制御用、音声合成用の 2 台の PC を積んでいる。

ロボット外部には、音声認識や画像認識用の PC、WS を置く。これら複数の計算機はネットワークを介して情報交換を行なう。

5.2 ソフトウェアモジュール構成

図 3 に本システムのソフトウェアモジュール構成を示す。ソフトウェアモジュールは、音声・画像処理を行なう認識系のモジュールと認識結果から身体表現や発話内容を決定する駆動系のモジュールに大別される。

音声処理としては、音声がどこから発せられたかを推定する音源定位モジュールと、各話者につき一つずつの音声認識モジュールを用意した。画像処理としては、 顔の位置を検出するモジュールと顔・顔方向を認識するモジュールを分けて用意した。 これは、視線の制御を高速な顔位置検出によるフィードバックループを形成することで話者とのスムーズなアイコンタクトを実現するためである。このループは、ほかの高度な情報処理 (顔・音声認識など) とは独立して動作する。各情報は情報統合モジュールによって統合された後、視線等ロボットの制御、応答文の生成・発話が行なわれる。

以下主な処理モジュールについて概要を述べる。

5.2.1 顔画像による個人と顔向きの認識

ロボット頭部に設置したカメラによって対話参加者の顔画像を取り込み、各対話参加者の個人認識（認識対象は研究室の学生 20 人）と顔向き識別 (識別対象は、正面、斜め前左右 30 度、60 度、90 度の 7 方向) を行なう。

統計的手法によって肌色尤度を求め、肌色領域を顔領域として抽出する。得られた肌色領域を 16 × 16 に正規化した後、独立成分分析にかけて、顔向き認識用には 7 次元, 個人認識用には 20 次元の特徴ベクトルに落とす。得られた特徴量を統計的な識別器にかけ、顔向きと個人とを認識する。実環境における顔領域抽出のロバスト性向上のために肌色尤度情報は、フレーム毎に MAP 推定に基づく適応処理が行われる[17]。

こうして得た顔向きの情報は各会話参加者が誰を注目しているかの判断に用いられ、個人情報は発話者の同定に用いられる (4.2 参照)。

5.2.2 音源の定位

頭部の両脇に設置した 2 つのマイクに入力される音の対数スペクトルの差を特徴ベクトルとし、統計的認識手法を用いて 10 度刻みの精度で音の到来方向を判定する。マイクに入力される音の周波数特性は、もともとの音源の周波数特性にロボットの頭部伝達特性の影響が畳み込まれた形となる。よって 2 系統の音の対数スペクトルの差をとれば、音源の特性によらず概ね左右のマイク位置での頭部伝達特性の差が得られることになる。この頭部伝達特性の差は音の到来方向に依存して決まるため、2 系統のマイクの対数スペクトル差を特徴ベクトルとして、統計的なパターン認識手法を適用することで音の到来方向を知ることができる[18]。音源の方向情報は、発話者の理解に用いられる (4.2.a. 参照)

5.2.3 音声の認識と音声応答の生成

音声認識では、バイグラム言語モデル、HMM 音響モデルを用いた語彙量約七百のフレーム同期の連続音声認識を行ない音声を単語の列へと変換する。また、GMM を用いた話者の認識 (対象は顔認識と同じ学生 20 人) を行う[19]。それに続く言語処理では、認識の結果得られた単語列から、キーワード系列をスポッティングし、さらにテンプレート処理によりキーワード列を意味表現へと変換する。各意味表現と、それが入力されたとき行なうべき音声応答と動作の関係は、表形式で与えられており、認識理解処理結果に応じて決められた音声応答と動作を行なう。応答音声の合成に関しては Windows ベースのテキスト読み上げソフト[20]を用いている。

6 システムの動作

上記のシステムを統合して、グループ会話を実現した。会話参加者数は 2 名 (ロボットを含めて 3 名) である。会話内容としては、ロボットの持つ機能についての一問一答形式の質疑応答とした。各ユーザはロボットの前にならんで座り、任意の会話参加者を相手に発話を行なう。システムは、自由背景、室内照明の条件の下で動作する[21]。

会話例を図 4 に示す。対話実験を通じてシステムの各機能は安定に動作することが確認できた。ロボットは、発話者の顔方向を認識することで、発話が自分に向けられているのか、または他のユーザに向けられているのかを判断し、望まれたタイミングで適切な応答を返すことができた。会話参加者は、身振りや呼びかけを交えながらロボットに語りかけ、また、ロボット自身も身振りで内部の状態を表出することによって、ごく自然な対話をすることができた。さらには、ロボットが観察者の役を担うときも、発話者を予測しながら交互に視線で追うことによって、会話に対する参加意識を表現することができ、対話の当事者たちもロボットを協力的な対話参加者として意識することができた。

7 検討

前章に紹介したように、ロボットは、対等に近い関係で人間同士の会話に加わることができ、参加者は、互いに一体感のあるグループ会話を実現できた。

本システムの特徴的な点として、1) 人間型のロボットを用い、擬人化インタフェースを実現したこと、2) それを 3 次元空間上に構成したこと、3) 聴者のシステムにアウェアネス改善のための行動をとらせたこと、などが挙げられる。

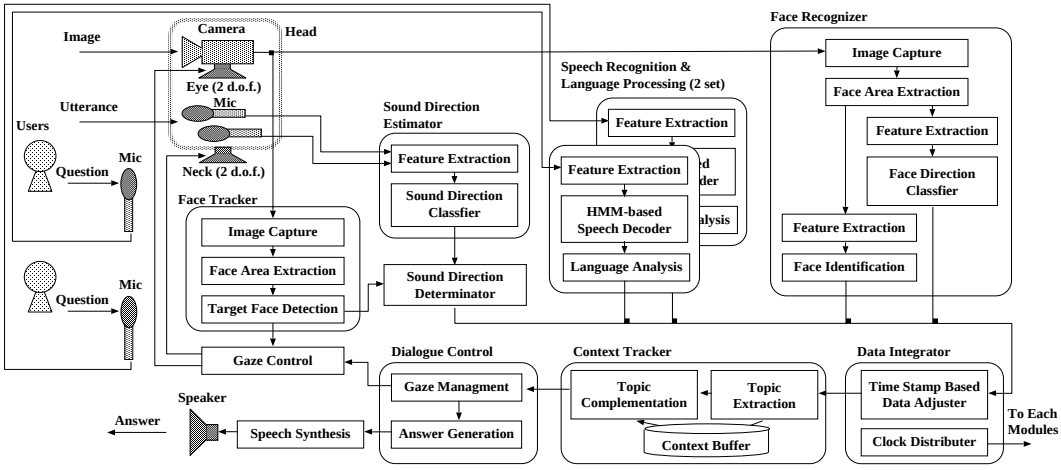


Figure 3: Software module composition of the system.

一般にインタフェースは透過的でなければならない[22, 23]。この意味で会話参加者は対話の状況を明確化するなんらかの努力を求められることを既に述べた。また、対話参加者が観察者の役を担うとき、自分に対するアウェアネスを高めるための努力をすることの必要性についても述べた。これらの目的のためにシステムがとる手法が人間が通常行なう手法と同じであれば、違和感なくシステムの意志が伝わることが期待でき好ましいことではあるが、そうしなければならない必然的な理由はない。例えば、対話が人と機械の一対一の関係で行なわれるのであれば、人間が機械の指定するやり方に合わせることが可能であり、人間はそれに慣れることができる。しかしながら、ここで扱ったグループ会話においては、人間と人間との対話に機械が加わる形で会話が構成される。状況の明確化やアウェアネスの改善のために人間が人間に対しとる手法が、加わった機械の都合で（人間同士にとっては不都合な）機械に合わせた手法になることは考え難い。また、会話全体としての一体感を実現するためには、局面毎にとられる手法に一貫性を持たせる必要がある。これらの意味で、グループ会話においては、擬人化インタフェースを用い、人間と同様な手法で状況の明確化を行なう能力を持たせることは非常に重要と考える。

また、この擬人化インタフェースを 3 次元の実空間上に実現したことの意味も大きいと考える。擬人化インタフェースとしては、2 次元ディスプレー上の擬人化エージェントを用いた研究が盛んに行なわれている[24]。しかしながら、グループ会話では様々な場所に位置どったユーザ個々に対して、その位置に依存した情報を正確に伝えることが重要であり、平面的な画像情報によって、これを伝えることは困難である。例えば、視線ひとつとってみても、2 次元の情報で特定のユーザだけを見ている状態を全ての位置にいるユーザに対して正確に伝えることは非常に難しい。グループ会話にとって、実ロボットを使ったことは、この観点からも有効であった。

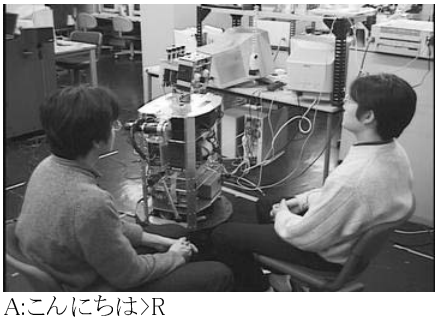
会話システムにアウェアネスの改善を意図した行動をとらせたことも他に例を見ない特徴である。会話実験の被験者によれば、被験者はロボットを協力的な対話参加者として意識することができ、共に会話をしているという感覚を持つことができたとしているが、これは、システムが観察者の役に回っているとき、アウェアネス改善の努力をしたことの効果によるものと考える。

8 むすび

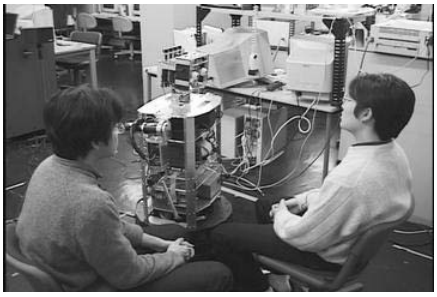
グループ会話の特徴の分析に基づいて、円滑なグループ会話を実現するために必要なシステムの振舞いについて検討し、さらに、マルチモーダルな情報処理機能を持つ人間型ロボットを用いて、グループ会話に参与できるシステムを実現した。

ロボットは会話状況に関する判断能力を持つことで、望まれるタイミングで自然な応答をすることができた。また、ロボットが会話状態の明確化のために行なう身体表現の効果により、会話参加者も戸惑うことなくグループ会話を進めることができた。ロボットには、単に聞かれたことに答えるだけではなく、ユーザ同士の対話に割込んで発話を行なう機能も持たせた。また、ロボットが観察者の役に回っているときでも、アウェアネス改善に向けて対話当事者に働きかける機能を持たせた。これらのことにより、グループとしての会話を活性化することができた。

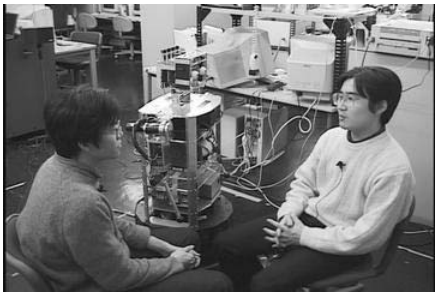
ここで実現したシステムでは、会話において人がとる振舞いをかなり単純化して捉えており、また扱った対話タスクも小規模なものではある。しかし、グループ会話の本質を捉えたシステム構成になっているものと考える。今後このシステムを基礎として、より自然で有用なグループ会話システムを実現していきたい。



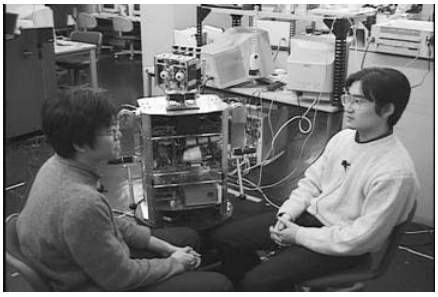
A:こんにちは>R
R:こんにちは>A



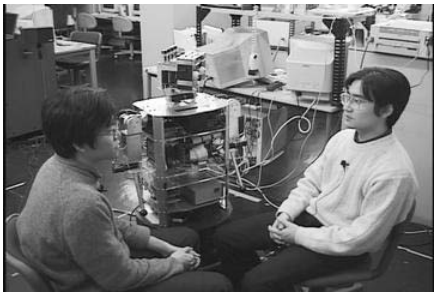
A:何ができますか>R
R:私は複数の話者と対話ができます>A



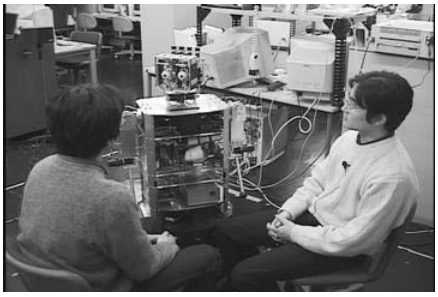
A:こんな感じでしゃべることが
できます>B



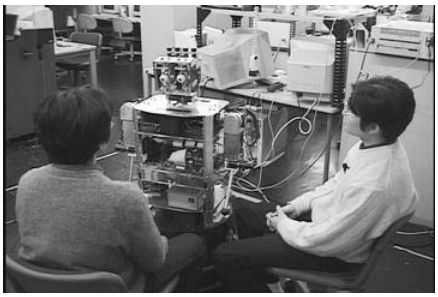
R:(interest)>B
B:何を聞いても大丈夫ですか>A



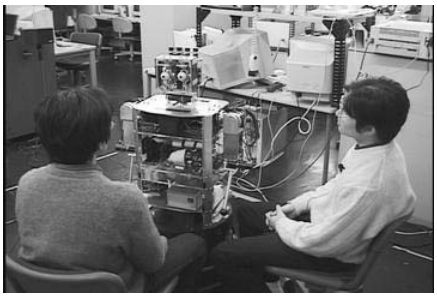
R:(interest)>A
A:はい、何か聞いてみてください>B



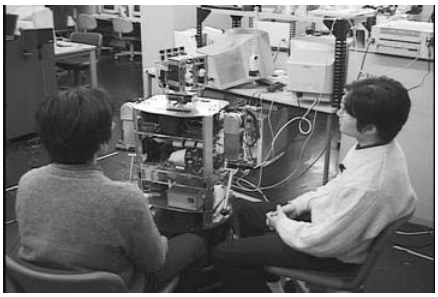
R:(interest)>B
B:何歳ですか>R



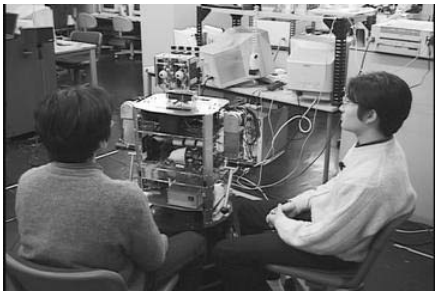
R:私は4歳です>B



B:背の高さは ...>R
A:すいません>R



R:(side-glance)
少し待ってください>A



R:何でしたっけ>A
...

Figure 4: Sample experiment.

参考文献

[1] 中西英之, 吉田力, 西村俊和, 石田亨, “FreeWalk: 3次元仮想空間を用いた非形式なコミュニケーションの支援,” 情処学論, Vol.39, No.5, pp.1356-1364, 1998.

[2] R. Vertegaal, “Look who's talking to whom ? Mediating Joint Attention in Multiparty Communication and Collaboration,” Cognitive Ergonomics Department University of Twente, 1998.

[3] 白井克彦, 小林哲則, 岩田和則, 深沢克夫, “ロボットの柔軟な対話を目的とした音声入出力システム–WABOT-2 における会話系,” 日本ロボット学会誌, Vol.3, No.4, pp.104-113, 1985.

[4] D.Goddeau, E.Brill, J.Glass, C.Pao, M.Phillips, J.Polifroni, S.Seneff and V.Zue “GALAXY: A Human-Language Interface to On-Line Travel Information,” Proc. ICSLP'94, pp. 707-710, 1994.

[5] 松井俊浩, 麻生英樹, John Fry, 浅野太, 本村陽一, 原功, 栗田多喜夫, 速水悟, 山崎信行, “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] Nuance Home (http://www.nuance.com/)

[7] K. Nagao and A. Takeuchi, “Social interaction: Multimodal Conversation with Social Agents,” Proc. AAAI'94, pp.22-28, 1994.

[8] 城塚音也, 桑田喜隆, 安地亮一, 小泉宣夫, “遠隔会議を対象とした音声対話モニタリングによる対話支援システム,” 情処学論, Vol.39, No.5, pp.1240-1247, 1998.

[9] P. Dourish, S. Bly, “Portholes: Supporting Awareness in a Distributed Work Group,” Proc. ACM CHI'92, pp.541-547, 1995.

[10] R. Vertegaal, B. Velichkovsky, G. van der Veer, “Catching the Eye: Management of Joint Attention in Cooperative Work,” SIGCHI, Vol.29, No.4, 1997

[11] 本田新九郎, 富岡展也, 木村尚亮, 岡田謙一, 松下温, “在宅勤務者の疎外感の解消を実現した位置アウェアネス・アウェアネススペースに基づく仮想オフィス環境,” 情処学論, Vol.38, No.7, pp.1454-1464, 1997.

[12] 横山真男, 青山一美, 菊池英明, 帆足啓一郎, 白井克彦, “人間型ロボットの対話インタフェースにおける発話交替時の非言語情報の制御,” 情処学論, Vol.40, No.2, pp.487-496, 1999.

[13] N. Ward, “Using Prosodic Clues to Decide When to Produce Back-channel Utterances,” Proc. ICSLP'96, pp.1728-1731, 1996.

[14] J. Hirasawa, N. Miyazaki, M. Nakano, T. Kawabata, “Implementation of Coordinate Nodding Behavior on Spoken Dialogue System,” Proc. ICSLP'98, Vol.6, pp.2347-2350, 1998.

[15] K. Watanuki, K. Sakamoto and F. Togawa, “Analysis of Multimodal Interaction Data in Human Communication,” Proc. ICSLP'94, pp.899-902, 1994.

[16] H. Sacks, E.A. Schegloff and G. Jefferson, “A simplest systematics for the organization of turn taking in conversation,” Language, Vol.50, No.4, pp.696-735, 1974.

[17] 久保田千太郎, 松坂要佐, 小林哲則, “グループ会話ロボットにおける顔画像処理システム,” 信学技法, NLC99-85 PRMU99-268, pp.49-56, 2000.

[18] 田宮大介, 松坂要佐, 小林哲則, “ロボット頭部に設置した 2 系統のマイクによる音源定位,” 日本音響学会春季研究発表会 講演論文誌, pp.469-470, 1999.

[19] 村井則之, 小林哲則, “話者性と発話交代を考慮した複数話者対話音声の認識,” 信学論 (D-II), Vol.83-D-II, No.11, pp.2465-2472, 2000.

[20] 東芝音声システム (http://www2.toshiba.co.jp/pc/service/download/mimi/index_j.htm)

[21] ROBITA オンラインビデオライブラリ (http://www.tk.elec.waseda.ac.jp/robita/)

[22] D. A. Norman (野島久雄訳), “誰のためのデザイン?–認知科学者のデザイン原論,” 新潮社, 1990.

[23] 西本卓也, 志田修利, 小林哲則, 白井克彦, “マルチモーダル入力環境下における音声の協調的利用–音声作図システム S-tgif の設計と評価–,” 信学論 (D-II), Vol.J79-D-II No.12, pp.2176-2183, 1996.

[24] 石塚満, “マルチモーダル擬人化エージェントシステム,” システム/制御/情報, Vol.44, No.3, pp.128-135, 2000.

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-0317-2

身体表現を用いたロボットの発話生成
Utterance Generation with Physical Expression

今井 倫太 小野 哲雄 石黒 浩
Michita Imai Testuo Ono Hiroshi Ishiguro
ATR 知能映像通信研究所
ATR Media Integration and Communications Research labs.
michita@mic.atr.co.jp

Abstract

This paper proposes a speech generation system named Linta-III, which generates an utterance dependent on a real world situation. To generate the situated utterance, Linta-III has a joint attention mechanism, which develops joint attention between a person and a robot. The joint attention mechanism employs eye-contact and an attention expression. The eye-contact promotes the relationship between the person and the robot. The attention expression manifests relevant sensor information with a physical expression. With the eye-contact and attention expression, the joint attention mechanism is able to draw the person's attention to the same sensor information as the robot. As a result of the joint attention, Linta-III is able to omit obvious words in the situation from an utterance description. We also conducted a psychological experiment on the development of joint attention. The results indicated that the eye-contact and attention expression are significant factors in the development of joint attention.

1 はじめに

近年、ペットロボットに代表されるコミュニケーション型ロボットが多く開発されている。近い将来、一般の人々が日常生活の場面でコミュニケーション型ロボットと触れ合う機会も出てくるであろう。本稿では、人間とロボット間の柔軟なコミュニケーションの実現を目指し、センサ情報を利用した発話生成システムについて取り上げる。

人間は、そばの人と会話をする場合、状況から自明な事柄を省き状況に依存した発話文をしばしば用いる[Barwise and Perry, 1983]。実世界の状況を人間と共有し活動するためには、ロボットも同様に、回りの状況を参照しつつ、発話文を生成したり、人からの発話文を解釈する必要がある。つまり、ロボットは、状況に依存した発話文を扱うために、文脈や状況から重要となるセンサ情報を獲得できなければならない。

センサ情報を用いた対話システムの研究は、従来から行われてきた[Chapman, 1991] [Grosz, 1977][今井 他, 1994]。これらのシステムでは、センサ情報の中で重要な情報に注意を向ける機構（注意機構）を用いている。注意機構によって参照すべきセンサ情報が予め絞られているため、発話処理のための計算コストを削減することができる。特に、対話システム Linta-II[今井 他, 1994]では、ロボットの行動や周りの状況に応じてセンサ情報を選択し、発話処理に必要となる状況を獲得する。例えば、ロボットが障害物に近づいている場合には、注意機構が前のセンサ情報に注目し、障害物に関連して発話処理が行われる。ここで、人間が「だめ！」と発話すると、 Linta-II は、前のセンサ情報に関連して発話文の処理を行い、「前進するのがだめである」と解釈することができる。さらに、コミュニケーションにおける状況の役割は、言語学や認知科学の分野でもその重要性が古くから研究されてきた。コミュニケーションにおける状況の役割を明確にしている理論の一つに関連性理論がある[Sperber and Wilson, 1986]。関連性理論では、コミュニケーションにおける状況の参照を、人間同士の情報伝達の観点からモデル化した。関連性理論では、相互顕在化という用語を用いて、二人以上の人間が同じ状況へ注目していく（心の中で状況を顕在化する）過程で人間のコミュニケーションを捉えている。つまり、人間が、メッセージ（発話等）によって、同じ状況に相手を注目（相互顕在化）させることにより情報を伝えあっていると主張している。実世界の情報を利用したロボット用対

話システムを実現するためには、関連性理論で扱われる相互顕在化の概念も非常に重要になってくる。

しかし、注意の機構によってセンサ情報を選択するだけでは、状況を利用した発話生成システムを構築するにあたって不十分である。構築の際には、以下の三つの課題が考えられる。

1. ロボットが注目している情報にどのようにして人間の注意を引くのか？
2. ロボットが持つ情報伝達の意図をどのようにして人間に気づかせるのか？
3. 注意機構において人間の注意をどのように扱うべきか？

課題 1 は、従来の注意機構が注目しているセンサ情報を人間に対してまったく表出していなかったことに起因する。ロボットの注意が表出されないため、人間は、ロボットの注目している情報に気づくことがない。ロボットの注意に人間が気づいていないことは、発話生成システムが、実世界の情報に関する発話文を生成するうえで深刻な問題となる。例えば、博物館で案内ロボットが展示物を説明する際に注意を表出せずに指示語を用いたとすると、人間は、ロボットが何に注目し説明しているのかが分からず発話が理解できないことになってしまう。

課題 2 は、課題 1 を解決するにあたって生じる問題である。関連性理論では、聞き手が、話し手の情報伝達の意図を推測できるかどうかに依存して、相互顕在化（話者と聞き手が同じ情報に注目している状態）が生じると主張している。例えば、ロボット（話し手）が、箱の前で停止して、（前に進むために）「これどけて！」と発話したとする。人間が、この発話を理解するためには、ロボットの意図（前に進みたい）に気づく必要がある[Ono and Imai, 2000]。意図の推測によって、人間は箱へ注目し（相互顕在化）、ロボットの発話を理解することができる。つまり、聞き手が、話者の情報伝達の意図を推測するかどうかは、人の注意を引き込む際に確実に影響を与える。

課題 3 は、従来の注意機構で扱っていた注意の対象に関する問題である。従来の注意機構では、センサ情報を選択するだけであり、人間の注意まで扱っていなかった。しかし、発話生成システムは、状況依存の発話文を生成するために、人間の注意も扱わなければならない。なぜなら、人間が注意を向けていないときに、状況依存の発話文を生成したとしても人間が理解できない可能性が高いからである。

本稿は、人とロボット間の共同注意[Moore and Dunham, 1985]を実現するために共同注意機構を提案する。共同注意とは、二人以上の人間が同じ実世界の情報に注目している状態を指す。人とロボット間に共同注意を実現することにより、共同注意機構は、ロボットが、状況に依存した発話文を生成したとしても、人間に理解させることができる。共同注意機構では、課題 1 を解決するために、注意の表出の機能を持っている。この機能は、ロボットの身体表現によって、ロボットが注目しているセンサ情報の情報源を表出する。具体的には、注目している情報源へロボットの視線を向け、また、ロボットの腕によって対象物を指し示す。

課題 2 を解決するために、共同注意機構では、アイコンタクト機能を持つ。アイコンタクト機能によって、人間とロボットにコミュニケーションのための関係を与える。アイコンタクトは、ロボットの顔を人間の方向へ向けることによって実現される。人間と関係を形成することによって、人間に、ロボットのコミュニケーションの意図を推測するよう仕向ける。人間は、コミュニケーションの意図を察する結果、注意の表出によって示されている実世界の情報に気づくことができる。

課題 3 を解決するために、共同注意機構では、ロボットの注意と人間の注意を扱うための注意座標を持つ。人間の注意は、共同注意の達成過程に応じて注意座標上で表現される。人間の注意が注意座標上で表現されているので、共同注意機構は、人間が発話を理解しやすいように、発話文の状況依存性を変化させることができる。

また、本稿では、人間型ロボット Robovie 上に、共同注意機構を用いた発話生成システム Linta-III を構築した。

本稿の構成は、以下の通りである。二章で、状況依存の発話文を生成する際に生じる問題点および、我々が開発したロボット Robovie について説明する。三章では、共同注意機構による共同注意の達成の手法および Linta-III の構成について説明し、四章で、 Linta-III による発話生成について説明する。五章では、共同注意機構の評価実験について述べ、六章で、実験結果および共同注意機構の効果について考察する。七章で、まとめと今後の課題について述べる。

2 対話における共同注意

2.1 共同注意

人々は、同じ実世界の情報（物や、人物、景色等）に注目しながらコミュニケーションすることがよくある。社会心理学では、同じ物に注目することを共同注意[Moore and Dunham, 1985]と呼ぶ。ただし、共同注意とは、二人以上の人間が同じ情報に注目しているだけなく、相手が同じ情報に注目していることにもお互いに気づいている状態である。次の発話例は、共同注意が形成されていてはじめて解釈可能な発話である。

発話例 1

R1-1：ポスター見てね。
R1-2：面白いよ。

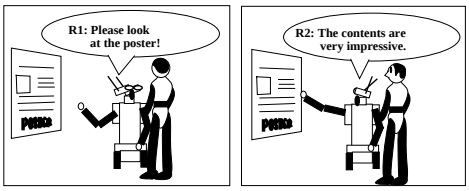


Figure 1: Joint attention and an utterance.

この例は、ロボットがポスターについて説明している場面である（図 1）。左の図は、人間が、ロボットにだけ注目しており、ポスターの存在に気づいていない場面である。ロボットは、この場面で発話 *R1-1* によりポスターについて触れている。右の図は、人間が、発話 *R1-1* によってポスターの存在に気づき、ポスターを見ている場面である。また、人間は、ロボットもポスターに注目していることにも気づいている。つまり、右の図では、人間とロボットが共同注意を持っている場面を示している。共同注意によって、人間は、ロボットの発話 *R1-2* がポスターについて語っていると理解することができる。

本稿では、人間とロボット間の共同注意を利用した発話生成システムを構築する。共同注意を基本とした発話のやりとりを実現するためにも、発話生成システムは、意図した情報へと人間の注意を引き付ける機構が必要となる。

2.2 伝達意図の推測

ロボットは、人間にとってコミュニケーションの対象となりづらい可能性があり、たとえロボットが注意を表出したとしても、人間は、ロボットと共同注意を本当に持つことができるのかといった疑問が残る。我々は、この疑問に対して心理実験を行なった[Ono and Imai, 2000]。実験結果から、被験者とロボット間の関係が共同注意の構成に影響を与えることが分かった。具体的には、被験者の前にロボットが突然現れゴミ箱にぶつかり、ゴミ箱をどかすように被験者に依頼するといった実験であった。結果、ロボットとの関係を与えられた被験者は、ロボットからの依頼を理解しゴミ箱をよけた。一方、関係を与えられなかった被験者は、依頼そのものを聞き取ることができなかった。つまり、被験者は、見ず知らずのロボットの意図（前に進みたいという依頼の意図）を推測しなかったので、ゴミ箱への共同注意を持たなかったといえる。以上の実験結果を踏まえると、ロボットが人間に共同注意を持たせ状況に依存した発話文を扱おうとすると、発話生成システムは、人間とロボット間に関係を与える機構を持つ必要があると言える。

2.3 発話生成における注意

状況に依存した発話文を生成するためには、注意機構で、ロボットの注意と人間の注意の双方を把握し扱う必要がある。従来の注意機構[Imai *et al.*, 1994] [Chapman, 1991] [Grosz, 1977]では、人間の注意を扱っていなかったので、発話例 1に代表される状況に依存した発話文を人間が理解可能かどうか判断することができなかった。従来の注意機構は、単純にセンサ情報を選択し注目するのみの機構であった。以下の発話例について考えてみる。

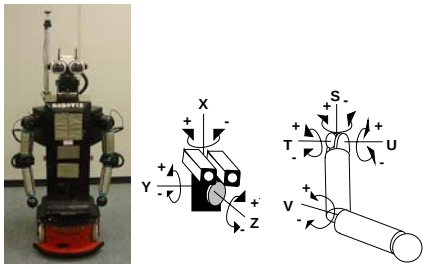


Figure 2: Everyday Robot “Robovie”

発話例 2 状況*:* 障害物の前で
R2-1: 前に進めない。

発話 *R2-1* を生成するためには、実世界の情報への共同注意を形成する必要がないので、人間の注意を扱う必要がない。ロボットが、障害物を検知するセンサさえ持っていれば、従来の注意機構でも生成可能な発話文である。一方、以下の発話文を人間が理解することができるか判断するためには、注意機構が、人間の注意を把握しなていなくてはならない。

発話例 3 状況*:* 人間と障害物の前で
R3-1: これどけて。

発話 *R3-1* は、障害物を動かしてもらうことを人間に依頼している。ここで、人間が「これ」を理解するためには、障害物に注目している必要がある。よって、状況に依存した発話文を生成するためには、注意機構で、人間の注意を把握している必要がある。

2.4 Robovie

ロボット用発話生成システムを構築するために、本稿では、人間型ロボット Robovie を用いる (図 2左)。Robovie は、片方 4 自由度の腕と 3 自由度の頭部によって、人間とほぼ同等の身体表現が可能になっている。また、3 輪の台車で移動することができ。多数のセンサ（触れられたことを判定するタッチセンサ、障害物を検知する超音波距離センサ、画像認識のための二種類のカメラ）を持つ。カメラの一つは、全方位カメラとなっており Robovie の周囲の画像を一度に取り込むことが可能になっている。人間の位置は、全方位カメラで取り込んだ画像から肌色領域を探すことによって検知している。二つ目のカメラは、二つの CCD カメラであり、人間の目の位置にあたる場所に設置され、立体視に用いられる。また、 Robovie の視線を人間

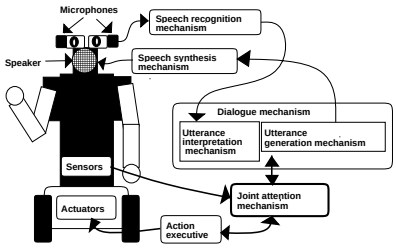


Figure 3: An overview of Linta-III

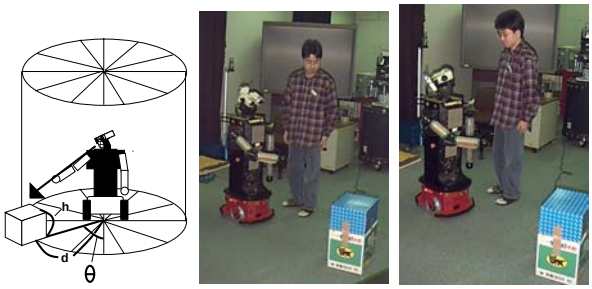


Figure 4: Attention expression and eye-contact by Robovie

に感じさせる効果も狙っている。Robovie は、音声認識および合成装置も持ち、音声によるコミュニケーションを人間ととることができる。

ここで、Robovie のジェスチャーを説明するために各部の姿勢を表す記号を導入する。頭部の姿勢は、$Head(X,Y,Z)$ で表す。X および、Y、Z は、それぞれ、頭を左右に向ける軸および、上下に向ける軸、左右に傾げる軸を表す（図 2右）。腕の姿勢は、$Hand_i(S,T,U,V)$ で表す。ここで、i は、左腕 $(i = l)$ か右腕 $(i = r)$ を表す。各軸 S および、T、U、V の可動個所は、図 2右に示す通りである。

3 Linta-III と共同注意機構

本稿では、人間とロボット間に共同注意を形成する共同注意機構および発話生成システム Linta-III を提案する。

図 3に Linta-III の概略を示す。Linta-III は、共同注意機構および、対話機構（発話生成機構および解釈機構）、行動制御機構で構成される。共同注意機構は、Robovie のセンサおよびアクチュエータを用いて人間と共同注意を実現する。構成された共同注意を利用して、Linta-III は、状況に依存した発話文を生成する。

共同注意機構は、Robovie が注目している実世界の情報に人間の注意を引き付けるために注意の表出機能を持つ。注意の表出機能は、Robovie の視線の動きおよび、腕による方向指示によって、実世界の情報への Robovie の注意を表出する（図 4真中）。

また、注意の表出機構は、人間とコミュニケーションするための関係を築くためにアイコンタクト機能を持つ。アイコンタクト機能によって築かれた関係によって、注意の表出機構は、人間に、Robovie の注意の表出を気づかせ、Robovie と同じ実世界の情報に注目させる。アイコンタクトは、Robovie の顔を人間がいる方向に向けることによって実現される（図 4右）。

さらに、共同注意機構は、Robovie の注意および人間の注意を扱うために注意座標を持つ。注意座標は、注意の表出機能およびアイコンタクト機能に対する人間の注意の変化を予想し、座標上に人間の注意を追加する。つまり、注意機構は、ある実世界の情報にたいして共同注意が構成されているかどうかの表現をもっている。Linta-III は、人間とのインタラクションで注意座標の状態を参照することにより、人間が理解しやすいように発話文の状況への依存度を変化させることができる。

3.1 注意の表出

注意の表出は、頭部と腕によって Robovie が注目中のセンサ情報を表出する。共同注意機構では、Robovie の各センサデータが筒上の座標系 (θ, d, h) にマッピングされる。筒上の座標系を用いることによって、共同注意機構は、注意を表出する Robovie の振舞いへと容易に変換することができる。

図 4左に、注意の表出の一例を示す。図では、Robovie が、距離 d で方向 θ にある箱への注意を表出している。ここで、箱の高さは h である。箱の位置に従って、Robovie の頭部の姿勢は $Head(\theta, \arctan(d/(L-h)),\ 0)$ となり箱の方向へ顔を向ける。ここで、L は、Robovie の頭の地面からの高さである。また、右腕が、$Hand_r(0, arctan(cos\theta tan\theta_1), arcsin(sin\theta_1 sin\theta), 0)$ となり、箱を指し示す。ここで、θ_1 は、Robovie の胴体と腕の間のなす角度 $arctan(d/(H-h))$ である。H は、Robovie の肩までの高さである。

3.2 アイコンタクトによる人間との関係

アイコンタクト機能は、ロボビーの視線を人間に向け、人間と関係を形成する。人間の位置は、全方位カメラで取り込まれ、筒上の座標上で $(\theta_p, \_, T)$ と表現される。ここで、座標上の θ_p は、Robovie の正面から人がいる方向への角度を表し、T は、全方位センサまでの高さを表す。座標上に、人間の位置までの距離情報がないのは、全方位カメラで捉えられるのが方向の情報だけだからである。アイコンタクトの際には、座標上の人間の位置情報に従って、Robovie の首の姿勢が $Head(\theta_p,\ arctan(D/T),\ 0)$ となる。ここで、D は、全方位カメラが、人の位置を検知できる限界の距離である。

Robovie の注意の表出の振舞いを人間に気付かせるために、共同注意機構は、注意の表出の際にアイコンタクトを

頻繁に行なう。例えば、 Linta-III が発話例 *R1-1* を生成するさいに、共同注意機構は、ポスターを腕で指し示しながら Robovie の頭をポスターと人間の方向に交互に動かす。アイコンタクトの結果、人間は、 Robovie からコミュニケーションの意図を感じ取り、 Robovie の注意の対象について気づくことができる。

3.3 センサ情報への注意

共同注意機構は、注意座標上の表現をセンサ情報から作り出し Robovie の注意および人間の注意を扱う。注意座標は、図 4左の円筒形の座標にセンサ情報の属性 f を付け加えた表現となっている。つまり、注意座標は、以下の表現となる。

$$f[(\theta, d, h)] \tag{1}$$

属性 f には、四つの種類がある（人間 p および、物体 o、ポスター e、Robovie　r）。例えば、 Robovie から -30 度（右側）の方向に 50cm 離れた位置にある物体は、$o[(-30, 50, U)]$ と表現される。ここで、 U は、地面からの超音波距離センサの高さである。

人間 p に対する注意座標には、実世界の情報への人間の注意も表現するために下に示す表現を用いる。

$$p[(\theta_p, \_, T), a], \tag{2}$$

ここで、 a は、人間が注意を向けている対象が入る項であり、物体等の注意座標が直接書き込まれる。例えば、人間が、前段落で例として挙げた物体に注意を向けているとき、注意座標は、 $p[(\theta_p, \_, T), (o[(-30, 50, U)])]$ となる。

共同注意が、注意の表出機能およびアイコンタクト機能に従って形成されるので、共同注意機構は、それぞの機能の実行過程に従って注意座標の各情報を変更する。以降、発話例 1に従って共同注意の構成過程を示す。この例では、 Robovie が、 θ_p にいる人間 p と θ_e にあるポスター e に注意を向けている。この注意に従って、共同注意機構は、以下の二つの注意座標を持つ。

$$p[(\theta_p, \_, T), \_], \quad e[(\theta_e, \_, T)] \tag{3}$$

ポスタの方向 θ_e は、人間と同様に、全方位カメラから色によって認識される。

ポスターについて説明するために、共同注意機構は、 Robovie の頭を人間の方向へ向け、アイコンタクトを行なう。人間が、アイコンタクトにより Robovie の存在に気づくので、共同注意機構は、人間の注意座標に、 Robovie への情報 $r[(0, 0, 0)]$ を追加する。結果、注意座標は、以下の形となり、人間が Robovie に気づいてることを表す。

$$p[(\theta_p, \_, T), (r[(0, 0, 0)])], \quad e[(\theta_e, \_, T)] \tag{4}$$

ここで、 Robovie r の位置情報は、 Robovie 自体の位置を表すため 0 としている。

Table 1: Utterance generation rules

Rule1	$v(\_, f)$	$\rightarrow$	$v(r, \_)$
Rule2	$v(\_, f), p[\_, \_]$	$\rightarrow$	$Hello, v(\_, f)$
Rule3	$v(\_, f), p[\_, (r, \_)]$	$\rightarrow$	$v(\_, f)$
Rule4	$v(\_, f), p[\_, (r, f)]$	$\rightarrow$	$v(\_, dp)/\ v(\_, \_)$

次に、共同注意機構は、 Robovie の視線をポスターへ向け腕で指し示し、 Robovie のポスターへの注意を表出する。注意の表出によって人間がポスターに注目するので、共同注意機構は、ポスターの情報 $e[(\theta_e, \_, T)]$ を人間の注意座標に加え、以下の注意座標を得る。

$$p[(\theta_p, \_, T), (r[(0, 0, 0)], e[(\theta_e, \_, T)])], \quad e[(\theta_e, \_, T)] \tag{5}$$

結果、共同注意機構は、ポスターに対して人間と共同注意を持っているといった表現を注意座標上に持つ（$p[(\theta_p, \_, T), (r[(0, 0, 0)], e[(\theta_e, \_, T)])]$）。

4 Robovie とのインタラクション

4.1 発話生成ルール

Linta-III では、身体的なインタラクションと言語的なインタラクションをより柔軟な形で実現するために、発話生成と Robovie の振舞いが平行に実行される実装形態を考える。よって、 Linta-III は、注意の表出機能およびアイコンタクト機能の実行と独立して、発話生成を行なうことができる。独立していることによって、実世界で起る予期しない出来事（腕がぶつかる等）に独立して、発話生成を遂行することができる。また、 Linta-III は、注意座標を参照し発話文を生成するので、共同注意の成立の度合を実時間で発話生成に反映することができる。

Linta-III は、表 1に示す発話文生成ルールを持つ。発話文生成ルールは、注意座標に適した形で発話文の状況依存性を変化させ、共同注意の達成の度合に適した形の発話文を生成することができる。表 1のルールは、$(utterance\ \ content), (attentin\ \ coordinate) \rightarrow (altered\ \ utterance\ \ content)$ といった形になっている。発話内容 $(utterance\ \ content)$ は、$v(subject, object/complement)$ といった形で、動詞 v および、主語 $subject$、目的語 / 補語 $object/complement$ といった要素を持つ。例えば、発話「ポスター見てね。」は、$look(p, e)$ といった表現となる。また、本稿では、共同注意を扱うことを目指しているので、表 1では、人間の注意座標 $p[\_, \_]$ のみを載せた。表では、 r は Robovie を、f は任意の種類の属性を、 dp は指示語を表す。また、 $\_$ は、任意の値を取ることができる。さらに、表のルールでは、注意座標の中から位置の情報を省略している。

発話を生成するときに全方位カメラの視野に人間がいない場合、 Linta-III は Rule1 を用いる。 Rule1 は、人間

がそばを偶然通る可能性もあるので、Robovie の視点で状況を報告する発話内容 $v(r,\_)$ を生成する。発話例 2は、Rule1 を用いた例である。この例では、物体をどけてもらう人間がそばにいないので、Robovie r は、前に進めないこと $(\neg proceed(r))$ を報告している。この発話文生成を Rule1 に当てはめると $move(p,o) \rightarrow \neg proceed(r)$ となる。

全方位カメラの視野に人間がいるが、Robovie とコミュニケーションの関係が成立していない場合、Linta-III は Rule2 を用いる。Rule2 では、最初に挨拶をして人間と関係を築く。挨拶の後で、Rule2 は、センサ情報 f を省略せずに発話文を生成する。なぜなら、この状況では、人間が、まだ実世界の情報 f に注意を向けていないからである。

人間が、アイコンタクトによって Robovie と関係を既に形成している場合、Linta-III は、Rule3 を用いる。しかし、この状況でも人間はまだ f に注意を向けていないので、Rule3 でも、f を省略せずに発話文を生成する。発話例 *R1-1* は、Rule3 を用いた例である。この例では、Linta-III が、ポスター e を見るよう人間 p に依頼している。しかし、人間がポスターにまだ注目していないので、Linta-III は、ポスター e を発話内容 $look(p,e)$ から省略していない。この発話文生成を Rule3 に当てはめると $look(p,e), p[\_,(r)] \rightarrow look(p,e)$ となる。

注意の表出機能およびアイコンタクト機能によって共同注意が実現されると Linta-III は Rule4 を用いる。共同注意が成立している場合、Linta-III は、指示語を用いて f を参照する（場合によっては、指示語も用いず発話文から省略する。）。発話例 *R1-2* および *R3-1* は、Rule4 を用いた例である。発話 *R1-2* では、Linta-III が、人間 p に、ポスター e が面白いことを伝えようとしている。注意座標を参照すると、ポスターへの共同注意 $p[\_,(r,e)]$ が既に形成されているので、Linta-III は、発話内容 $is\text{-}a(e,fun)$ から e を省略し、指示語「これ」を用いる。省略後の発話内容は $is\text{-}a(this,fun)$ となる。この発話文生成を Rule4 に当てはめると $is\text{-}a(e,fun), p[\_,(r,e)] \rightarrow is\text{-}a(this,fun)$ となる。

4.2 共同注意を用いた発話生成

本章では、発話例 3を用いて、共同注意に応じた発話生成例をいくつか紹介する。Linta-III が生成する発話文は、前章で触れた通り、注意座標に応じて選択された発話文生成ルールに従って生成される。

発話例 3の状況では、Robovie が障害物で前に行けなくなっている。この状況で、Linta-III は、人間に障害物をどけてもらおうと発話内容 $move(p,o)$ を持つ。ここで、全方位カメラに人間を捉えることができない場合、共同注意機構が、物（障害物）の存在だけに注意を向けているので注意座標は $o[\_]$ となる。人間への注意 $p[\_,\_]$ が注意座標にないので、Linta-III は、Rule1 を選び、発話内容 $\neg proceed(r)$ を生成する。

Figure 5: Experimental scenes: eye-contact between a person and Robovie (in the left figure), joint attention to the poster (in the center figure), and a person looking at Robovie's hand (in the right figure).

全方位カメラが人間の位置 $(\theta_p,\_,T)$ を捉えることができた場合、人間への注意 $p[(\theta_p,\_,T),\_]$ が注意座標に加えられる。このとき、共同注意機構は、Robovie の頭を動かし人間とアイコンタクトを取ろうとする。しかし、人間がすばやく移動してしまうこともあり、アイコンタクトは必ずしも達成されない。アイコンタクトが達成される前に Linta-III が発話生成を始めてしまうと、注意座標は、まだ、$p[(\theta_p,\_,T),\_]$ および $o[\_]$ の二つになっている。よって、Rule2$(move(p,o),\ p[(\theta_p,_,T),_] \rightarrow Hello, move(p,o))$ が選ばれる。結果、Linta-III は、「こんにちは！障害物をどけてください。」と発話する。

アイコンタクトが達成されると、注意座標は $p[(\theta_p,\_,T),(r(0,0,0),\_)]$ となる。よって、Linta-III は、Rule3$(move(p,o),\ p[(\theta_p,_,T),(r(0,0,0),_)] \rightarrow move(p,o))$ を選ぶ。結果、発話文「障害物をどけて下さい」を生成する。

発話の生成の前に、アイコンタクトと注意の表出の双方が達成されると、人間も障害物へ注意を向けている可能性が高くなるので、注意座標は、$p[(\theta_p,\_,T),(r(0,0,0),o(\theta_o,d,h))]$ となる。従って、Linta-III は、Rule4 $(move(p,o),\ p[(\theta_p,\_,T),\ (r(0,0,0),o(\theta_o,d,h))] \rightarrow move(p,this))$ を選び、結果、発話例 3を生成する。ここで、指示語「これ」は、Robovie と障害物との距離によって選択されている[Imai *et al.*, 1999]。

5 行動実験

本章では、共同注意の達成におけるアイコンタクトの影響を行動実験で検証する。本実験では、20 人の被験者を半数づつ二つの群に分けた。一つの群（実験群）は、被験者とアイコンタクトを行なう Robovie を与えた。もう一つの群（対照群）には、アイコンタクトを行なわない Robovie を与えた。腕による注意の表出は、両群共に行なった。Robovie が注意を向ける対象は、壁に貼られたポスターである。実験では、注意の表出にしたがって被験者がどこを

Table 2: Experimental result: comparison of the number of people who looked at a poster pointed out by Robovie by the effect of eye-contact. ($U = 5, p < .01$)

	Saw a poster	Saw Robovie's hand
With eye-contact	10	0
Without eye-contact	1	9

見たかを記録した。

実験は、以下の手順で行なった。始めに、 Robovie が被験者の前を通りすぎ、ポスターの前で止まる。ここで、被験者は、 Robovie とポスターの両方とも見える場所に立っている。次に、 Robovie は、被験者の方に振り返り、「これ見てね。」と発話しながら、腕でポスターを指し示す。この時点で、実験群の被験者には、 Robovie がアイコンタクトとポスターを見る動作を繰り返す。対照群の被験者には、 Robovie が顔を正面に向けたまま動かさない。

表 2に実験結果を示す。被験者は、 Robovie がアイコンタクトした場合（図 5左）、ポスターへ視線を向けた（図 5真中）。一方、 Robovie がアイコンタクトを行なわなかった場合、多くの被験者がポスターを見ずに Robovie の腕先を見た（図 5右）。結果、共同注意の成立にアイコンタクトが影響を与えることが確認された ($U = 5, p < .01$)。

6 考察

実験結果から、アイコンタクトをしない Robovie を与えられた被験者は、 Robovie が、実世界の物体を指し示す能力を持つことに気づかなかったといえる。つまり、被験者と Robovie の間は、二項関係となっており、そこに第三の実世界の物が入って来ていない。従来の人間とロボットのコミュニケーションは、二項関係をベースとした物であったといえる。一方、アイコンタクトによって、被験者は、 Robovie の注意の表出に気づき、ポスターを見ることができた。この実験結果は、被験者および、 Robovie、ポスターの間に三項関係が二項関係に代って生じていることを示している。つまり、 Linta-III は、三項関係を元にして人間とロボットの新しいコミュニケーションを実現していると言える。

また、共同注意機構は、 Robovie の注意を、身体表現（アイコンタクトと注意の表出）によって伝えた。一方、関連性理論[Sperber and Wilson, 1986]では、コミュニケーションを、論理的な推論のみによって説明している。しかし、実験結果から、アイコンタクトによって形成される関係は、論理的な推論や知識処理よりもより基本的な効果を持っていると思われる。 Linta-III は、身体表現によって、実世界の情報を参照しながらコミュニケーションすることができ、より自然なユーマンロボットインタフェースを実現することができる可能性がある。

7 まとめ

本稿では、共同注意機構を提案し、実世界の情報を利用した発話生成システム Linta-III を実装した。また、実装に際して人間型ロボット Robovie を用いた。共同注意機構が、 Robovie の注目している実世界の情報に人間の注意を引き付けることができるので、 Linta-III は、状況から自明な単語を発話文から省略することができる。

共同注意機構は、人間とのアイコンタクト機能および注意の表出機能（視線および腕による実世界の情報の指し示し）を持つ。また、獲得しているセンサ情報に対して共同注意が構成されているかどうか判断するために注意座標も持っている。 Linta-III は、注意座標を参照することにより共同注意の達成度に応じて発話文を生成することができる。また、本稿では、共同注意機構の有効性を検証するために行動実験も行なった。実験結果は、人間と Robovie のアイコンタクトが共同注意の成立に影響を与えていることを示していた。

本稿では、共同注意を利用した発話生成を提案した。今後の課題として、人間からの発話文の解釈に共同注意機構を用いて行く予定である。また、さらに、人間が注意を向けている実世界の情報に対して共同注意を成立させる手法も今後扱っていかなくてはならない課題である。

参考文献

[Barwise and Perry, 1983] J. Barwise and J. Perry. *Situations and Attitudes*. Mit Press, 1983.

[Chapman, 1991] D. Chapman. *Vision, Instruction, and Action*. Mit press, 1991.

[Grosz, 1977] B. J. Grosz. The representation and use of focus in a system for understanding dialogs. In *IJCAI-77*, pages 67–76, 1977.

[Imai *et al.*, 1994] M. Imai, K. Hiraki, and Y. Anzai. Human-robot interface with attention. *Tranc. of IEICE (D-II)*, J77-D-II:1447–1456, 1994.

[Imai *et al.*, 1999] M. Imai, K. Hiraki, and T. Miyasato. Physical constraints on human robot interaction. In *IJCAI99*, pages 1124–1130, 1999.

[Moore and Dunham, 1985] C. Moore and P. J. Dunham. *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associatcs, Inc., 1985.

[Ono and Imai, 2000] T. Ono and M. Imai. Reading a robot's mind: A model of utterance understanding

based on the theory of mind mechanism. In *Proceedings of AAAI-2000*, pages 142–148, 2000.

[Sperber and Wilson, 1986] D. Sperber and D. Wilson. *Relevance: Communication and Cognition*. Oxford: Basil Blackwell, 1986.

[今井 他, 1994] 今井, 開, and 安西. 注意機構を利用したヒューマンロボットインタフェース. 信学論 *(D-II)*, J77-D-II:1447–1456, 1994.

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-0317-3

パーソナルロボット PaPeRo の音声認識インタフェース

Speech Recognition Interface for Personal Robot "PaPeRo"

岩沢 透

Toru IWASAWA

NEC ラボラトリーズ 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 personal robot "PaPeRo". A Speech Recognition interface for mobile robot (including PaPeRo) needs to have unspecified speaker's voice recognition ability and adaptability for the distance between a microphone and a user. A major problem in realizing the speech recognition interface is to distinguish between noises and normal utterances of robot's vocabulary. In this paper we suggests a noise rejection method using "rejection dictionary". The rejection dictionary is added to a dictionary of the robot's vocabulary and recognizes noise and unknown utterance without interrupting a normal utterance of the robot's vocabulary. This paper also describes an implementation method and evaluation result of the rejection dictionary for PaPeRo.

1 はじめに

近年，人間と共生することを目的としたパートナー型ロボットの研究が盛んである．一方，音声認識技術の向上に伴い人間とのコミュニケーション手段に音声を利用する試みがカーナビや KIOSK 端末などで行われており，ロボットの分野にも応用がなされ始めている．我々は人間との音声対話機能をもつ自律移動型パーソナルロボット PaPeRo の研究開発を行っている．移動ロボットと人間の自然なインタラクションを実現するためには，利用者の身体的自由度の高い音声認識インタフェースが求められる．これは，利用者にマイクを装着させない，発話時に多少距離が離れていても音声認識可能といった利用者の負担を軽減させるための音声認識インタフェースである．このような音声認識インタフェースを構築する上で問題になるのが雑音や周囲会話，未知発話といった不要音声の棄却である．不要音声を音声認識語彙に誤認識するということはロボットが不要音声により何らかの誤動作をしてしまうことを意味する．その結果，ロボットは利用者にとって非常に使いづらいものになってしまう．従って，移動ロボット用の音声認識インタフェースを構築する上で不要音声の棄却は非常に重要な意味を持つ．本稿では，PaPeRo の音声認識インタフェースの特徴及び音声認識インタフェース構築の際に必要不可欠となる不要音声の棄却手法について述べる．

2 PaPeRo の音声認識インタフェースと不要音声棄却

2.1 PaPeRo の音声認識インタフェースの特徴

PaPeRo は，人を見分け個人適応した対話をする機能をもつ，一般家庭での使用を想定した自律移動型パーソナルロボットである．PaPeRo の外観を図 1 に，ハードの主なスペックを表 1 に示す．PaPeRo は，普段は視覚センサや超音波センサを利用し自律的に部屋内を走行している．そして，人間を見つけるか声をかけられると人間の方を向き，CCD カメラによる画像認識で利用者を識別し対話を開始する．利用者は音声対話によりロボットにダンスをさせたり他の人への伝言を残したりすることができる．また，PaPeRo は内部に自身の感情や利用者ごとの好感度や親密度を保持しており，利用者に適応した反応をすることができる．音声認識は離散単語音声認識を利用しており，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．

PaPeRo の音声認識インタフェースの特徴は，以下の 3 つである．

1. 不特定話者の音声認識
2. ハンズフリーな音声認識インタフェース
3. マイク間距離に融通性のあるディスタンスフリーな

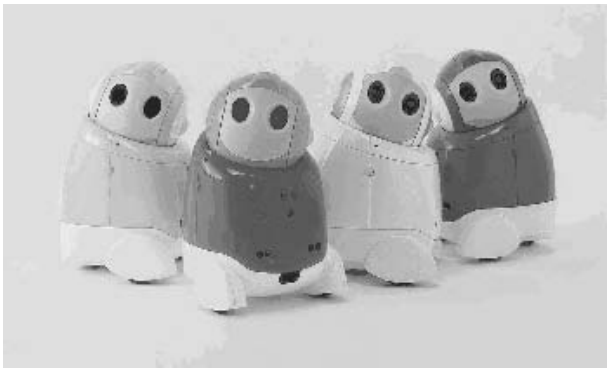


Figure 1: パーソナルロボット PaPeRo

Table 1: PaPeRo の主なスペック

サイズ (mm)	248(W) x 245(D) x 380(H)
重さ	5.0kg
稼働時間	2 時間
視覚センサ	CCD カメラ x 2
聴覚センサ	音源方向検出用マイク x 3 音源認識用マイク x 3
その他センサ	超音波センサ，段差センサ，持ち上げセンサ，頭センサ（たたかれ，なで検出）
その他	TV リモコン，画像・音声出力 I/F，インターネット接続

音声認識インタフェース

これらはいずれも利用者の身体的自由度を向上させるために必要と考えられる項目である．音声認識の話者適応に関しては，画像認識による話者照合の精度次第で話者学習させることも考えられるが，現時点では不特定話者の音声認識を対象としている．利用者とロボットのマイク間距離は，0.5〜2m で発話された音声が認識可能となることを想定しマイクの入力レベルを設定している．このマイク間距離は，最小値がロボット正面 0.5m 座姿勢の発話音声（近距離発声）を，最大値がロボット正面 1.5m 立姿勢の発話音声（遠距離発声）を想定したものである．また，マイクの指向性については，周囲雑音棄却の面からは正面からの発話にフォーカスするのが好ましい．しかしながら，PaPeRo は対話中に移動動作を行うため床のすべりなどの影響で移動後に正面を向かなくなる可能性がある．このことから正面以外の発話が完全に棄却されるのはあまり好ましくない．このため PaPeRo では側面や背面の入力レベルがやや低下する程度の指向性マイクを利用している．なお，音声認識には NEC 製の音声認識エンジン Ver4.1(SmartVoice Ver3.0 付属の音声認識エンジン) を使用している．

このように，PaPeRoでは利用者との距離に融通を持たせた音声認識インタフェースを有するため環境雑音や周囲会話を棄却する手段が必要不可欠となる．また，利用者が発話した未知発話も同様に棄却できることが望まれる．さらに環境雑音と未知発話を区別できれば，両者に対するロボットの反応を変える（例えば雑音は無視し未知発話に対しては「分からない」と答える）ことでロボットにより的確な反応をさせることが可能となる．そこで PaPeRo の音声認識インタフェースでは，利用者の発話した音声と周囲雑音を区別して棄却することを目標とした．以下，離散単語音声認識における雑音棄却の問題点と PaPeRo の雑音棄却手段について説明する．

2.2 離散単語音声認識における不要音声棄却

離散単語音声認識において不要音声の棄却を行う際には，正しく発話された音声の認識と不要音声の棄却という 2 つの要求を満たす必要がある．不要発話棄却の従来研究としては，音節的・言語的な尤度を利用した未知語検出の方法が研究されている[1][2][3]．一般的な音声認識エンジンでは，標準パターンとのマッチング距離や尤度の値が棄却判定の閾値を超えたか否かで棄却判定が行われる．PaPeRo のような不特定話者で距離に融通性を持たせた音声認識インタフェースを想定した場合，音声認識率を向上させるためこの棄却判定閾値を下げる必要がある．この概念を平面図で擬似的にイメージ化したものが図 2 である．音声認識語彙（「おはよう」「こんにちは」「こんばんは」）は「◆」で空間にマッピングされており，周囲の点線の円で囲われた音声認識空間を音声認識エンジンが正しい認識結果と解釈するものとする．なお，図中の「こんにちは」「こんばんは」のように円に重なりがある場合は，距離の近い方の音声認識語彙が認識結果となる．この点線円の半径が音声認識結果を返すための閾値であり，棄却判定閾値が低いほど半径が大きくなる．これに対し，円の内部にある実線枠で囲われた空間は正しく音声認識すべき発話音声の空間を意味する．PaPeRo の音声認識インタフェースでは，不特定話者かつ近距離発声から遠距離発声まで様々な距離からの発話音声が正しく認識すべき対象に含まれるため実線枠の空間が必然的に大きくなる．この実線枠の空間を吸収するためには閾値を下げ半径を大きくする必要があり，その結果として不要音声が点線円の内部に入りやすくなるため誤認識が多くなる．

これに対し，実環境での音声認識利用を想定した不要音声棄却手法としては，タスクを講演中の音声認識に限定し講演音声とコマンド発話音声の話し言葉の違いから講演音声を棄却しようとするもの[4]や顔画像認識を併用し話者がマイクの方角を向いているときのみ音声認識をするもの[5]が報告されている．しかし，これらはいずれも近距離発声を前提としており遠距離発声に適用した際の音声認識や周囲雑音の棄却に関しては未知である．これに対し，移動ロボット jijo-2 では，音源方向検出とマイ

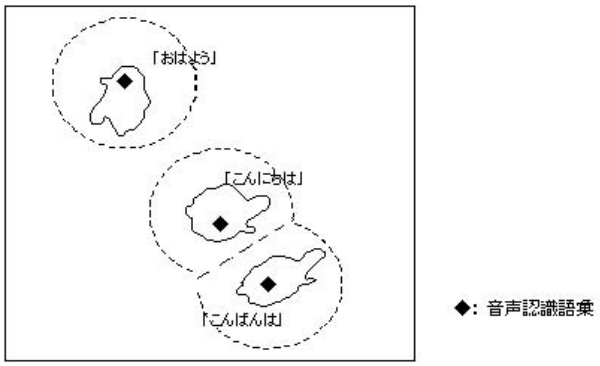


Figure 2: 音声認識空間のイメージ

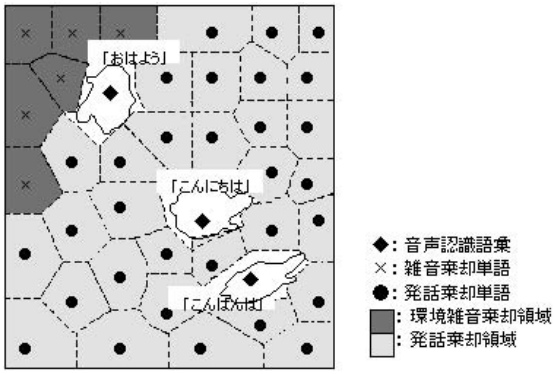


Figure 3: 棄却辞書構築後の音声認識空間イメージ

クフォンアレイを利用することで正面からの発話のみを受理し周囲雑音を棄却する試みがなされている[6]．しかしながら，正面指向性が強化された場合でも遠距離発声を前提とする場合は，正面から入力される雑音に関する対処といった問題が存在する．そこで PaPeRo では，想定される使用環境や音声認識語彙に適応した不要音声を棄却する音声認識辞書（以下棄却辞書と呼ぶ）を構築し音声認識語彙に追加することで不要発話の棄却を行うことを試みた．以下，棄却辞書による棄却性能向上の手法とその構築方法について述べる．

2.3 棄却辞書の概念と構築方法

棄却辞書は，音声認識空間に棄却専用の語彙（以下棄却単語と呼ぶ）を散布し不要発話を棄却単語に認識させ棄却するものである．棄却辞書構築にあたり，家庭環境における環境雑音を連続音声認識させる簡単な予備実験を行った．その結果，環境雑音が単音節や人間の発話し得ない語彙（「んんん」など）に認識しやすい傾向を得た．この結果から，我々は以下の 2 つの仮定を立てた．

仮定 1： 環境雑音と発話音声は区別可能である
仮定 2： 環境雑音は比較的短い音節数で吸収可能である

この仮定に基づく棄却辞書構築後の音声認識空間のイメージ図を図 3 に示す．棄却単語の散布により音声認識語彙の周囲以外の空間が棄却単語の音声認識範囲により覆われているというのが棄却辞書構築のイメージである．そして，音声認識空間が環境雑音が認識される領域とそれ以外の主に人間の発話音声が認識される領域に分かれており，環境雑音領域をカバーする雑音棄却単語（図中の×）からなる雑音棄却辞書とそれ以外の領域をカバーする発話棄却単語（図中の●）からなる発話棄却辞書を別々に構築することが可能であるというのが仮定 1 である．

次に雑音棄却辞書と発話棄却辞書の構築方法について説明する．棄却辞書構築システムの基本構成図を図 4 に示す．棄却辞書は，雑音棄却辞書と発話棄却辞書共に音声認識語彙とは無関係に構築された棄却単語データベースから音声認識語彙の認識に影響を及ぼす語彙を除去する

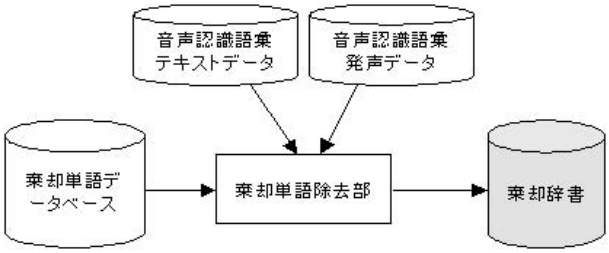


Figure 4: 棄却辞書構築システムの構成図

方法により構築される．棄却単語の除去には，音声認識語彙のテキストデータと音声認識語彙の発話音声を収録した音声認識語彙発声データを利用する．雑音棄却辞書と発話棄却辞書の違いは両者の棄却単語データベースが異なる点である．雑音棄却辞書の棄却単語データベースは想定される環境雑音を連続音声認識させ得られた語彙からなる．これに対し，発話棄却辞書の棄却単語データベースは基本的に音声認識空間全体が対象となるが，音声認識空間は無限大であるため音節数を限定した音声認識空間とする．また，仮定 2 に基づき，雑音棄却辞書の音節数は発話棄却辞書の音節数以下とする．

棄却単語除去は次の 3 つの処理過程にて行われる．

過程 1： 音声認識語彙にマッチする棄却単語の除去
過程 2： 音声認識語彙の類似語彙を利用した過程 1 と同様の棄却単語除去
過程 3： 複数話者の音声認識語彙発声データを利用した棄却単語除去

棄却単語の除去は，音声認識語彙の先頭から棄却単語の音節数分を切り出したものとのマッチングを行い除去するものとする．これを部分マッチによる棄却単語除去と呼ぶ．ただし，音節数の短い棄却単語に部分マッチの方法を適用すると棄却単語が消滅してしまう可能性があるため，棄却単語の音節数に応じて除去方法を切り替える．音節数の短い棄却単語は音声認識語彙に完全にマッチしているもののみを除去し，音節数の長いものは音声認識語彙の先頭から音節数分切り出した文字列と部分マッチしているものを除去するものとする．音節数の長短の境

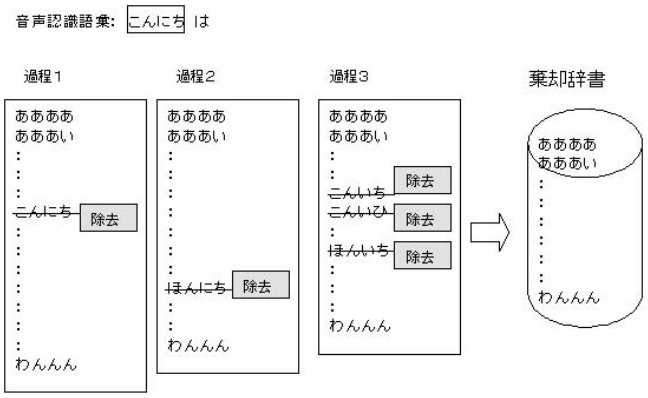


Figure 5: 棄却単語除去処理の流れ

目は音声認識語彙セットにも依存するが一般的には 3 か 4 以上の音節数を長い音節数と捉えるのが妥当と考えられる．PaPeRo では 4 音節以上の棄却単語を部分マッチによる除去の対象としている．

また，過程 3 で利用する音声認識語彙発声データは，棄却単語なしの状態で正しく認識される正解音声であることが必要不可欠である．音声認識語彙発声データの発話音声自体が誤認識するものであると，本来除去すべきではない棄却単語が除去されてしまうためである．

次に図 5 を元に棄却単語除去の個々の処理過程について説明する．図 5 は，音節数を 4 に固定した発話棄却辞書に対し音声認識語彙「こんにちは」を正しく認識する棄却辞書を構築する例である．棄却単語の音節数 4 は部分マッチの対象であるので,「こんにちは」の先頭 4 音節「こんにち」と部分マッチを行う．まず，過程 1 においては棄却単語「こんにち」が除去される．過程 2 の例は，k と h の子音が誤認識しやすいというルールが別途存在するか得られたと想定した場合の例である．この例では,「こんにちは」に対し k と h を入れ替えた類似語彙として「ほんにちは」が生成され，棄却単語「ほんにち」が除去される．最後に過程 3 では，正解音声を音声認識語彙と過程 1,2 を通過した棄却単語からなる音声認識辞書で認識させ，棄却単語に誤認識した場合にその棄却単語を除去した音声認識辞書を再構築し再び認識させる処理を繰り返す．図 5 の例は，棄却単語「こんいち」「こんいひ」「ほんいち」が誤認識された後正しい「こんにちは」が認識された例である．以上の 3 段階の過程を経て残った棄却単語が棄却辞書に登録される．

3 PaPeRo への棄却辞書の実装

ここでは，前節で説明した棄却辞書の PaPeRo への実装方法について説明する．

棄却辞書構築及び性能評価に用いる発声データとしては，0.5m 座姿勢の近距離発声データと 1.5m 立姿勢の遠距離発声データ（距離数値は話者と PaPeRo の水平距離）を利用した．これらの発声データは，近距離及び遠距離から PaPeRo と自然なインタラクションを行う上で想定される姿勢のデータであり，いずれもロボットの正面からの発声音声を収録したデータである．棄却辞書の実装にあたり，棄却辞書構築用に男性 2 名，女性 1 名の成人の優良話者による全認識語彙の発声データを実機にて収録し利用した．

また，棄却辞書構築に際しては，簡略化のため発話棄却辞書の音節数を 4 に，雑音棄却辞書の音節数を 4 以下に限定した．また，音節は濁音を除く単音節に限定した．同時に，音声認識エンジンの棄却判定閾値を近距離・遠距離発声データの認識率が劣化しない（音声認識率最大時との差 1 %以内）範囲の最大値に調整し，音声認識エンジンが判定する棄却を併用した．雑音棄却辞書は家庭環境の環境騒音を収録した音声を音節数を 4 以下に限定した連続音声認識を用いて音声認識させ約 600 単語からなる棄却辞書を構築した．

次に PaPeRo の棄却辞書構築における棄却単語除去方法について説明する．まず，テキストベースの棄却単語除去過程（過程 1 と過程 2）について述べる．PaPeRo の発話棄却辞書の棄却単語データベースは，音節数を 4 に限定しているものの棄却単語総数は音節数の 4 乗（濁音なしの単音節で約 500 万）と音声認識語彙約 650 に対し非常に膨大である．一方，音声認識語彙の発話音声を連続音声認識させ得られる音節列の精度は十分に高いとは言えず，特に PaPeRo の音声認識インタフェース特有の遠距離発声により精度はさらに低下する．以上のことから，発話棄却辞書に対するテキストベースの棄却単語除去はある程度まとまった単位で行うことにした．具体的には，4 音節の音節列を前 2 音節と後 2 音節に分けどちらかが音声認識語彙とマッチするものを除去する方針を取った．前述の「こんにちは」に対する棄却単語除去を例に挙げ説明すると，1,2 文字目が「こん」である棄却単語と，3,4 文字目が「にち」である棄却単語が全て除去される．ただし，雑音棄却辞書に関しては棄却単語数が少ないのでこの方法は取らなかった．また，過程 2 の類似語彙生成については，事前評価において誤認識傾向の強い音素を子音に限定して抽出したものを利用し類似語彙を生成した．過程 3 の正解発声データを用いた棄却単語除去に関しては，近距離発声データのみ前述した男性 2 名女性 1 名の正解発声データを利用し棄却単語を除去した．棄却単語除去を近距離発声データに限定した理由は，遠距離発声データに対する棄却単語除去の収束が芳しくなく大量の棄却単語が除去され棄却性能が劣化する現象が見られたためである．遠距離発声に対しては，あらかじめ収録したデータは使用せず近距離発声データの収録話者 1 名が直接発話し認識不良語彙を抽出，棄却単語除去を行った．

以上の手順により，約 450 単語の雑音棄却辞書と約 14000 単語からなる発話棄却辞書を構築した．以下，棄

却辞書の性能評価について示す.

4 棄却辞書の性能評価

4.1 評価方法

棄却辞書の性能評価には 3 つの指標が必要である．1 つ目は音声認識語彙を正しく発話した場合の音声認識率，2 つ目は音声認識語彙以外の語彙発話に対する棄却精度を示す発話棄却率，3 つ目は周囲雑音や環境雑音に対する棄却精度を示す雑音棄却率である．棄却辞書の性能評価は，棄却精度に加え棄却辞書追加による音声認識率の劣化を評価する必要がある．これら 3 つの指標に加え，発話棄却率の実験における発話棄却成功時の雑音棄却辞書と発話棄却辞書の棄却分類精度について評価を行った.

評価はいずれも認識語彙数約 650 の対話メインモードで行った．音声認識率及び発話棄却率の評価は，評価データとして収録した男性 23 名，女性 6 名（いずれも成人）の計 29 名分の近距離発声データ及び遠距離発声データを利用し行った．評価用の発声データの内訳は音声認識評価用約 150 単語，発話棄却評価用約 50 単語である．一方，雑音棄却の評価に関しては，テレビのニュース番組を 10 分間収録した音声をマイク距離 1m から再生し音声認識させた時の誤認識の頻度で棄却性能を評価した．評価対象としては，棄却辞書なしの状態で棄却判定閾値を上下させたものを利用し，音声認識と不要音声棄却のバランスを比較した.

4.2 結果

最初にマイクの設置・音量調整など物理的な音声認識インタフェースの評価を兼ねた音声認識エンジン単体の評価結果について示す．棄却辞書なしで音声認識エンジンの棄却判定閾値を上下させた時の音声認識率と発話棄却率の関係をグラフで示したものが図 6 である．左側が近距離のグラフで右側が遠距離のグラフである．音声認識率と発話棄却率はトレードオフの関係にあり，反比例のようなグラフとなる．グラフより発話棄却精度を高めた時の音声認識率の劣化は，近距離よりも遠距離の方が激しいことが分かる．なお，音声認識率の最大値は近距離 93.5 %，遠距離 78.9 %であった.

次に，棄却辞書の有無による音声認識率と発話棄却率の比較について述べる．棄却辞書ありの場合の音声認識率と発話棄却率を表 2 に示す．参考値として棄却辞書なしで音声認識率または発話棄却率が棄却辞書ありと同程度となる棄却判定閾値の時の値を併せて示す．棄却辞書を追加したことによる音声認識率の劣化は近距離で 3.5 %，遠距離で 8.2 %であった．また参考値との比較から，音声認識率をほぼ同程度にした場合は発話棄却率が約 25 %，発話棄却率をほぼ同程度にした場合は音声認識率が近距離で約 5 %，遠距離で約 12 %棄却辞書ありの方が勝っていることが分かる.

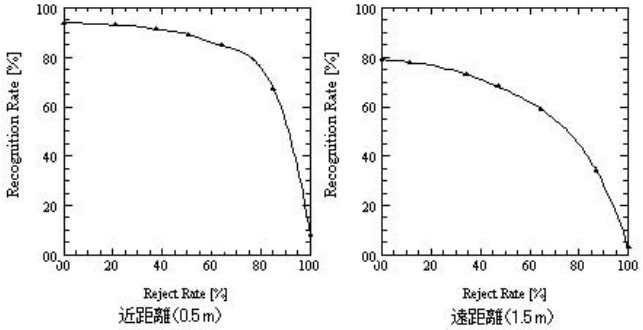


Figure 6: 棄却辞書なしの場合の音声認識率と発話棄却率の関係

Table 2: 棄却辞書ありの場合の音声認識率と発話棄却率の比較

	音声認識率（%）		発話棄却率（%）	
	近距離	遠距離	近距離	遠距離
棄却辞書あり	90.0	70.7	61.4	63.4
棄却辞書なし（音声認識率が同程度）	91.4	73.3	37.8	34.0
棄却辞書なし（発話棄却率が同程度）	84.8	59.0	63.9	64.6

次に，雑音棄却率について棄却辞書ありと棄却辞書なしの場合の比較結果を示す．表 3 は，音声認識率と雑音棄却の関係を棄却辞書ありと棄却辞書なしで比較したものである．棄却辞書なしの場合は棄却判定閾値を 4 段階に切り替え実験を行った．雑音棄却の指標には，雑音棄却率と 10 分間のニュースを再生中に誤認識をした回数を示す．棄却辞書ありの場合は 1 分に 1 回程度の誤認識で棄却率が 90 %を超えているのに対し，棄却辞書なしの場合は棄却判定閾値が最も低い棄却辞書 A が約 8 秒に一回，最も高い棄却辞書なし D でも約 14 秒に一回誤認識をしていることが分かる．音声認識率の比較で見ると，棄却辞書なし D は棄却辞書ありの場合に対し遠距離発声で約 35 %劣っているにも関わらず雑音棄却精度まで劣っていることが分かる.

最後に発話棄却成功時の棄却分類精度に関する評価結果を示す．PaPeRo の棄却辞書は雑音棄却辞書が 4 音節以

Table 3: 棄却辞書ありと棄却辞書なしの場合の雑音棄却性能比較

	音声認識率（%）		誤認識回数	雑音棄却率（%）
	近距離	遠距離		
棄却辞書あり	90.0	70.7	12	91.8
棄却辞書なし A	93.1	78.0	72	51.7
棄却辞書なし B	91.4	73.3	59	60.9
棄却辞書なし C	84.8	59.0	55	65.2
棄却辞書なし D	67.3	34.3	44	69.0

Table 4: 発話棄却成功時の棄却分類精度

発声	音節数	雑音棄却辞書 (%)	発話棄却辞書 (%)
近距離	3 以下	67.6	32.3
	4	54.3	45.7
	5 以上	13.9	86.1
遠距離	3 以下	85.8	14.2
	4	73.4	26.6
	5 以上	37.1	62.9

下，発話棄却辞書が 4 音節固定で構築されている．ここでは棄却分類精度を音節数が 3 以下，4，5 以上の音声認識語彙に分けそれぞれの棄却分類精度を調べた．表 4 に棄却分類精度の結果を示す．近距離発声について見ると，音節数 3 以下では雑音棄却辞書による棄却が多く，音節数 4 でほぼ同等に，音節数 5 以上では圧倒的に発話棄却辞書による棄却が圧倒的に勝ることが分かる．これに対し遠距離発声では，音節数 4 以下まで雑音棄却辞書による棄却が圧倒的に多く，音節数5 以上で発話棄却辞書による棄却が勝るようになることが分かる．以上のことから，棄却分類精度は近距離発声に関しては意図した通りの結果が得られているが，遠距離発声に関しては雑音棄却辞書の影響が強いことが分かる．

4.3 考察

今回構築した棄却辞書は，発話棄却，雑音棄却とも棄却辞書なしの場合に比べ音声認識と棄却のバランスの面で優れていることが分かった．特に雑音棄却に関しては優れた棄却性能を示している．発話棄却辞書に関しては，棄却単語数が除去前 500 万であったものが除去後に 14000 まで減少しており音声認識空間的にかなり疎になっていた可能性がある．これに関しては，今後棄却単語データベースの簡略化方法や棄却単語除去過程，特に類似語彙の生成方法を改良することで棄却精度を向上させることが可能と考えている．また，認識不良の音声認識語彙は多くの棄却単語を除去してしまう傾向が見られることから，認識不良語彙に対し類似語彙を音声認識辞書に追加することで音声認識語彙強化を行う方法も棄却精度を向上させる上で有効であると考えられる．特に遠距離発声は音声認識率が低く認識不良語彙も多いことから，音声認識語彙強化の併用が効果的と考えられる．

次に発話棄却時の棄却分類精度については，遠距離発声の棄却において雑音棄却辞書の影響が強くなるものの音節数 5 以上であれば 6 割以上の発話が発話棄却辞書で棄却可能であることが分かった．ここで一般的な誤発話の傾向について考えると，誤発話は音声認識語彙の未修得や言い間違い，言い淀みによるものが多くそのほとんどが 5 音節以上であると考えられる．従って，一般的な誤発話は雑音と誤認識されにくく，PaPeRoが人間の発話した音声を雑音と誤認識し無視する現象は起きにくくなっているものと考えられる．

5 おわりに

本稿では，パーソナルロボット PaPeRo の音声認識インタフェースの特徴及び必要不可欠となる不要音声棄却の手法について報告した．不要音声の棄却に関しては，棄却専用の音声認識辞書を音声認識語彙に適合させ構築することにより音声認識率を劣化させることなく効果的な不要音声棄却を行うことが可能であることが分かった．

今後の課題としては，考察の際にも述べたように遠距離発声への対応方法を改善し棄却辞書の性能向上を検討する予定である．さらには，棄却辞書構築過程の正解発声データによる棄却単語除去を行わず音声認識語彙から棄却辞書を自動構築する方法を検討する予定である．棄却辞書の自動構築が可能になれば，状況依存の音声認識語彙に適応した棄却辞書を構築でき棄却辞書の有用性がさらに高まると考えられる．

謝辞： 棄却辞書の構築方法，特に棄却単語データベースの構築方法について多くのご教示を頂いた，足利工業大学客員研究員の上山武明氏に感謝いたします．

参考文献

[1] A. Asadi, R. Schwarts, and J. Makhoul: Automatic detection of new words in a large vocabulary continuous speech recognition system, Proc. ICASSP'90, pp.125-128, (1990).

[2] 甲斐, 廣瀬, 中川: 単語 N-gram 言語モデルを用いた音声認識システムにおける未知語・冗長語の処理, 情報処理学会論文誌 Vol.40 No.4, pp.1383-1394, (1999).

[3] 渡辺, 塚田: 音声認識を用いたゆう度補正による未知発話のリジェクション, 信学論 D-II Vol. J75-D-II No.12, pp.2002-2009, (1992).

[4] 河原, 石塚, 堂下: 発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化, 情報処理学会論文誌 Vol.40 No.4, pp.1491-1498, (1999).

[5] 西本, 志田, 小林, 春山, 小林: 音声利用効果の経時変化と顔向認識による不要発話の棄却 - マルチモーダル作図システム S-tgif における評価 - , 信学技法 SP96-32, pp.89-96, (1996).

[6] 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)

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-0317-4

Auditory Processing for a Mobile Telerobot

Jie Huang
School of Computer Science and Engineering
The University of Aizu
Tsuruga, Ikkimachi, Aizu-Wakamatsu, 965-8580 Japan
E-mail: j-huang@u-aizu.ac.jp

Abstract

In this article, we described an autonomous mobile robot equipped with audition as well as vision and other sensors. The ultimate goal of this project is a multimodal telerobot, which can recognize and interact with the environments in multimodalities. The robot will also be linked to the Internet. Through the Internet, people can operate the robot and receive the remote auditory and visual scenes. With the first phase development, a prototype robot has been completed. It contains multimedia interfaces, and has the auditory functions of spatial sound localization and object tracking. Other auditory and visual functions such as sound source serapation, sound understanding, 3D sound presentation are now under development.

1 Introduction

The technology of mobile robot is an emerging field with wide applications. For example, a mobile robot can serve as a guard robot which can detect suspicious objects by audition and vision. The robot can also be used as an Internet-connected agent robot by which the user can explore a new place without being there. The robot can even attend a meeting instead of its users so that the users can get the remote auditory and visual scenes of the meeting room

For the above mentioned purposes, the robot must be capable of treating multimedia resources, especially sound media to complement with vision [1]. Visual sensor is one of the most popular sensors used today for mobile robots. Since a robot generally looks at the external world from a camera, difficulties occur when a object does not exist in the visual field of the camera or when the lighting condition is poor. A robot cannot detect a non-visual event which in many cases may, however, be accompanied by sound emissions. In these situations, the most useful information is given by audition. Audition is one of the most important senses used by humans and animals to recognize their environments. Although the spatial resolution of audition is relatively low compared with that of vision, the auditory system can complement and be cooperative with vision systems. For example, sound localization can enable the robot to direct its camera to a sound source.

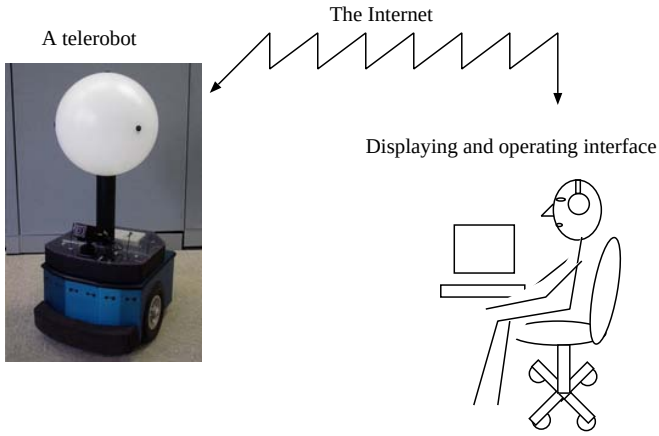


Figure 1: A multimodal interactive telerobot

Considering a guard robot patrolling around the campus of our university, we would need the robot to work in two different navigation modes. One is the autonomous mode, the robot will automatically move around the campus and check for possible suspicious objects. There were many researches concentrated on the autonomous aspect of mobile robots. The other one is the human-piloted (controlled) interactive mode. When the robot has found some suspicious objects, the robot will be better to be controlled by its owner who will order the robot to approach the objects and check more details about the objects. Even when the robot is in the controlled mode, some basic tasks, such as obstacle avoidance, will remain to function autonomously, because of that the Internet is not a perfect environement for real time operation, and a semi-autonomous mode can decrease the operation load of the users.

By integrating the robot with other parallelly developed telerobot technologies, the robot will also be linked to the virtual environments. So, the robot can be used as a actual mobile interface to the real would to provide the real visual and auditory scenes to the virtual environments. [2, 3].

2 System and Desired Auditory Functions

2.1 System

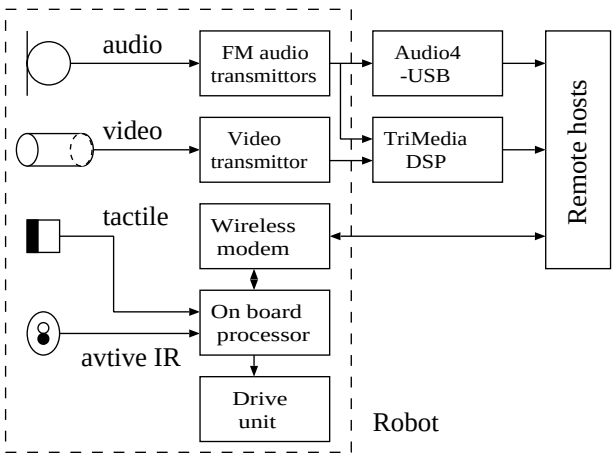


Figure 2: Diagram of signal flow

Base Platform: The robot is based on the LABO-3 platform of Applied AI Systems, Inc. The LABO-3 platform is a wheel-based mobile robot driven by two-wheel differential steering with zero turning radius. It has 10 active infrared sensors (6 in front, 2 on both sides and 2 in back) and tactile sensors (in front and rear bumpers) for obstacle avoidance, and a micro-processor for steering control and signal processing of built-in sensors. The battery can provide eight hours power supply when fully charged. The maximum payload is about 30 kg.

Auditory Sensors: Four microphones are arranged in the surface of a sphere, where three microphones (x,y,z) in a same horizontal plane with a shape of regular triangle, and the fourth microphone (o) arranged in the center with a height so that the four microphones (o-xyz) form Cartesian coordinates (i.e. ox-oy-oz in right angles). The diameter of the sphere is about 30 cm.

Vision Sensor: A fixed-directional CCD camera is mounted on the front of the robot, which is planned to be upgraded to an active camera system with a pan/tilt servo platform.

Communication: A radio bi-directional modem is used for host-robot communication. Four FM transmitters and one video transmitter are used for audio and video signal transmission respectively.

Preprocessing Units: Audio4-USB is a DSP-based data acquisition system with a Motorola DSP56307 (24bit fixed-point) processor. It has four channels audio input and output and is connected to the host PC via USB interface. TriMedia-TM1300 is a PCI plug-in DSP board which supports video and two channel audio input/output with a TriMedia TM1300 DSP processor (Philips) for digital video, audio and telecommunications processing.

2.2 Desired Functions

As shown in Figure 3, integration of multimodalities may greatly increase the capability and the flexibility of the environmental recognition. Some pilot developments and researches have been done for integrated object localization and tracking [4, 5]. However, since auditory study is behind that of vision, our purpose is to create a full set of auditory functions for robots, and hence to promote vision-audition integration as the next step.

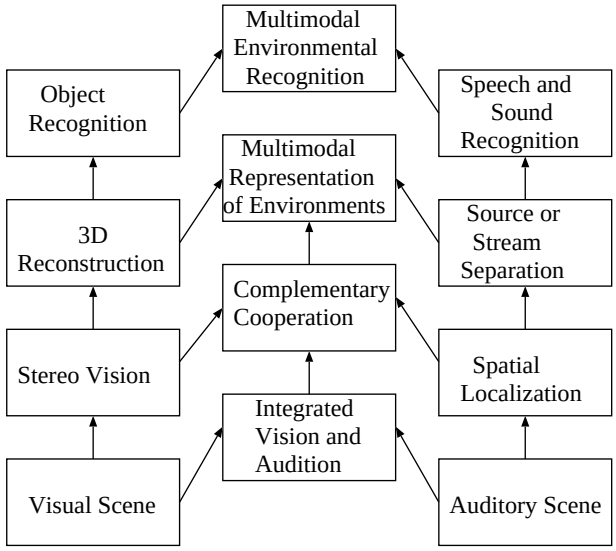


Figure 3: Multimodal integration

Spatial Sound Processing: Sound localization is an important function of the auditory system of a mobile robot and can also be used for a teleconference system to guide its camera to pick up the faces of speakers automatically [6, 7]. In such practical applications, sound localization is usually performed in reverberant environments. In the latter sections of this paper we will focus on the techniques of spatial sound processing and echo avoidance in reverberation environments [8, 9, 10].

3D Sound Presentation: There are two reasons for that we have to create binaural 3D sound. One is that the robot will be used as multimedia interace, for example to a teleconference system. The other one is that as described above, an autonomous mobile robot also needs to be piloted and controlled by humans to achieve practical tasks. In this case, the human operator will need to get 3D auditory scenes of the real environments where the robot is working. This requirement in some case can be satisfied by using a dummy head, a binaural microphone system to emulate the acoustics of the human head. However, the microphone system of the mobile robot has other important tasks such as sound localization and source separation. A simple head shape

like a sphere will simplify the localization and separation tasks and enhance the localization accuracy. In this case, we will need a preprocessing to convert the sound received by the robot head to binaural 3D sound which can be perceived by humans. In our robot, the microphone head is designed as a sphere with wider interval distance than that between two ears of humans. Since we do not used any structures like pinnas of human auditory systems, elevation cues need to be added by incoporating spatial cues created by all of the spatial arranged four microphones [3].

Blind Source Separation: Talking to other persons in a noisy room, walking in the street, the auditory system faces the problem of separating complex sound signals into different sound streams. In spite of the difficulties that the acoustic events produced by different sources may overlap in time and frequency domain, the human auditory system can identify and separate the sources without effort. What methods are used in the human auditory system?

As a mathematic model, the task of separating multiple sources is often described as of solving a multiple input and multiple output system by only knowing the output but without knowing the input and the system itself [11]. The principle of those methods is that the signals of different sources are usually mutually independent. One weakness of this method is the need of intensive computation power, especially when the positions of sound sources or the positions of receivers are time variant. This weakness, however, can be improved by incorporate with sound localization method to inform the system the real source positions initially.

Perceptual Auditory Scene Organization: Comparing to the above mentioned blind separation methods, the computational auditory scene analysis methods are based on the human auditory model of sound stream organization [12, 13, 14, 15]. The process can be described as the following steps. Firstly, the sound signals are coded and processed by a primary organization stage. This stage is considered as a bottom-up processing which is innate but not a learned effect. The time-frequency represented sound pieces will be organized and grouped into different substreams according to the so-called primary cues. After the first step of primary cue processing, there will be a middle stage, where the different substreams will be integrated into different sound events. Spatial cues will also be used for the integration [7, 16]. The sound substreams, then, will be processed by a high level stage which is based on the learned characteristics of different sound events or based on the sound recognition and understanding, the so-called schema-based organization. This stage is considered to be a top-down processing which will finally effect the outputs of the lower stage. Although those methods have achived some effects for automatic sound organization and separation, they are still not as flexible as the human auditory system. It is partially because of the discovered psychological rules are much qualitative rather than quantitative. In the future development, we will try to construct a quantitative computational model for the primary and middle sound organization stage of the auditory system [17].

Sound Understanding: Sound classification or understanding is important for an autonomous robot to recognize its environments. This is especially true for this project because we use the auditory system as one of the major part of the sensing system. Here, sound understanding does not include speech recognition which refer to the understanding of the meaning from a spoken language. Sound understanding will enables the robot to act in response to what is happening. We will restrict the target to some special sounds, e.g. sound of human voice, phone bell, door knocking, door open, walk step, siren, crash, and so on. The neural network technology will be used.

Auditory Navigation: Compared to vision, audition is all-directional. When a sound source emits energy the sound fills the air, and a sound receiver (microphone) receives the sound energy from all directions. Some specialized cameras can also receive an image from all directions, but still have to scan the total area to locate a specific object [18]. Audition mixes the signals into a one-dimensional time series, making it easier to locate any urgent or emergency situation. Audition requiring no illumination enables a robot to work in darkness or low light condition. Audition also is less effected by obstacles. So, a robot can perceive auditory information from sources behind obstacles [6]. One example is to localize a sound source outside of a room or around a corner. The robot will first localize the sound source in the area of the door or corner, and then travel to that point and listen again, finally have located the sound source. The ability of Localizing sound source can also be used for cancellation of positioning errors. Instead of visual landmarks, auditory landmarks will be used so that the robot can cancell the positioning error by localize the auditory landmarks.

3 Spatial Sound Processing

In this section, we describe the first phase development on the spatial sound processing technology for the multimodal mobile robot.

3.1 Sound localization cue processing

The auditory system used in the multimodal robot has some similar properties comparing to the human auditory system. The microphones are arranged in the surface of a sphere "head" and with interval distance 30 cm of about 1.5 times of that of humans. Spatial cues including the time difference and intensity difference cues are used for the multimodal mobile robot.

However, the multimodal mobile robot is not designed to simulate the human auditory system. It has some proper features which are based on the engineering needs of efficiency and accuracy. The sphere shaped head can simplify the formulation of time difference calculation. We use four microphones which form the Cartesian coordinates with the origin at the top of the sphere head. Different pairs of microphone will provide localization

cues for three orthogonal dimensions. By using the top-mounted microphone, we can localize the elevation of sound sources based on the time difference and intensity difference cues without the use of the relative uncertain spectral difference cue. To analysis the intensity difference cue of the sphere head, the head related transfer functions were measured for all azimuths and elevations of a 5 degree step.

3.2 Echo Avoidance for sound localization

One important problem of sound localization in real environments is how to cope with echoes. According to the EA model of the precedence effect (see Figure 4), the inhibition of sound localization depends on the sound-to-echo ratio [8, 9]. Suppose the sound intensity is $a(f,t)$ and the estimated echo is $a_e(f,t)$, (both $a(f,t)$ and $a_e(f,t)$ are amplitude envelop or short-term Fourier transform of signal in a narrow-subband f). Then the inhibition will correspond to the ratio $a(f,t)/a_e(f,t)$.

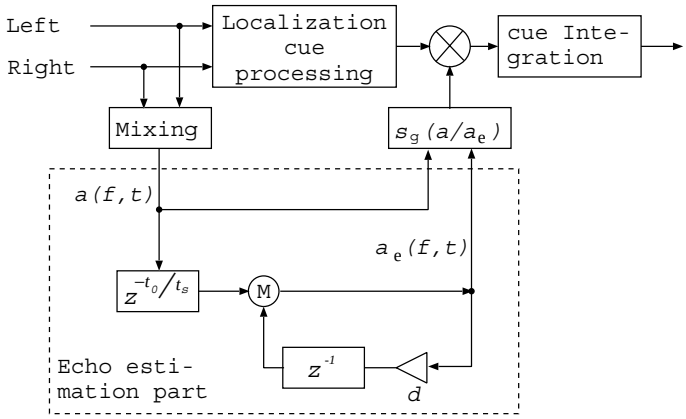


Figure 4: Echo Avoidance model of the precedence effect

Denote the impulse response from the sound source to a receiver (microphone or ear) as $h(f,t)$, and the part caused by echoes (removed the effect of the direct sound from the the impulse response) by $h_e(f,t)$. The echoes $a(f,t)$ can be estimated as

$$a_e(f,t) = h_e(f,t) * a(f,t). \quad (1)$$

Although the impulse response is unknown, we can give an generalized approximation which reflects the delay and decay features of the impulse response,

$$g_e = k\ e^{-(t-t_0)/\tau}. \quad (2)$$

Thus,

$$a_e(f,t) = M\{a(f,t-t')\ g_e(t')\} \quad \text{for } 0 < t' < \infty. \quad (3)$$

For easy treatment, the summation in the convolution is replaced by the maximum operation 'M'. The delay time t_0 and decay factor τ are chosen to match the most general cases in an ordinary environment. (In the human auditory system, it should be done by the learning effect [19, 20].) The decay factor τ, however, does not severely effect the result of echo estimation. It is because in the above approximation, the signal $a(f,t)$ contains not only direct sound but also the echoes. Thus, $a(f,t)$ itself has the decay feature which due to the echo portions. Because of this feature, the time decay factor can be much smaller (about 2 to 5 ms) than that of real environments.

By using the exponential decay feature of the generalized impulse response, the echo estimation algorithm can be implemented by a feed-back manner as shown in Figure 4, where t_s is the sampling time interval and $d = e^{-t_s/\tau}$ and sg is a sigmoid funtion. This algorithm is very fast with only two multiplication and one comparison operations to predict the echoes.

3.3 Integration of localization cues

Since the observed sound signal is usually not constant, but is continuously time variant. In different time, there will be different sound sources which mixed together to form a single one dimensional sound wave. Thus, a method is needed to integrate the localization cues over time and distinguish different sound sources.

3.3.1 Weighted cross correlation method

One traditional approach is the use of cross correlation and multi-sensor array beamforming methods [21, 22, 23, 24]. The cross correlation based methods are popularly used to calculate the time delay of two similar signals. However, those methods do not distinguish the direct sound and its reflections. To eliminate the influence of echoes, we will need a long time period of data for statistic average. Here, we propose an improved method, the weighted cross correlation method [25]. By this method (as shown in Figure 5), the signals are transfered into the time-frequency space, and weighted by the estimated sound-to-echo ratio. Finally, the signals are returned back to the time domain, and cross correlation is performed to the weighted signals.

Experiments of time delay estimation were performed in an ordinary room with walls , floor, and ceiling made of concrete. The area of the room is about 30 m^2. The room was empty when the experiments were conducted, so that the reverbaration in the room was very strong. The testing sound is radio weather forecast presented by a male announcer. Two microphones (number 1 and 3) were detached from the robot head, and placed with a interval distance of about 13.5 cm. Testing sound was played from a speaker positioned about 2.9 m from the center of the two microphones and with an angle of 60 degrees to the center line. In the Figure 6, conventional and weighted cross correlation for sound between microphone 1 and 3 are indicated by solid and dashed lines. A dotted line vertical line near the time point of about 0.333 ms indicates the real arrival time delay of testing sound. It is clear that while the conventional cross correlation was strongly influenced by the echoes and reverberations, the weighted cross correlation showed the correct time delay.

The disadvantage of cross correlation based method is the low spatial resolution, because of the gentle-slope-peaks. It is difficult to localize multiple sound sources

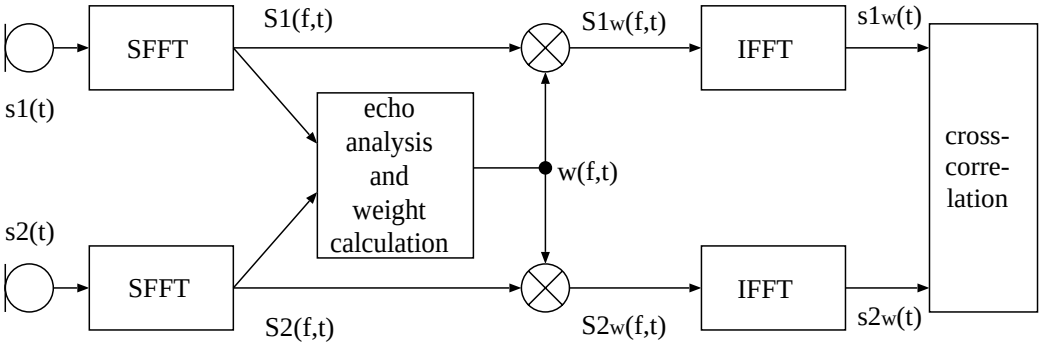


Figure 5: Weighted cross correlation method for time delay estimation

when the interval distance between microphones is small. However, this disadvantage can be overcome by enlarging the microphone interval distance.

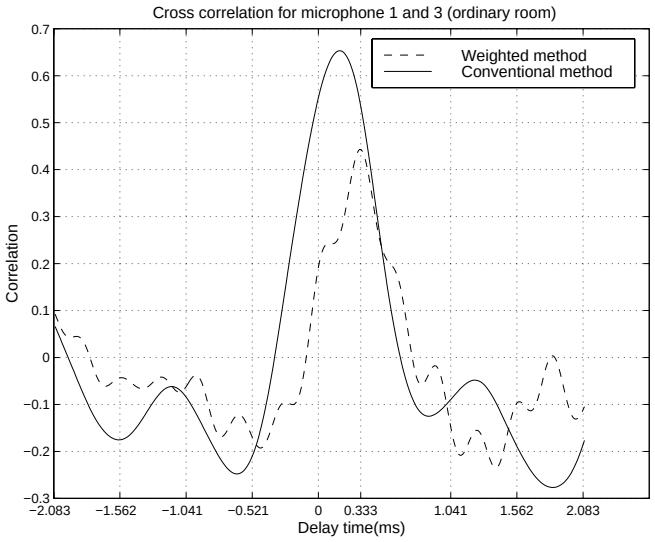


Figure 6: Comparison for conventional cross correlation and weighted cross correlation method

3.3.2 Time difference histogram method

Another approach is the use of a time difference histogram [9]. This method is much similar to the characteristic delay calculation discovered in the owl's auditory system [26]. There are possibly more than one time difference candidates from a value of phase difference

$$|\Delta t_c| = |\frac{1}{2\pi C f}(\Delta\omega_f + 2\pi n_c)| \leq \frac{D}{C} \qquad (4)$$

where n_c is an integer. To reduce the ambiguity, an azimuth histogram for sound source direction is used. Firstly, time difference histograms are formed for all microphone pairs by adding all of the possible candidates Δt_c of all frequency bands over a certain time segment. The time difference histograms of different microphone pairs are then mapped to the azimuth domain with a spread depends on the azimuth estimation sensitivity to the time difference $d\hat{\theta}/d(\Delta t)$. The histograms in azimuth domain are then integrated by the arithmetical average to form the single one azimuth histogram. by the final azimuth histogram, sound sources can be localized as the positions of peaks. Compared to the cross correlation based method, the histogram method has higher spatial resolution and is available for multi-source localization. Experiments showed this method could localize two simultaneous sound sources in an ordinary room with an accuracy of ± 2 degrees.

3.3.3 Intensity difference cue

In the above mentioned two methods, the intensity difference cue does not play any role. It is because in the previous version of robot, the microphones were mounted directly in the space without any kind of "head" between them. Sounds can directly go through the microphones without any shadow effect. The intensity difference is only caused by the distance difference and is trivial to be used as sound localization cue. In the new version of robot the microphones are arranged in the surface of a sphere "head". Due to the shadow effect of the sphere head, the intensity difference between different microphones became significant. The intensity difference can be used to help determination of the unique time difference from the phase difference in the above mentioned histogram method. Suppose the azimuth and elevation are estimated to be $\hat{\theta}$ and $\hat{\phi}$, and the measured intensity difference is d_i then it must satisfy the following equation

$$d_i = |H_x(\hat{\theta}, \hat{\phi}, f)| \qquad (5)$$

where H_x is the pre-measured interaural transfer function. This restriction is expected to reduce the redundancy and improve the localization accuracy.

4 Conclusion

In this article, we described a prototype design of a multimodal autonomous mobile telerobot. This robot is equipped with audition, vision and other sensors. Multimedia interface through the Internet is implemented. By its first phase development of spatial sound processing, different sound localization methods which are robust against reverberant environment were proposed and comparison between the different methods was made.

The important feature of this project is that we concentrate on the sound technologies as well as visual processing. It is because that interaction through sound media between robots and humans is indispensable. With the auditory functions, the robot will be aware when a person is talking and to understand where is the speaker. When dangerous accidents occur with a crash sound, the robot will be aware and direct the camera to the sound source.

Although it is difficult to trust robot with complex tasks at a full-time autonomous robot, we combined with the autonomous robot with a semi-autonomous mode, e.g., while the robot keeps some basic autonomous functions, the human users can also control the robot to complement its ower functions. This design will make it easier to apply mobile robots to practical applications.

References

[1] Jie Huang, "Spatial sound processing and a hearing robot," in *Proc. Int. Conf. Information Society in the 21st Century*, Aizu-Wakamatsu, Nov. 2000, U. Aizu, pp. 281–287.

[2] Michael Cohen, "A design for integrating the internet chair and a telerobot," in *Proc. Int. Conf. Information Society in the 21st Century*, Aizu-Wakamatsu, Nov. 2000, U. Aizu, pp. 276–280.

[3] William L. Martens, "Pseudophonic listening in reverberant environments: Implications for optimizing auditory display for the human user of a telerobotic listening system," in *Proc. Int. Conf. Information Society in the 21st Century*, Aizu-Wakamatsu, Nov. 2000, U. Aizu, pp. 269–275.

[4] H. G. Okuno, K. Nakadai, T. Lourens, and H. Kitano, "Sound and visual tracking for humanoid," in *Proc. Int. Conf. Information Society in the 21st Century*, Aizu-Wakamatsu, Nov. 2000, U. Aizu, pp. 254–261.

[5] P. Aarabi and S. Zaky, "Integrated vision and sound localization," in *Proc. 3rd Int. Conf. Information Fusion*, Paris, July 2000.

[6] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie, "A model based sound localization system and its application to robot navigation," *Robotics and Autonomous Systems*, vol. 27, no. 4, pp. 199–209, 1999.

[7] J. Huang, N. Ohnishi, and N. Sugie, "Building ears for robots: Sound localization and separation," *Artificial Life and Robotics*, vol. 1, no. 4, pp. 157–163, 1997.

[8] J. Huang, N. Ohnishi, X. Guo, and N. Sugie, "Echo avoidance in a computation model of the precedence effect," *Speech Communication*, vol. 27, no. 3-4, pp. 223–233, Apr. 1999.

[9] J. Huang, N. Ohnishi, and N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Trans. Instrum. and Meas.*, vol. 46, no. 4, pp. 842–846, Aug. 1997.

[10] J. Huang, N. Ohnishi, and N. Sugie, "Spatial localization of sound sources: Azimuth and elevation estimation," in *Proc. Instrum. Meas. Technol. Conf.*, St. Paul, May 1998, pp. 330–333, IEEE.

[11] M. Kawamoto, K. Matsuoka, and N. Ohnishi, "Realwould blind separation of non-stationary signals," in *Proc. Int. Conf. Information Society in the 21st Century*, Aizu-Wakamatsu, Nov. 2000, U. Aizu.

[12] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, London, 1990.

[13] M. Cooke, *Modeling Auditory Processing and Organisation*, Cambridge University Press, Cambridge, 1993.

[14] D. P. W. Ellis, "A computer model of psychoacoustic grouping rules," in *Proc. 12th Int. Conf. on Pattern Recognition*, 1994.

[15] T. Nakatani, H. G. Okuno, and T. Kawabata, "Multiagent based harmonic stream segregation for auditory scene analysis," *Journal Japanese Society for Artificial Intelligence*, vol. 10, no. 2, pp. 68–77, Mar. 1995.

[16] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3-4, pp. 209–222, 1999.

[17] K. Yoshida, "Interaction between different primary cues for sound integration and segregation," Graduation thesis, Univ. Aizu, 2001.

[18] Y. Nishizawa, Y. Yagi, and M. Yachida, "Generation of environmental map and estimation of free space for a mobile robot using omnidirectional image sensor COPIS," *J. Robotics Soc. Japan*, vol. 11, no. 6, pp. 868–874, 1993, (Japanese).

[19] K. Saberi and D. R. Perrott, "Lateralization thresholds obtained under conditions in which the precedence effect is assumed to operate," *J. Acoust. Soc. Am.*, vol. 87, pp. 1732–1737, 1990.

[20] R. K. Clifton, B. A. Morrongiello, and J. M. Dowd, "A developmental look at an auditory illusion: The precedence effect," *Dev. Psychobiol.*, vol. 17, pp. 519–536, 1984.

[21] Y. T. Chan, R. V. Hattin, and J. B. Plant, "The least squares estimation of time delay and its use in signal detection," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 3, pp. 217–222, 1978.

[22] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol. AP-34, no. 3, pp. 276–280, 1986.

[23] P. Stoica and K. C. Sharman, "Maximum likelihood method for direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-38, no. 7, pp. 1131–1143, 1990.

[24] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, PTR Prentice-Hall, NJ, 1993.

[25] T. Goto, "A weighted cross correlation method for sound localization in reverberant environments," Graduation thesis, Univ. Aizu, 2000.

[26] T. Takahashi and M. Konishi, "Selectivity for interaural time difference in the owl's midbrain," *J. Neuroscience*, vol. 6, no. 12, pp. 3413–3422, 1986.

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-0317-5

Entertainment Robot における音によるインタラクション

--- On the Interaction with Sound Signal by Entertainment Robot

藤田 雅博
Masahiro FUJITA
ソニー(株) デジタルクリーチャーズラボラトリー
Digital Creatures Laboratory, Sony Corp.

石井 和夫
Kazuo ISHII
ソニー(株) ER カンパニー
ERC, Sony Corp.

ABSTRACT

ペット型ロボットとの音を用いたインタラクションに関して、反射、情動、理性の3つの側面をロボットの行動制御アーキテクチャとインタラクションの種類の両方の点から議論する。具体的に、前記の3つの側面を考慮した行動制御アーキテクチャの実現と、口笛のような音階信号による実環境に強いインタラクションの実現、音声に含まれるプロソデックな情報を真似ることでインタラクションをさらに豊かにする試みに関して報告をする。

1. はじめに

我々は、ロボットエンターテインメントの 1 例としてペット型ロボットの開発を行っている。ペット型ロボットにおいて重要な用件は、ユーザーとのインタラクションである。ユーザーは、ロボットとインタラクションをすることでロボットに愛着を感じ、また愛着を感じるからインタラクションを行うようにポジティブなスパイラルを形成する。

自律型ロボットのインタラクションとしては、人間の手の動きなどによる指示や、道具を使ってのインタラクションがあるが、おそらく最も典型的に思い浮かぶインタラクションは言葉によるものであろう。例えば、ペット型ロボットが“犬”をメタファーとして作られたものであれば、“おすわり”、“お手”などの声による指示、あるいは“犬”が理解しているかどうかに関わらず、自然に語りかける、ということを行いたくなるであろう。

自然言語によるインタラクションは確かに理想としては存在するが、現在の音声認識、対話や談話処理の技術を考えると、まだ技術的課題が多く残っている。我々は、1997年にロボットエンターテインメントシステムの1例としてペット型ロボットの開発を報告しているが[1][2]、その時のデザインコンセプトは、言葉による不自然な対話よりも、口笛などを使った自然なインタラクションというものを重要視した。その根本的な考えは、人間が発生する信号には、いくつかの階層があり、“言葉の意味”といったレベルをいきなり用いるよりも、信号の存在や位置、信号に含まれる情動的情報、信号に含まれる意図、といった情報を用いてインタラクションを行うことの方が必要かつ現実的であるというものであった。

この論文では、このような見地から我々が開発したいくつかのロボットとのインタラクションのうち、特に音響信号に関わる部分に焦点を絞って議論する。

以下のこの論文では、まず1997年に開発を発表した4脚自律型ロボット MUTANT の音をもちいたインタラクションに関して簡単に説明する。この技術は第1世代の AIBO **ERS-110** においても用いられた。さらに、第2世代の AIBO **ERS-210** において開発された音声とそのプロソディを用いたインタラクションに関して説明する。

最後に、反射、情動の上に来る言葉の意味を、従来の対話のように辞書として予め与えるのではなく、自分の行動、情動的経験などを通して獲得する重要性を議論する。そこで、従来の物理接地記号に加え、情動接地という概念を導入しその実現例を紹介する。

2. MUTANT における音によるインタラクション

2.1 自律型ロボットにおける反射、情動、理性

人間の脳の構成自体が“反射脳”、“情動脳(辺縁系)”、“理性脳(新皮質)”から構成されている事実[3]を2つの側面から考えてみよう。1つは、ロボットの行動制御アーキテクチャに生かす、という面。もう1つは、人間の認知活動がこのような3つの階層から構成されているという面である。

歴史的なながれから考えれば、古典的 AI による熟考型行動制御アーキテクチャをもつ自律ロボットが現実の環境のダイナミックな変化に対応しきれず破綻をしており、それに対してサブサンプションアーキテクチャ[3]が提案された。この時点で、反射的な行動あるいは本能的な行動のモジュール構成が重要視されてきた。しかし、計画的な行動が難しいことから、それらを融合するアーキテクチャが提案されてきた(例えば、[5])。

我々が1997年に開発を発表した MUTANT[1][2]は、アーキテクチャとして反射行動と熟考行動の融合を図り、さらに本能情動といったものを自律ロボットに取り入れることによって、反射、情動、理性というものをロボットの行動制御アーキテクチャの側面から意識して作られたものである。ただし、インタラクションという面からみると、表出系は反射、情動、理性を考慮していたが、人間の情動の検出というものは意識的に行われてはいない。これは、後述の AIBO 第2世代において実現している。

2.2 音によるインタラクション

さてここで MUTANT の音に関するインタラクションに焦点を絞る。この試作機において最も重要視されたのは、実世界、実時間でインタラクションできるロボット、というコンセプトである。

従来考えられてきている“音声”によるインタラクションには、いくつかの問題が存在した。

1. 環境ノイズや自分の発生するモーターノイズによる音声認識の難しさ
2. デモンストレーション会場や TV などのさまざまな人の声がするような環境下での音声認識の難しさ

である。さらに、対話技術そのものである。音声認識に誤りを認めながら、ドメインを限らないような対話をすることはきわめて難しい問題である。

しかしながら、前述のような反射、情動、理性という3点に帰って考えるのであれば、音声によるインタラクションは、(1) 音の存在、発生の位置、(2) 音に含まれる情動、(3) 音に含まれる意図、というものをいかに検出し、それに反応するか、ということが重要であることがわかる。自然言語の音声認識が技術的に困難であるならば、自然言語以外の音でこのようなインタラクションを構成できないであろうか？ 我々は、口笛あるいは一般に音階信号によるインタラクションという結論にいたった。

2.3 音階信号によるインタラクション

2.3.1 信号の存在検出

口笛あるいは人間の声である一定の音程を発生することを想定する。前述した2つの音声認識の問題に対して、以下のようなアプローチにより解決が可能である。

1. 環境、モーターノイズ: ホワイトノイズのような時間－周波数軸上で構造を持たないで分布している信号に対して、口笛は、ある程度の長い時間、周波数軸に倍音構造で横線状の構造を持つ。時間－周波数空間において周波数軸方向の低域フィルタを施すことにより、口笛の検出が可能である。
2. 人の発話は通常ピッチおよびフォルマントが変化をする。ピッチやフォルマントが約 200msec 以上一定である会話の部分はほとんどない。口笛でこれ以上一定の音声を保って発生することで前述した時間－周波数空間における低域フィルタで口笛信号の抜き出しが可能である。

さらに、口笛などの自然言語を用いないインタラクションには別の意味が存在する。ヒューマンインタラクションにおいては、従来の人対人のコミュニケーションを期待させる自然音声によるコミュニケーションをロボットに行わせると、期待に反してその性能の低さがコミュニケーションとしての悪さを与えてしまう。逆に、事前に人に若干の不自然さを感じさせることにより、期待を低めに設定させ、それによるコミュニケーションを計るほうがより人に違和感を感じさせないインタラクションを構成できる。

このように、口笛には、ノイズ耐性を強化する、人の声を妨害音とした場合定常スペクトラムを利用して信号の抜き出しが容易、人の期待値がそれほど高くならない、といった利点がある。

2.3.2 信号の方向検出

前述したように、特定の構造を持つ音響信号を抜き出すことが可能であることから、ステレオマイクを用いて、まず、この信号だけに注目し抜き出す。それを時間－周波数空間において左右の信号の位相を比較することで、ノイズや周囲の人の会話に妨害されずに、口笛の方向を検出することが可能である。

2.3.3 信号による意図伝達

さらに、音程をもつ信号をアルペジオのように構成することで3音和音などを構成し、それに意味をつけることで、意図を伝達するようにした。例えば、“ドミソ”は、“座れ”、“ドファラ”は、“立て”という意味をつけた。3音同時に発生する通常の和音を用いたほうがノイズや妨害に強いことは電話の2音による DTMF 信号を考えても明らかであるが、そのような和音は人間が発生できない、ということからアルペジオのような構成とした。

2.3.4 実験

図 1に男性の音声と音階信号 (シンセサイザーによる) を同時に発生させた場合の時間－周波数空間の表現を示す。音声は、“おはようございます”という発話で、音階は“ファラド”という音階である。また、図 2にその信号を時間－周波数空間において時間軸方向に低域フィルタをかけた結果を示す。“おはようございます”という音声が低減ざれ、“ファラド”という音階信号が強調されていることがわかる。また、図 3にある時刻 (図 1、図 2で縦線で示した時刻、100フレーム目) における周波数特性を示す。左は、図 1に対応するもの、右が図 2に対応するものである。グラフよりわかるように、フィルタを通す前は、“ファ”に相当する 349Hz あたりのピークが音声の基本周波数の倍音にあた

る 500Hz あたりのピークより 15dB ほど小さな値であるが、フィルタを通したあとは、349Hz のピークが 500Hz のピークより 25dB ほど大きな値となっており、音階信号が簡単に抜き出せることを示している。

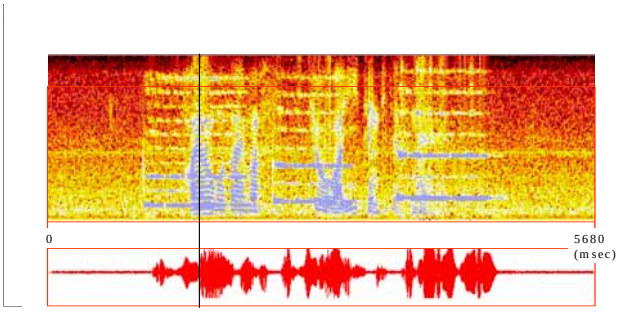


図 1 人の音声と音階信号の時間―周波数空間の表現（“おはようございます“と”“ファラド”という音“）

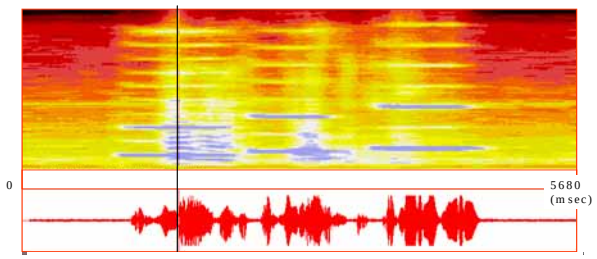


図 2 時間方向に低域フィルタを通した結果

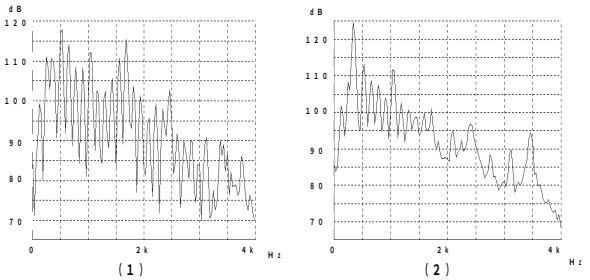


図 3 100 フレーム目の時刻における周波数特性

3. 第2世代 AIBO の音によるインタラクション

3.1 Vocal Communication

AIBO の第1世代のユーザーからの要望に、音声によるインタラクションを求める声は非常に多かった。これは、第1世代 AIBO が口笛によるインタラクションよりも、サウンドコマンダーという機器を用いたインタラクションの説明に重点をおいてしまっていることも大きな要因となっている。

しかし、そこで求められているものは、人間との対話のように音声を用いた知的な対話よりも、動物のペットに話しかけるように音声をメディアとしてロボットとインタラクションをしたいという情動的な側面があると思われる。われわれはペットロボットにおける音声インタラクションのデザインとして、連続音声認識を用いた音声対話システムよりもむしろ音声を介した"Vocal communication"によってお互いの理解を深めることを第一の目的と考えた。これは、MUTANTにおける口笛を用いたコミュニケーションを1歩進めた目標である。

人間のペットに対しての発話様式("Pet Directed Speech")は乳幼児に対しての発話("Infant Directed Speech")いわゆるマザリーズ(motherese)に近いことが知られている[7]。マザリーズは、乳幼児の知覚能力に最適な発話様式としてわれわれが通常行う話し方であり、高いピッチ周波数、大きなピッチ周波数の変化を特徴とする。近年、乳幼児の音声知覚能力は音声におけるプロソディの認識から音素の分節化に発達していくことが知られてきた[8]。われわれは学習と発達をベースにしたペットロボットをデザインして開発しており、このようなロボットにおいての音声インタラクションにおいて、人間が自然にペットや乳幼児に対して行うインタラクションのやり方に適合させることは非常に重要なことである。

われわれは、第2世代 AIBO において音声のプロソディを知覚しその情報から音を生成して応答するプロソディ分析合成システムとフレーズ音声認識とを融合して Entertainment Robot としての新しい形の Vocal communication を実現した。

3.2 音声対話との違い

従来の音声認識は音声におけるテキスト表現可能な情報を認識することを目標とし、近年は PC においてディクテーションソフトウエアとして実用化されている。音声対話システムもテキスト的情報をもとに対話を生成することになるため、知識や情報の伝え合いが音声対話の目標であった。

しかし、日常のコミュニケーションにおける音声の役割にはさまざまなものがある。われわれがある音声を聞いた場合、例えば以下のような情報を同時に得る。

- テキスト的な情報
- 音声に含まれる発話行為的情報（疑問、断定、依頼）
- 音声の個人性（個人識別、性別）
- 音声に含まれる感情や精神状態（怒り、疲れた）

また、複数の発話を対象に対話の側面から見ると発話のタイミングの中にも同意/非同意などの情報を読み取ることが可能である。

われわれはこれらの情報を伝え合いながらコミュニケーションをしており、日常のあいさつやたわいもないお喋りなどは，テキスト情報の伝達以外の情動的な側面のコミュニケーションが行われているといえる。また、対話の目的に対しても対話をすること自体が目的でたのしいという側面があり、Entertainment Robot としては情報伝達の正確さよりも対話というインタラクションの場を与えることが必要な役割だとかんがえられる[9]。

Entertainment Robot での音声を用いたコミュニケーションを考えた場合、

1. 音声を媒介としたコミュニケーション自体の楽しみ
2. 発話を理解してくれたという反応から得られる共感
3. 情動的な発話が行える状況

が必要と考えられる。

ペットロボットを考えた場合、それはペットや乳幼児に対するのと同じインタラクションを求められる対象となる。このような発話に対する処理は従来の音声対話ではとらえられない。

そこで、これらのことから音声のプロソディを積極的に用いた対話を生成することとした。

3.3 プロソディを用いる対話

通常の音声認識では、

- 音声パワーの正規化
- 音声ピッチ情報の除去
- DTW などによる発話の時間軸の正規化
- 統計的手法やマルチテンプレートによる認識

を行うことで、発話が変動してもその中にあるテキスト的情報を認識することが目標とされてきた。そのため音声分析部においては音声パワーとピッチ情報は捨てられることが多い。音声パワーとピッチの時間的変化はプロソディを構成する主たる物理量である。プロソディは音声における話者の意図や感情などを含む情報であり、従来の音声対話システムでは切り捨てられたタイプのコミュニケーションを実現する可能性がある。

3.4 模倣による共感

プロソディ分析合成音をもちいた非分節音においても模倣によるコミュニケーションの好感度が向上することが報告されている[10]。まねをするロボットに対して、理解してくれた、理解しようとしている、学習しているなどの志向性を人間は感じ、自分のことばを真似て復唱してくれるるロボットに対して肯定的な感じを受ける。また、応答のタイミングも重要なので処理の遅延なしに模倣的な音声を生成することが必要である。

3.5 プロソディの分析合成

音声のプロソディを分析し、その情報から音声を生成して音素的には非分節化された音声のやり取りとしてのインタラクションを提供する。音声のプロソディ情報として今回使用するのは、音声パワーとピッチの時系列である。通常 AIBO が発声する電子音との親和性をかんがみて、分析されたピッチ周波数の6倍程度の周波数の正弦波を合成することにした。正弦波の振幅は分析された音声のパワーに基づいて計算される。

図 4 には、男性の「おはよう」と言う音声とそれを真似て分析合成した合成音の波形とスペクトルをしめす。

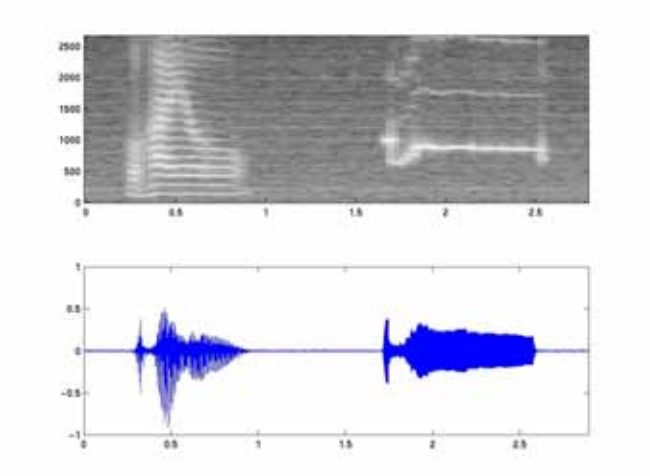


図 4 OHAYOU に対する合成音の波形とスペクトル

3.6 音声認識との融合

3.6.1 音声認識とプロソディ分析合成

AIBO ではプロソディ分析合成と音声認識を融合して、従来の音声認識だけのシステムにはなかった Entertainment Robot としての音声インタラクションを実現した。プロソディ合成は、ユーザーの音声を真似て発話する「物まねモード」と「名前登録」でつけられた名前を発話するという 2 つの機能として使用する。

以下に各機能について説明する。

3.6.2 音声認識

AIBO に搭載した音声認識は、50 ほどのフレーズを認識する不特定話者認識である。「AIBO」、「お手」、「おはよう」などのフレーズを認識し、認識した状況や成長度合いに応じて反応が変化する。

3.6.3 名前登録

名前をつけることは Entertainment Robot とそのオーナーとの個人的関係を高める。不特定話者音声認識に加えて新規単語の登録機能も装備した。「名前登録」というフレーズを認識すると、AIBO は音声登録状態になりユーザー

のタッチセンサ入力を待つ。タッチセンサ入力後の音声を登録音声とみなす。登録音声のスコアが既存フレーズのスコアと近い場合は、コンフュージョンを起こすので登録を却下する。登録が OK の場合は、前述のプロソディ音によって復唱することでユーザーに知らせる。名前が登録された後、「お名前は」というフレーズを認識すると、AIBO はプロソディ音によって自分につけられた名前を発話する。

3.6.4 物まねモード

AIBO の通常の動作は、音声認識された結果やセンサ情報に対して行動選択モジュールが行動を選択して運動や発音をおこなう。「物まねモード」においては、通常の行動に加えて、音声が入力されると入力音声のプロソディ分析をした後、反射的に合成音を生成する。反射的に合成音を生成するためにリフレクティブパスから直接音声発音部に発音データを送り、音声認識処理と行動選択モジュールにおける処理の遅延に依存せずに反射的にプロソディ合成音を発音することが出来る。

物まねによるインタラクションは、人間にとってみるとロボットが情動的な部分を抜き出して理解しているように感じる。音声の中の“意味“の部分をあえて捨て去り、情動的な部分でのインタラクションが成立する場面は、ペットや幼児に対して話しかけているのと同じようである。

3.6.5 システム構成

ステレオマイクから音声は入力される。入力された音声にたいして音声分析部では、音声区間検出用特徴量、音声認識用特徴量、プロソディ分析の計算をする。音声区間として判定された音声に対しては、音声認識モジュールにおいて、音声認識のスコア計算が行われ、認識されたフレーズは行動選択モジュールに出力される。行動選択モジュールは、確率的オートマトンであり、入力センサ情報に対して状態の遷移が起こり、同時にその遷移ノードに対応付けられている行動が出力する。

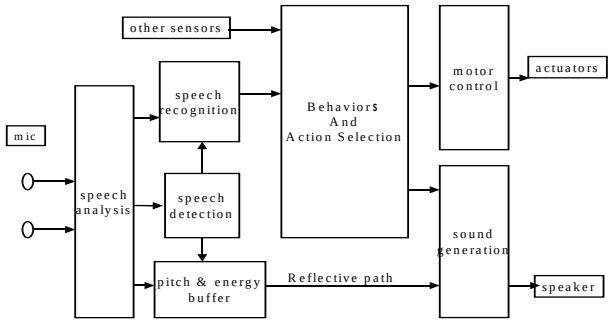


図 5 第 2 世代 AIBO の音のインタラクションに関するシステム構成

物まねモードにおいてはリフレクティブパスによる反射応答が生成される。2 パス構成によって行動選択モジュールによる通常反応と反射的な反応が融合した音声インタラクションを実現している。

4. Emotionally Grounded Symbol Acquisition

音によるインタラクションとして反射、情動、理性のうち反射と情動に関する部分の重要性を特に説明してきた。理性に関する部分に関しては、簡単なコマンドの理解というレベルにとどめて説明をしてきた。ここで最後に“意味”の獲得に関して若干の説明を加える。

自然言語処理に関しては、音声対話を含めて盛んに研究がなされている。しかしながら、予め辞書に存在しない単語をいかに獲得するか、ということは常に変化する Open な実世界で動作する自律型ロボットにおいて非常に大切な事柄となる。すでに、実ロボットにおいて言語獲得、特に物理接地の問題と関連付けてのシンボル獲得の報告がなされている[11][12]。

我々は、さらにこれを自律型行動の中に“情報を食べる”行動として定義し、好奇心という内部欲求に関連したホメオスタシス行動として統合した[13]。また、従来から提案されている物理接地は複数の知覚チャネル（例えば視覚と音韻）の連想記憶によるものであり、シンボルの意味はその知覚チャンネルにより接地させた物理世界での表現としての意味であった。しかし、これでは、自律型ロボットがこのシンボルに対してどのような行動を行えば良いかを決定することができない。そこで、物理接地に加え、そのシンボルに対しておこした行動から得られる内部状態の変化(空腹が満たされるとか、好奇心が満たされるとか)も連想させることにより、その意味を与えることを行った。この場合、そのシンボルを観察しただけで、かつて体験した情動的な経験を想起させ、その状況においてその物体に適切な行動をおこす事も可能となる。これを情動接地記号と呼び、物理接地と情動接地の両者が物体に対する意味として重要であることを主張している。

物理接地シンボルに関しても、[11]で行われていた既に音声認識部で登録されている音韻系列と視覚処理の結果のカテゴリを連想させるのではなく、新しい音韻を獲得する、ということで真の新しいシンボル獲得を実現している。用いた音声認識部は前述の第2世代 AIBO と同じ物で、これを対話の中で用いている。以下が対話の例である。

1. H(uman)： 指で物体を指しながら、“これなんだ？”
2. R(obot)：“知らない、なんていうの”
3. H：“りんご”
4. R：“りんごなの？”
5. H：“そうだよ”
6. R：“わかった。りんごだね”

このようにして、“りんご”という音韻系列を獲得し、それを視覚処理のカテゴリー（赤で丸い）と連想記憶させる。この状態では、りんごに対してどのような行動をするのかが不明であるが、好奇心がましてくると、りんごに対して、“食べる”、“蹴る”といった動作を試し、食べることによって内部状態の満腹感が増加することを記憶する。それ以降、りんごを見ただけで、それが満腹感を増す物体であり、それを食べれば満腹感がまし、快刺激が得られ、喜びを感じられる、ということが想起される。もし適度に食欲があれば、りんごを食べる、という行動が選択される。

前述の対話の中で、“指で物体を指して”という行動でいわゆる共同注意を達成している。すなわち、ロボットと人間が共通のものに注意をはらって行動あるいは対話をすることを可能にしている。これらは、ロボット側からなされることも可能であり、人間とのインタラクションをより容易にするものである。

5. まとめ

自律型ロボットと人間のインタラクションとして、人間の脳が反射、情動、理性の3つの側面を持っていることを考慮して、自律型ロボットの行動制御アーキテクチャと、この 3 つを考慮したインタラクションの重要性に関して議論した。特に、音響信号に関しては、音の存在、方向、意図の認識および反射、情動、理性的なインタラクションの実現を試みた MUTANT の紹介と、同じく人間の音声のプロソディ要素の物まねによるインタラクションの実現例を第 2 世代 AIBO を用いて紹介した。また、情動接地シンボルという概念を提案し、新しい知識を獲得する方法に関しても簡単に報告した。

6. REFERENCES

[1] M. Fujita and K.Kageyama, Open Architecture for Robot Entertainment, Proc. of Autonomous Agents pp.435—440 (1997)

[2] M. Fujita and H. Kitano, Development of Autonomous Quadruped Robot for Entertainment, Autonomous Robots, Vol.5, pp8—18, Klawer Publisher (1998)

[3] P.D. MacLean, Primate Brain Evolution, Method and Concept (eds. E. Armstrong and D. Falk), pp291—317, Plenum Press, (1982)

[4] R. A. Brooks, A Robust Layered Control System for a Mobile Robot, IEEE Transactions on Robotics and Automations, Vol.RA-2,pp14—23, (1986)

[5] R..J. Firby, Task Networks for Control Continuous Process, Proc of the 2nd International Conference on AI Plannning, (1994)

[6] R.W.Picard, Affective Computing, The MIT Press, 1997

[7] Denis Burnham, et.al..,"Are You My Little Pussy-Cat? Acoustic, Phonetic And Affective Qualities Of Infant- And Pet- Directed Speech", ICSLP'98,pp.453--456 (1998)

[8] 正高信男，「子どもはことばをからだで覚える」，中公新書 (2001)

[9] 鈴木紀子，竹内勇剛，石井和夫，岡田美智男：“状況に引き出された発話による対話の形成とその心理的評価”，情報処理学会論文誌，Vol. 40, No.4，pp.1453--1463 (1999)

[10] 鈴木紀子，竹内勇剛，石井和夫，岡田美智男：“非分節音による反響的な模倣とその心理的影響”，情報処理学会論文誌，Vol. 41，No.5，pp.1328--1338 (2000)

[11] F. Kaplan, Talking AIBO: First experimentation of verbal interactions with an autonomous four-legged robot. In proceedings of the CELE-Twente workshop on interacting agents, October, 2000

[12] Roy, D. and Pentland A. Learning words from natural audio-visual input, in proceedings of International Conference on Spoken Language Processing, 1998

[13] M.Fujita, G. Costa, R. Hasegawa, T.Takagi, J.Yokono, and H.Shimomura, Architecture and Preliminary Experimental Result for Emotionally Grounded Symbol Acquisition, Proc. of the 5th Autonomous Agents, pp.35—36, (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-0317-6

視聴覚情報の階層的統合による実時間アクティブ人物追跡

Real-Time Active Tracking by Hierarchical Integration of Audition and Vision

中臺 一博 [1], **日台 健一** [1], **奥乃 博** [1,2], **北野 宏明** [1,3]

Kazuhiro Nakadai[1], **Ken-ichi Hidai**[1], **Hiroshi G. Okuno**[1,2] **and Hiroaki Kitano**[1,3]

[1] 科学技術振興事業団 ERATO 北野共生システムプロジェクト

[2] 京都大学大学院情報学研究科, [3] ソニーコンピュータサイエンス研究所

[1]**Kitano Symbiotic Systems Project, JST,** [2]**Kyoto Univ.,** [3]**Sony CSL**

{nakadai, hidai, okuno, kitano}@symbio.jst.go.jp

Abstract

We developed real-time human tracking system for humanoid robots by integrating various modalities obtained from sensors. We use sound direction, speaker ID, face location, face ID, object location by stereo vision and motor direction for modalities, which are extracted from binaural microphones, stereo cameras and a potentio meter with a motor. The modalities make a stream by taking their time series into account. The stream belongs to a name or a location stream layer according to abstract level of its modality. When several streams are close, they are associated in a name or a location stream layer. In addition, the associations can occur between stream layers. The status of streams influences on "focus-of-attention" control of robot. As a result, we achieve robust human tracking even when two persons speaks simultaneously.

1 はじめに

昨今, HONDA ASHIMO, Sony AIBO, SDR-3X といったロボットが相次いで登場するなど, 研究のみならず世間一般にロボットが脚光を浴びている. これらのロボットはロバストな二足歩行を行ったり, 簡単な, しかし, 時には愛らしい仕草を可能にしたが, 知覚や認識の面からは, 研究課題は山積みされているのが現状である. このようなロボットには, 状況に即した行動を実現するために, 様々なセンサ情報を適切に扱って自律的に周囲の状況を把握することが求められる. しかし実際には, 周囲の状況をロバストに把握し, その状況を基準に行動を選択することは難しく, 現時点では, "人間のパートナー" として振舞えるように人間とソーシャルインタラクションを行うことは困難である.

ロボットが人間とソーシャルインタラクションを行うための最低限の要件は, 人間と同様のセンサを持つことであろう. このようなセンサとして視覚は一般的であるが, 聴覚は人間では主要なセンサであるにもかかわらず, これまであまり積極的に使用されてこなかった. これは, 聴覚のセンサ情報としての扱いにくさが原因となっている. 一般に聴覚処理は正確であるとは言いがたい. また, 単一音源からの音を収音することは難しく, たとえ指向性マイクを使用しても, 一般に複数の音源からの音が混入してしまう. さらに聴覚情報は部屋の反響など環境の変化に非常に敏感であり, 方向同定や音声認識にしばしば悪影響を及ぼすためである. 一方, 聴覚機能を備えたロボットの研究も複数存在しており, MIT AI Lab の *Kismet* [Breazeal and Scassellati, 1999] や, 早稲田大学の *Hadaly* [Matsusaka *et al.*, 1999] などが挙げられる. 両者とも音声認識を行い, 後者はさらにマイクロホンアレイを用いて音源定位を行う事ができる. しかし, マイクロホンアレイは動かないことが前提となっている. また, 両者とも音声認識は各話者の口元に取り付けられたマイクを利用しなければならないという制約があり, そのため音源分離能力も備えていない.

こうした状況の中, 我々はソーシャルインタラクションを行う能力を備えたロボットを目指して顔認識とアクティブオーディションを利用し、実時間に人物を追跡するシステムを開発し、これをヒューマノイド *SIG* に実装した[中臺 *et al.*, 2001]. このシステムでは, 従来あまり取り組まれてこなかった聴覚処理に注目しているとともに, センサ情報には曖昧性が含まれているという前提に立ち, 聴覚以外のセンサ情報も積極的に使用し, 情景分析におけるストリームとそのアソシエーションという概念を導入した情報統合を行っている. その結果, 複数の音源が混在する状況下であっても, 情報統合により, ロバストに人物追跡が行えるばかりか, オクルージョンや視野が狭い場合など, 視

覚情報が使用できない場合にも聴覚情報を用いて情報を補うことでロバストな処理を実現した.

しかしながら, 従来のシステムでは, アソシエーションは音と顔の方向情報のみで行っていたため, 情報統合の効果は, 限られた状況に特定されていた. そこで, 本稿では, よりロバストで適用範囲の広いシステムを構築するために, 話者情報およびステレオビジョンを用いた正確な位置情報を新たな統合情報として既存システムに追加した. これにより, 複数の情報の一部が欠けていてもロバストな人物追跡を可能にするとともに, より人間の知覚に近い階層的な情報統合を実現することにより, 人物抽出の精度を向上させることができた.

以下, 2 章では, 本稿で使用したヒューマノイドについて述べ, 3 章では, 本システムの概要, 各モジュールの詳細について説明する. 4 章でシステムの実験と評価を行い, 5 章でまとめる.

2 ヒューマノイド *SIG*

Figure 1: Humanoid *SIG*

研究のテストベッドとして, Fig. 1 の上半身ヒューマノイド *SIG* を使用している. FRP 製の外装は, 音響的にロボットの内外を区別できるよう設計されており, カメラには, 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, 外界からの音響信号を収音するよう *SIG* の耳の位置に一対, また主にモータ動作によって発生する内部ノイズを収音するよう外装の内部に一対ずつ配置されている. これを用い, 自己の発する音を認識し, キャンセルすることによって, アクティブオーディションを実現している [Nakadai *et al.*, 2000]. また, *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, 速度の制御が可能な DC モータを用いている.

3 システム概要

システムの各モジュールの構成を Fig. 2 に示す. システムは, “音源分離・定位”, “顔抽出・認識”, “話者同定”, “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” の大きく 8 つのモジュールから構成されている. 実装上は, 8 つのモジュールを 100Base-TX の LAN で接続された 3 台の Pentium III ベースの Linux ノードに分散させている. 各モジュールは複数のサブモジュールから構成され, モジュールの内部およびモジュール間では様々なレベルの情報の通信が非同期に発生する.

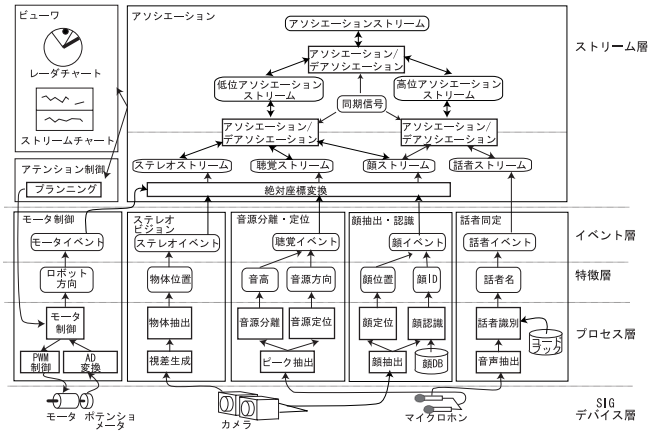


Figure 2: システムのモジュールとその階層構造

モジュール内のサブモジュールや情報は, 下位層から順番に, “*SIG* デバイス層”, “プロセス層”, “特徴層”, “イベント層”, “ストリーム層” の 5 つの階層に分けられている. *SIG* デバイス層は, *SIG* が備えているカメラ, マイク, モータシステムなどのセンサデバイスを指す. これらのセンサから得られたローレベルデータがプロセス層へ入力され, 位置, 名前情報といった特徴として特徴層に出力される. 各特徴は, 抽出されたタイミングでモジュール単位にまとめられ, 発生時刻を付与されたイベントという形でイベント層に出力される. イベント発生のタイミングは, 非同期であり, 各モジュールごとに異なる. ストリーム層では, イベントを各種類ごとに時間方向に接続し, ストリームを形成する. ここで, ストリームは自身を形成するイベントの抽象度に応じて, 高位と低位のストリームレイヤに分けられる. さらに, ストリーム間の距離に応じて, 複数のストリームを束ねて, アソシエーションストリームを生成する. 次節では処理の詳細を説明する.

3.1 音源分離・定位モジュール

音源分離・定位モジュールでは, 入力信号は, 異なる方向からの混合音を仮定しており, 48 KHz, 16 ビットでサンプリングされる. その後, 高速フーリエ変換 (FFT) による周波数解析を行い, 左右のチャネル毎にスペクトルを生成し, そのスペクトルを入力として, 音源分離および定位の処理が行われる.

ピーク抽出と音源分離: 入力スペクトルのうち, 閾値以上のパワーを持ったローカルピークを抽出する. 閾値は周波数毎に異なり, 部屋の暗騒音を計測することによって求められる. この際, バンドパスフィルタを用い, ノイズの大きい 90 Hz 以下の低周波域とパワーの小さい 3 KHz 以上の高周波域を計算量低減のためカットしている.

次に, 周波数が低いものから順番に, 抽出したローカルピークを取り出し, その周波数 $F0$ と 6% 以内の誤差で整数倍とみなせる周波数 Fn を持つローカルピークを, $F0$

の倍音としてクラスタリングを行う. このクラスタリングによって集められた最終的なピークの集合を一つの音とみなすことによって, 音源分離を実現している.

音源定位: 一般に, 両耳聴における音源定位には, 頭部伝達関数 (HRTF) から求められる両耳間位相差 (IPD) と両耳間強度差 (IID) が使用される. しかし, HRTF は頭部の形状や環境に大きく依存し, 環境が変わる都度, 計測が必要であるため実環境アプリケーションには不向きである. 筆者らは, HRTF に依らない IPD を利用した音源定位法として, ステレオ視におけるエピポーラ幾何の概念を聴覚に拡張した聴覚エピポーラ幾何に基づく方法を適用している [Nakadai *et al.*, 2000]. この際, (1) 音の倍音構造の利用, (2) IPD を用いた聴覚エピポーラ幾何による定位結果と IID を用いた定位結果の Dempster-Shafer 理論を用いた統合, (3) モータ動作中でも正確な音源定位を可能とするアクティブオーディションの導入によって, 音源定位のロバスト性を向上させている.

音源定位は音源分離によって分離された調波構造を有した各音に対して行う. *SIG* では, 左右のマイクのベースラインから 1.5 KHz 以下の周波数域に対しては IPD, それ以上の周波数域では IID による音源定位が 有効である. このため, 入力音のうち 1.5 KHz 以上の倍音成分と 1.5 KHz 以下の倍音成分の 2 つに分けて処理を行う. まず, 入力音のうち 1.5KHz 以下の周波数 f_k をもった各倍音成分に対して, 聴覚エピポーラ幾何を使用して, *SIG* 正面に対して $\pm 90°$ の範囲で $5°$ おきに IPD 仮説 $(P_h(\theta, f_k))$ を生成する. 次に, Eq. (1) に示す距離関数より, 入力の各倍音における IPD $(P_s(f_k))$ と各仮説間の距離 $(d(\theta))$ を計算する. ここで, $n_{f<1.5KHz}$ は周波数が 1.5 KHz 以下である倍音数である.

$$d(\theta) = \frac{1}{n_{f<1.5\mathrm{KHz}}} \sum_{k=0}^{n_{f<1.5KHz}-1} \frac{\left(P_h(\theta, f_k) - P_s(f_k)\right)^2}{f_k} \tag{1}$$

得られた距離に対し, Eq. (2) によって定義される確率密度関数を適用し, 距離を IPD を用いた場合の音源方向を支持する確信度 BF_{IPD} へ変換する. ここで, m と s は, それぞれ, $d(\theta)$ の平均と分散であり, n は d の個数である.

$$BF_{\mathrm{IPD}}(\theta) = \int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{\frac{s}{n}}}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \tag{2}$$

入力音のうち 1.5 KHz 以上の周波数をもった倍音に関しては, IID の総和の正負に応じて, Table 1 に示す値を IID を用いた場合の音源方向を支持する確信度 BF_{IID} として与える.

IPD, IID それぞれの処理によって得られた音源方向を支持する値から, これらを Eq. (3) で示される Dempster-Shafer 理論によって統合し, IPD と IID の両方から音源方向を支持する新しい確信度を生成する.

Table 1: IID の確信度, $BF_{\mathrm{IDD}}(\theta)$

	θ	$90° \sim 35°$	$30° \sim -30°$	$-35° \sim -90°$
I	+	0.35	0.5	0.65
	-	0.65	0.5	0.35

$$\begin{aligned} BF_{\mathrm{IPD+IID}}(\theta) &= BF_{\mathrm{IPD}}(\theta) BF_{\mathrm{IID}}(\theta) + \\ &\left(1 - BF_{\mathrm{IPD}}(\theta)\right) BF_{\mathrm{IID}}(\theta) + BF_{\mathrm{IPD}}(\theta)\left(1 - BF_{\mathrm{IID}}(\theta)\right) \end{aligned} \tag{3}$$

最終的に音源分離・定位モジュールは, 分離した音ごとに, 音高情報, 確信度付き音源方向 (確信度の高い順に上位 20 位まで) および観測時刻からなる聴覚イベントを生成する.

3.2 顔抽出・認識モジュール

顔抽出・認識モジュールは, 顔抽出と抽出した画像から人物同定を行い, 顔イベントを生成する. この際, 追跡時の環境変化に対応するため, 顔の位置, 大きさ, 明るさが動的に変化するような条件下におけるロバストな顔抽出, および実時間追跡のための高速処理が要件となる.

顔の抽出: 顔の抽出は, 肌色検出と相関演算に基づくパターンマッチングの組合せにより行っている. MMX 命令の使用により, 顔が複数存在している場合でも, 200 ミリ秒周期の顔領域検出を実現している [Hidai *et al.*, 2000].

顔領域を抽出した際に画像平面上の顔位置 (x, y) を, Eq. 4 を用いて距離 r, 方位角 θ, 仰角 ϕ として 3 次元の実空間上に変換し, 顔の位置同定を行っている.

$$r = \frac{C_1}{w},\ \theta = \sin^{-1}\left(\frac{x - \frac{X}{2}}{C_2\ r}\right),\ \phi = \sin^{-1}\left(\frac{\frac{Y}{2} - y}{C_2\ r}\right) \tag{4}$$

ここで, 画像平面上の大きさを $w \times w$ ピクセル, 探索画像の大きさを $X \times Y$ とし, C_1, C_2 は, 探索画像サイズ X, Y, カメラの画角, 実際の顔の大きさによって定義される定数である.

顔の認識: 顔の認識では, 抽出された顔領域画像を判別空間に射影し, 事前に登録された顔データとの距離 d を求める. 距離 d は, 登録顔の数 (L) に依存するので, Eq. (5) によってパラメータに依存しない確信度 P_v に変換している.

$$P_v = \Gamma\left(\frac{1}{2}, \frac{d^2}{2}\right) = \int_{\frac{d^2}{2}}^{\infty} e^{-t}\, t^{\frac{L}{2}-1} dt \tag{5}$$

判別空間の基底となる判別行列は, オンライン LDA によって求める [Hiraoka *et al.*, 2000]. オンライン LDA では, 通常の LDA と比べ少ない計算で判別行列の更新が可能であり, リアルタイム処理が可能である. 最終的に顔抽出・認識モジュールは, 各顔毎に, 上位 5 つの確信度付きの顔 ID(名前) と位置 (距離, 方位角, 仰角) からなる顔イベントを生成する.

3.3 話者同定モジュール

話者識別は, ベクトル量子化 (VQ) 歪みに基づいた方法による話者識別が可能な Juno [秋田 *et al.*, 2000] をベースにして, ストリーム処理が可能なように改変したものを使用している. この手法では, コードブックと呼ばれる事前登録した話者データを使用する. コードブックは登録話者数を S, コードブックのクラスタ数を C とした時, $B(i,j)(i=1,...,S, j=1,...,C)$ と表される. 話者識別は距離関数 Eq. 6 を用いて行う. 入力音声を T 個のフレームに分割し, ベクトル列 $v(k)(k=1,...,T)$ に変換した後, フレーム毎に各話者との量子化歪 (ユークリッド距離) をすべてのクラスタに対して計算し, その中で最も小さくなる量子化歪みを出力する. この計算を全フレームにわたって行い, その累積和 $S(i)$ を話者 i との距離としている.

$$S(i) = \sum_k Min_j |v(k) - B(i,j)| (k=1,...,T) \quad (6)$$

すべてのクラスからの距離が等しいとき, 各クラスに属する確信度が 50% であるとして, $S(i)$ を Eq. 7 によって確信度に変換する.

$$B(i) = \frac{1}{\bar{S}\sqrt{2\pi}} e^{\frac{-S(i)^2}{2\bar{S}^2}} \quad (7)$$

$\bar{S}$ は $S(i)$ の平均である. 話者同定モジュールは, 現時点では複数話者には対応していないが, 尤度の高い話者順に確信度を付与してアソシエーションモジュールへ出力する.

3.4 ステレオビジョンモジュール

ステレオビジョンモジュールは, ステレオ視による視差画像から人物らしい物体を抽出し, その正確な 3 次元位置を得る. 具体的には, 左右のカメラの視差画像の生成, 視差画像からの物体抽出, 物体定位, ステレオイベント生成の順に処理が行われる.

視差画像は, 局所領域のマッチングによる対応点探索に基づいて生成される. この際, PC 上で実時間処理を達成するため, 再帰相関演算手法と Intel アーキテクチャ固有の最適化 [岡田 *et al.*, 2000] を用いている. また, 事前にアフィン変換を用いた補正を施している. 視差画像からの物体抽出は, 人体は縦長であることを利用して, 細かいノイズに左右されない人体およびそれに類する形状・大きさを持った物体の抽出を実現している. つまり, 2 次元の視差画像に対し, 視差値の縦軸方向のメディアンを横軸に沿って求めていくことによって, 視差画像を 1 次元化し, その 1 次元視差画像に対して視差の近い領域を分割することで, 物体の抽出を行う. 抽出した物体はエピポーラ幾何により定位を行い, 最終的に, 距離, 方位角, 物体幅および, 観測時刻からなるステレオイベントを生成する.

3.5 アソシエーションモジュール

アソシエーションモジュールは, *SIG* がロバストに周りの状況を把握するために, 様々なイベント情報を統合し, ストリーム, およびアソシエーションストリームを生成する. ストリームはイベントを時間方向に接続することによって生成され, アソシエーションストリームは, ストリーム間の状態によって発生するアソシエーションによって生成される高次のストリームである. 聴覚情報と顔情報のアソシエーションを行う場合の流れを Fig. 3 に示す.

イベントの絶対座標変換: 話者イベント以外のイベントは位置情報を含んでいるが, この位置情報はイベントが観測された時刻にロボットから見た座標系 (*SIG* 座標系) における情報である. そこで, モータイベントを利用してこれらのイベントを絶対座標へ変換する. この際, 各イベントは遅延時間, 到着周期が異なる非同期イベントであるため (Table 2 参照), 各イベントは 一旦, 2 秒間の短期記憶

Table 2: 各イベントの到着時間

視覚イベント	200 ミリ秒
聴覚イベント	40 ミリ秒
モータイベント	100 ミリ秒
ネットワークレイテンシ	10 〜 200 ミリ秒

に格納され, 同期がとられる. 短期記憶に格納されたイベントは, ネットワークレイテンシおよび処理時間を考慮して, 実際の観測時刻と比較し 500 ミリ秒の遅延をもつように取り出される. 取り出された聴覚, 顔, ステレオイベントが発生した時刻のロボットの方向をモータイベントの一次線形補間によって推定し, 各イベントの位置情報を絶対座標系に 変換する (Fig. 3(a) 参照).

ストリームの生成, 消滅: 絶対座標系に変換されたイベントは, Fig. 3(b) に示されるように, ストリームに接続される. ストリーム接続のアルゴリズムを以下に示す.

- **聴覚イベント:** 音高が, 同等もしくは倍音関係にあり, 方向が ±10° 以内で最も近い 既存の聴覚ストリームに接続される. この値は, 聴覚エピポーラ幾何の精度を考慮し定められた値である.
- **顔イベント:** 顔イベントは, 共通の顔 ID をもち, 40 cm の範囲内で最も近い既存の顔ストリームに接続される. この値は, 秒速 4m 以上で人間が移動しないことを前提にして定めている.
- **話者イベント:** 話者イベントは複数話者に対応していないため, 同時刻には高々 1 つのイベントしか発生しない. 従って, 話者ストリームが存在していれば無条件に話者イベントを接続する.
- **ステレオイベント:** ステレオイベントは, 40 cm の範囲内で最も近い既存のステレオストリームに接続される. この値は, 顔イベントと同様の基準である.

すべての既存ストリームに対して探索を行い, その結果, 接続可能なストリームが存在しないイベントが存在した

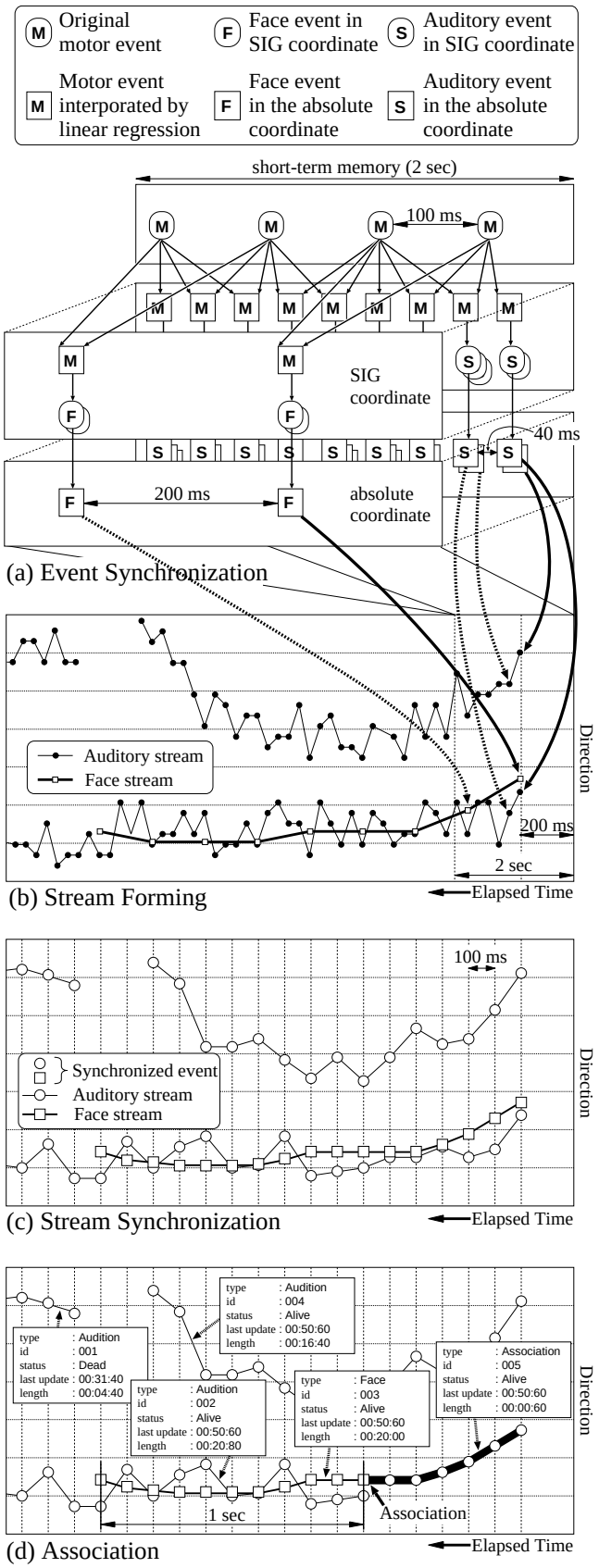


Figure 3: アソシエーションモジュールにおけるストリームの形成

場合, そのようなイベントから新しいストリームが生成される. ただし, 顔イベントに関しては, 既に存在している顔ストリームの ID と同じ ID を持ったストリームは生成せず, 第二候補以降の ID でストリームを生成する. また, すでに存在しているストリームは, 接続するイベントが全く存在しない場合でも, 最大 500 ミリ秒間は存続することができる. 500 ミリ秒以上全くイベントが接続されない状態が続いた場合, そのストリームは消滅する.

このような時間の流れを考慮したストリーム形成の利点は以下の通りである.

- 音：基音の取得失敗が訂正されうる
- 顔：名前が異なってもイベント間のユークリッド距離が近い場合は同一ストリームと見なせるため, 名前の誤り訂正がされる. またストリーム全体にわたって名前情報をチェックすることにより, ストリームを代表する名前の誤りを訂正できる. 例えば, ストリーム生成時に ID が間違っていても, 時間の経過とともに訂正されるうる.
- 話者：話者は同一ストリームであれば処理の性質上後からやってくるイベントの方が確信度が高くなる. このため, ストリーム生成時に間違っていた話者名が, 時間の経過に伴い訂正されてより正確になる.
- ステレオ：物体 (人物) の動きを把握できるようになる (アソシエーション時に有効).

アソシエーション: 複数のストリームが同一の人物に対するストリームであると判断された場合, Fig 3(d) で示されるようにこれらのストリームはアソシエーションされ, より高次のストリーム表現であるアソシエーションストリームを形成する. また, アソシエーションストリームを形成するストリームが消滅した場合, もしくは, 同一人物に由来するストリームであると判断されなくなった場合, アソシエーションストリームはデアソシエーションされ, 複数のストリームもしくはアソシエーションストリームに分割される. アソシエーションの際, ストリーム間の距離を算出するため, 前処理として, Fig. 2 に図示される同期信号を用いて 100 ms 単位で Fig 3(c) に示されるようにストリームを同期させる. また, ストリームは構成するイベントの抽象度に応じてストリームを高位ストリーム, 低位ストリームの 2 レイヤに大別される. 前者は, 名前情報 (話者, 顔) によるストリーム, 後者は位置情報 (音, 顔, ステレオ) によるストリームを意味し, 顔ストリームに関しては高位, 低位の両方の情報を含んでいるため, 両方に属すものとする. このような階層的なストリーム構造を構築した後に低位ストリーム間, 高位ストリーム間, および低位と高位ストリームにまたがった階層的な 3 段階のアソシエーションを行う.

低位ストリームのアソシエーション: 低位ストリームは位置情報を含んでいるため, 一定時間以上距離が近いスト

リームが存在した場合アソシエーションを行う. それぞれの低位ストリームは Table 3 に含まれる位置情報を含んでいる. しかし同じ情報であっても抽出方法が異なるため, その精度が異なる. 方位角はすべてのストリームに含まれるため, 距離の算出の際に有効であるが, その精度はステレオストリームに含まれるものが最も高く, ついで顔ストリーム, 聴覚ストリームとなっている. 聴覚ストリームは, 視覚ベースの位置情報と比較して精度が低いものの複数候補を確信度付きで持っているため, 複数の候補が利用できる. また, 距離 (r) の抽出精度は, ステレオストリームの方が顔ストリームよりも高い. このため, 以下のような基準でアソシエーションの判断を行っている.

聴覚 - 顔 方位角が $\pm 10^\circ$ 以内に近接する状況が 1 秒間のうち 500 ミリ秒以上観測される

顔 - ステレオ ユークリッド距離で 10cm 以内に近接する状況が 1 秒間のうち 500 ミリ秒以上観測される

ステレオ - 聴覚 方位角が $\pm 10^\circ$ 以内に近接する状況が 1 秒間のうち 500 ミリ秒以上観測される

精度の異なる同じ種類の情報を複数利用するため, 抽出精度に応じて距離関数のチューニングが必要である. このため, 一番精度の高い情報だけ利用すれば, 一見, 簡易であるし, 処理上も問題ないように考えられるが, 実際には, 情報を互いに補い合うことで, そのロバスト性を高めることができる. 例えば, 視野外にいる人からの声, 物陰に隠れている人からの声は, 視覚ベースの方法では定位することができない. また, 顔モジュールでは壁にかけてある写真を, 人物として抽出してしまうなど, 精度の高い方法であっても誤抽出があり, その際のエラー訂正の情報として, 他の同じ種類の情報が有効に利用できる. さらに, 方位角しかわからない聴覚ストリームであっても, 顔ストリームとアソシエーションできれば, 正確な方位角が取得できるばかりか, 距離, 仰角といった情報も取得することができる. このように低位ストリームのアソシエーションには, 大きな利点がある.

Table 3: 低位のストリームに含まれる位置情報

	聴覚	顔	ステレオ
距離 (r)		○	○
方位角 (θ)	○	○	○
仰角 (ϕ)		○	

高位のストリームアソシエーション: 高位ストリームには, 顔ストリーム, 話者ストリームが該当し, 名前情報を含んでいる. アソシエーションは両ストリームの話者名と顔 ID が一致した場合に行われる. 高位のアソシエーションにより, 一方の情報が欠けている場合でも名前情報を継続的に保持することができる.

低位と高位ストリームのアソシエーション: 低位と高位のストリーム間についても視覚ベースの位置情報と名前情報, および聴覚ベースの位置情報と名前情報についてアソシエーションを行う. 前者については, 顔ストリームが, 低位と高位の両方のストリームに属しているため, 顔ストリーム内において, 既に達成されている. これは, 顔ストリームでは, 顔の位置と名前は常に同時に, 顔抽出・認識モジュールで抽出されるためである. 後者については, 現時点では話者ストリームが複数話者に対応していないため, 以下のルールに基づいてアソシエーションを行う.

1. 聴覚ストリームが一つである場合はそのストリームと話者ストリームがアソシエートとする.
2. 聴覚ストリームが複数ある場合は, 話者ストリームの開始時刻と近い方のストリームをアソシエートする.

アソシエーションモジュールでは, 矛盾を防ぐ手段として, 以下のような制約を用いている.

- アソシエーションされていないストリーム同士は, 種類が異なる場合のみアソシエーション可能である.
- アソシエーションされたストリームは, そのアソシエーションストリームに属しているストリームと同じ種類のストリームを含んでいない任意のストリーム, および, アソシエーションストリームとアソシエーション可能である.

上述の制約を用いても, 矛盾が発生する場合, 基本的にそのような矛盾のあるアソシエーションは行わない. ただし, そのような状況が, 一定時間以上継続した場合には, アソシエーションの誤り, もしくは, ストリーム生成の誤りが生じている可能性が考えられる. 前者については, 最尤のアソシエーション状態になるようにアソシエーションストリームの再構築を行うことにより対処を行っている. 後者については今後の課題としている最終的に, アソシエーションモジュールは, 把握している状況をストリームおよびそのアソシエーション状態として保持し, アテンション制御およびビューワモジュールの問い合わせに応じて, ストリームの情報を送信する.

3.6 アテンション制御モジュール

アテンション制御モジュールでは, アソシエーションモジュールのストリームの状態に応じて *SIG* の行動を決定し, モータ制御モジュールへモータイベントを送出する. これにより, ビジュアルサーボや聴覚のみによるサーボと比較し様々な状況にロバストなトラッキングを達成することができる視聴覚サーボを達成している. 視聴覚サーボは, 状況に応じて複雑な制御も可能である. 現状では, 以下の 2 原則に基づいた制御を行っている.

1. アソシエーションストリームの存在は, *SIG* に対し正対して喋っている人が現在も存在している, もしくは近い過去に存在していたことを示している. したがって, 一般にそのような人間に対して, 高い優先度でアテンションを向け, トラッキングを行うのは妥当である.

2. マイクは無指向であるためカメラのような視野角は存在せず, 広範囲な情報を得ることができる反面, 情報の精度が低い. 情報の精度を高くするために, 聴覚ベースのストリームは視覚ベースのストリームより優先度を高くすべきである.

具体的には, アソシエーションされていない聴覚ストリーム, アソシエーションストリーム, 視覚ベースのストリームの追跡という優先順位でアテンションを制御する.

4 実験と評価

本システムの評価には, Fig. 4 に示すシナリオをベンチマークとして使用した. このシナリオでは, A, B の 2 話者が約 20 秒に渡って以下に示す様々なアクションを行う. なお, t_n は Fig. 4 における時刻を示すものとする.

t_1: *SIG* の視野外で A が話を始める. これにより聴覚ストリームが作られ, *SIG* は音の方向へ体を向ける.

t_2: A の声により話者ストリームが生成され, 聴覚ストリームとアソシエーションする.

t_3: A も *SIG* の正面に向かって移動し, A 氏が *SIG* の視界に入ったことを契機に視覚的に A 氏が検出され, 顔ストリームとステレオストリームが作られる.

t_4: 顔およびステレオストリームがアソシエーションする.

t_5: A に関するすべてのストリームがアソシエーションする. *SIG* はこのアソシエーションストリームに注意を向け追跡を続ける.

t_6: A をトラッキング中に *SIG* の視野外から B が話を始める. 同時に聴覚および話者ストリームが生成され, アソシエーションされる. B はそのストリームを確かめるために音のする方に向く.

t_7: *SIG* が B の方向に向いたため, A が視野からはずれ, デアソシエーションされる.

t_8: B が視界に入ったため顔とステレオストリームが生成される. 同時に B は短時間, 話を中断する.

t_9: 顔とステレオストリームがアソシエーションされる.

t_{10}: B に関するすべてのストリームがアソシエーションされ, *SIG* は B の追跡を続ける.

Fig. 4 より, t_5 から t_7, および t_{10} 以降では, 全種類のストリームがアソシエーションされたアソシエーションストリームにより, ロバストな人物の追跡が実現できており, t_6 から t_8 については 2 話者が同時に存在する状況で正確なストリーム分離が達成できている. この際, 複数のストリームが存在する状況下で, 視野外の聴覚ストリームが存在しているため, 視覚情報を用いて, より正確な情報を得るようにアテンションチェンジが行われている (t_6).

このシナリオにおける *SIG* の視野と正面方向を Fig. 5 に表す. アテンション制御モジュールが細かいストリームの動きを吸収して, 滑らかに人物追跡を可能にするとともに, 話者が見えないような状況においても聴覚情報によ

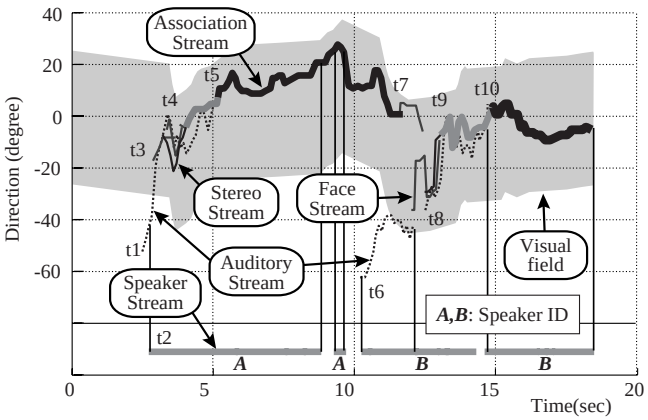


Figure 4: 2 話者におけるトラッキング結果

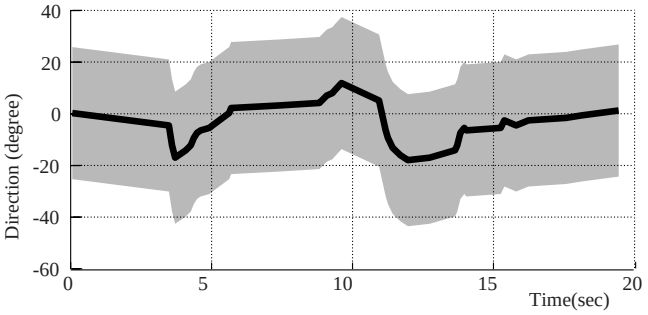


Figure 5: Fig. 4 における *SIG* の視野と正面方向

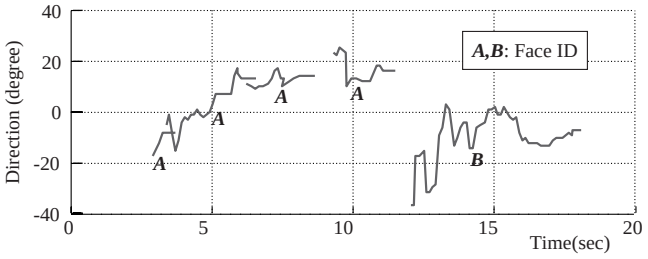


Figure 6: Fig. 4 における顔ストリーム

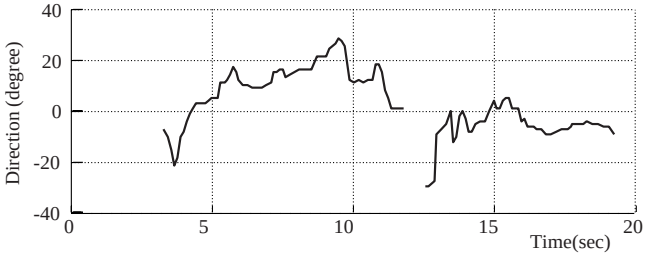


Figure 7: Fig. 4 におけるステレオストリーム

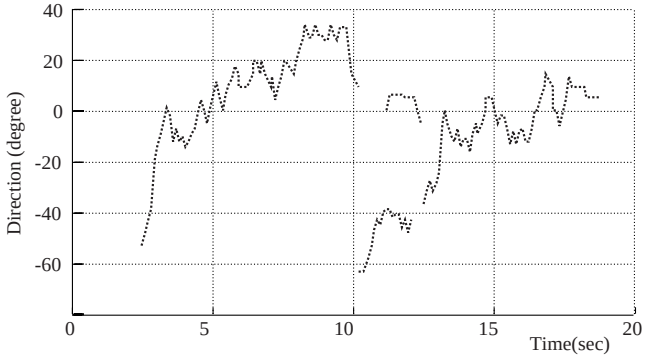


Figure 8: Fig. 4 における聴覚ストリーム

り, 的確な追跡を達成していることがわかる. 顔ストリームを Fig. 6 に示す. 顔ストリームは比較的正確な位置情報と名前情報が得られるという利点があるが, 横や下を向いてしまったり, ライティングなどの影響で, しばしば抽出に失敗するケースがある. このため 3.5, 6, 9 秒付近ではストリームが分断されてしまっている. このような場合でもアソシエーションによって追跡が継続できることを Fig. 4 は示している. Fig. 7 は, ステレオストリームを示している. ステレオストリームは比較的近い物体に関しては非常に精度のよい位置情報を取得できる. しかし, ステレオ情報は, 両眼で捕らえることができる場合のみ有効であるため, 顔ストリームと比較して視野が狭い. Fig. 4 では, 狭い視野もアソシエーションによって, カバーできることを示している. Fig. 8 は聴覚ストリームを示す. 聴覚ストリームはセンサの性質上, 全方位に渡って音情報を検出できる反面, Fig. 8 の 5 〜 10 秒の間のようにそれほど正確な音源方向の精度を得ることはできない. このような場合でも, アソシエーションによって視覚情報を補うことによって聴覚情報の曖昧性の解消が達成されている.

また, 本実験では扱わなかったが, 3 話者以上の場合やオクルージョンにより視覚情報が欠如するなど, 一部の情報が欠如した場合でも, アソシエーションによるロバストな追跡をすでに実現している [中臺 *et al.*, 2001]. さらに, 名前情報のないステレオストリームや聴覚ストリームは, 2 本のストリームが交差したり, 近接したりするような場合には, 誤りが生じる可能性がある. このような場合、顔ストリームや話者ストリームといった名前情報をもったストリームとアソシエーションを行うことによって誤り訂正が実現できることが期待できる.

Fig. 4 において, t_6 から始まる話者ストリームの後半部分は, t_8 から始まる聴覚ストリームとアソシエーションすることが妥当であり, この部分に関しては, アソシエーションの構築もしくはストリームの生成に失敗しているといえる. このような場合に対処するには, 一度アソシエーションモジュール内で高位の処理であるアソシエーションストリーム生成部から低位の処理であるストリーム生成部へフィードバックを行い, ストリームの再構築を行うような機構が必要であろう.

5 結論

本稿では, 音源方向, 話者情報, 顔情報, ステレオ視による物体情報, モータ情報を統合し, 実環境で複数の人物が存在しても実時間に追跡できるシステムを構築した. 各情報には, それぞれ特徴があり, 精度も区々である. しかし, これらをストリームおよびアソシエーションという形で実時間で統合することにより, 互いの特徴を生かし合い, 精度を高めることに成功した. 特に, 混合音を扱う研究である音環境理解 (CASA) に基づいた聴覚情報処理は, ロボットに対してそれ自体有効であり, 視覚情報の視野の不足を補うという意味でも有効であることを示した.

将来的に, より高度なソーシャルインタラクションを目標に, ロボットの知覚や認識のロバスト性を向上させるためには, 話者識別の複数話者対応, および動的に変化する環境へ対応するための学習の枠組みが必要である. さらに, 波形レベルでの相関を利用したローレベルの統合や, 音声認識を利用した意味レベルの情報との統合を行うことは有効であろう. また, 実装面では, 遅延を短くするためにギガビットイーサ等の高速なネットワークの導入も効果的であろう.

謝辞

話者識別プログラム Juno の使用許可および議論に関して, 京都大学大学院情報学研究科の秋田祐哉氏に感謝する.

参考文献

[Breazeal and Scassellati, 1999] C. Breazeal and B. Scassellati. A context-dependent attention system for a social robot. In *Proceedints of the Sixteenth International Joint Conference on Atificial Intelligence (IJCAI-99)*, pages 1146–1151, 1999.

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

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

[Matsusaka *et al.*, 1999] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi. Multi-person conversation via multi-modal interface — a robot who communicates with multi-user. In *Proceedings of 6th European Conference on Speech Communication Technology (EUROSPEECH-99)*, pages 1723–1726. ESCA, 1999.

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

[岡田 *et al.*, 2000] 岡田 慧, 加賀美 聡, 稲葉 雅幸, 井上 博允. PC による高速対応点探索に基づくロボット搭載可能な実時間視差画像・フロー生成法と実現. **日本ロボット学会誌**, 18(6):138 – 143, 2000.

[中臺 *et al.*, 2001] 中臺 一博, 日台 健一, 溝口 博, 奥乃 博, 北野 宏明. 顔認識とアクティブオーディションを利用した実時間人物追跡. **第 *11* 回 *AI* チャレンジ研究会**, SIG-Challenge-01-5:27 – 34, Mar. 2001.

[秋田 *et al.*, 2000] 秋田 祐哉, 河原 達也. 会議音声の自動アーカイブ化システム. In **電子情報通信学会技術研究報告** *(NLC2000-37/SP2000-85)*, 12 2000.

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-0317-7

事情通ロボットにおける音響信号処理

Acoustic Signal Processing in Jijo-2 Robot

浅野太，原功，本村陽一，伊藤克亘，速水悟，後藤真孝，麻生英樹，松井俊浩

F. Asano, I. Hara, Y. Motomura, K. Ito, S. Hayamizu, M. Goto, H. Asoh, and T. Matsui

産業技術総合研究所

AIST,

f.asano@aist.go.jp

Abstract

A real-time sound localization/separation system for mobile robot application was constructed and evaluated in a real office environment. As for the sound localization, the experimental results showed that the direction of the two sources was estimated with high accuracy while the range of the sources was estimated with moderate accuracy. As for the sound separation, a recognition rate of 70% for an on-line recognizer on a network and of 90% for an off-line recognizer were achieved, respectively.

1 Introduction

オフィス内のように様々な音源がある環境で，音声インターフェイスを用いて移動ロボットとのコミュニケーションを図る場合，音源の位置を推定する技術や，興味のある音声などの信号を，背景雑音や他の信号から分離する技術が重要である．我々は，これまで，オフィス内での案内や秘書のような働きをする移動ロボットの研究を行っており[1]，その音声インターフェイスの前処理系として，マイクロホンアレイシステムの開発を行ってきた[2]．本稿では，現在開発中である，ロボットの近傍における複数の音源の位置推定と分離を行うシステムを紹介する．

ロボットが話者と対話する場合，話者は，ロボットの近傍にいる場合も多い．この Near field と呼ばれるマイクロホンアレイの近傍では，音波が球面波として伝搬し，マイクロホン間に生じる位相差は，音源の到来方向だけではなく，距離の関数となる．このため，マイクロホンの近傍にある音源の信号を，マイクロホンアレイで歪み無く収録するためには，音源の距離に対する配慮が必要である．現在，開発しているシステム（第２世代システム）では，ロボットの近傍における音源からマイクロホンアレイまでの伝達関数の詳細な情報を事前に測定することにより，Near field に存在する音源を比較的低歪みで，分離収録できるのが特徴である．また，距離の変化により生じるマイクロホン間位相差を，逆に手がかりとして，複数音源の距離推定もある程度できるようになっている[3]．本稿では，これを実現するリアルタイムシステム及びその実環境評価について，報告する．また，これまでの開発の経緯，及び今後の展望などについても，簡単に触れる．

2 方法

2.1 信号のモデル

今，空間にある N 個の音源を，M 個のマイクロホンを用いて観測する場合を考える．第 m 番目のマイクロホンへの入力のフーリエ変換を $X_m(k,t)$ とし，これを用いて，次式のように入力ベクトルを定義する．

$$\mathbf{x}(k,t) = [X_1(k,t), \cdots, X_M(k,t)]^T \tag{1}$$

この入力ベクトルは，次式のようにモデル化することができる．

$$\mathbf{x}(k,t) = \mathbf{A}_k \mathbf{s}(k,t) + \mathbf{n}(k,t). \tag{2}$$

ここで，$\mathbf{s}(k,t) = [S_1(k,t), \cdots, S_N(k,t)]^T$ は音源のスペクトルから構成されるベクトル，$S_n(k,t)$ は，第 n 番目の音源のスペクトルである．k 及び t は，離散周波数及び時間フレームのインデクスである．また，行列 $\mathbf{A}_k$ は，その第 (m,n) 番目の要素に，第 n 番目の音源から第 m 番目のマイクロホンまでの伝達関数を持つ，伝達関数行列である． この行列の第 n 番目の列ベクトルは，第 n 番目の音源に対応しており，音源の位置ベクトルと呼ばれる．ベクトル $\mathbf{n}(k,t)$ は，各マイクロホンで観測される背景雑音のフーリエ変換から構成される雑音ベクトルである．

2.2 音源定位

音源位置は，次式で示される空間スペクトルから求めることができる．

$$\bar{P}(r,\theta) = \frac{1}{K}\sum_{k=k_L}^{k_H} P(r,\theta,k). \tag{3}$$

これは，各周波数で独立に求まる MUSIC 空間スペクトル $P(r,\theta,k)$[4]の周波数平均を取ったものである[3]．ここで，k_L 及び k_H は，周波数平均の範囲を示す．r 及び θ は音源の距離及び方向を示す．また，$K = k_H - k_L + 1$．各周波数における MUSIC スペクトルは，次式で与えられる．

$$P(r,\theta,k) = \frac{1}{|\tilde{\mathbf{a}}_k^H(r,\theta)\mathbf{E}_k^n|^2}. \tag{4}$$

ここで

$$\tilde{\mathbf{a}}_k(r,\theta) = \frac{\mathbf{a}_k(r,\theta)}{\|\mathbf{a}_k(r,\theta)\|} \tag{5}$$

は，スキャニングポイント (r,θ) に対する仮想的な位置ベクトルである．ただし，距離推定におけるバイアスを避けるため，ノルムで正規化してある[3]．また，行列 $\mathbf{E}_k^n$ は，空間相関行列

$$\mathbf{R}_k = E[\mathbf{x}(k,t)\mathbf{x}^H(k,t)] \tag{6}$$

を次式のように固有値展開し，

$$\mathbf{R}_k = \mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^{-1}, \tag{7}$$

その固有ベクトルを次式のように 2 つに分けたものの一方である．

$$\mathbf{E}_k = [\mathbf{e}_1,\cdots,\mathbf{e}_N|\mathbf{e}_{N+1},\cdots,\mathbf{e}_M] = [\mathbf{E}_k^s|\mathbf{E}_k^n] \tag{8}$$

ただし，固有値は，大きい順にソートされ，固有ベクトルもこれに対応してソートされているものとする． すなわち，$\mathbf{E}_k^n$ は，小さいほうから $M-N$ 個の固有値に対応した固有ベクトルである．

(4) により空間スペクトルが求まる原理を簡単に述べる．空間相関行列 $\mathbf{R}_k$ の固有値展開の結果得られた固有ベクトル $\mathbf{E}_k^s = [\mathbf{e}_1,\cdots,\mathbf{e}_N]$ は，音源ベクトル $\mathbf{A}_k = [\mathbf{a}_1,\cdots,\mathbf{a}_N]$ が張る部分空間 (Signal subspace) の基底ベクトルとなる性質がある[5]．固有ベクトル $\mathbf{E}_k^s$ と $\mathbf{E}_k^n$ は互いに直行するから，(4) における仮想的な位置ベクトル $\tilde{\mathbf{a}}_k^H(r,\theta)$ が真の位置ベクトル $\{\mathbf{a}_1,\cdots,\mathbf{a}_N\}$ のいずれかと一致した場合，(4) の分母は 0 となり，空間スペクトル上にピークが現れる．(3) では，このスペクトルを周波数平均してるが，これは，周波数により一方の音源の音源スペクトルが弱く，MUSIC 空間スペクトル上にピークが出ない場合があるためである．

Table 1: Differences of MV1 and MV2.

	MV1	MV2
Correlation	$\mathbf{K}_k$	$\mathbf{Q}_k$
Advantage	high noise reduction	high tracking capability
Disadvantage	absence of target must be detected	performance depends on localization accuracy

2.3 音源分離

最小分散法（例えば[5]）により，n 番目の音源のスペクトルは，以下のように推定される．

$$\hat{S}_n(k,t) = \mathbf{w}^H(k)\mathbf{x}(k,t) \tag{9}$$

ここで，ベクトル $\hat{\mathbf{a}}_{n,k}$ は第 n 番目の音源の位置ベクトルであり，音源定位モジュールにより推定される． ここで，フィルタ係数は，以下に定義される．

$$\mathbf{w} = \frac{\mathbf{R}_k^{-1}\hat{\mathbf{a}}_{n,k}}{\hat{\mathbf{a}}_{n,k}^H\mathbf{R}_k^{-1}\hat{\mathbf{a}}_{n,k}}. \tag{10}$$

(10) は，音源方向に全域通過の拘束条件をつけた上でシステムの出力を最小にする最適化問題の解として与えられ，結果として，ある音源の方向からくる音を透過させ，他の音源の方向に死角を持つ指向性が合成される．本稿では，従来の最小分散法をベースにした 2 つの手法（それぞれ MVBF1 及び MBVF2 と呼ぶ）を用いて，音源分離を行う．

MVBF1 では，(6) で定義される空間相関行列の代わりに，ターゲットの音源（ここでは，第 n 番目の音源とする）$S_n(k,t)$ が休止している区間で観測される空間相関行列 $\mathbf{K}_k$ を用いる．すなわち，

$$\mathbf{w}_{MV1} = \frac{\mathbf{K}_k^{-1}\hat{\mathbf{a}}_{n,k}}{\hat{\mathbf{a}}_{n,k}^H\mathbf{K}_k^{-1}\hat{\mathbf{a}}_{n,k}}, \tag{11}$$

この場合，$\mathbf{w}_{MV1}$ は，音源 $S_n(k,t)$ を推定する maximum likelihood (ML) estimator となる [5]．ここでは，便宜上，MVBF1 が従来の最小分散法のサブセットのような書き方をしているが，本来は，最小分散法が最尤推定法の近似である．(9) を用いることにより，少数のデータを用いて推定した $\mathbf{K}_k$ を用いても，$\mathbf{R}_k$ を用いた場合に比べ，高い雑音除去能力が得られる．一方，この方法の欠点は，表 1 に示すように，第 n 番目の音源 $S_n(k,t)$ が休止している区間を探さねばならないことである．

MVBF2 は，筆者らが提案している方法である[3]．この方法では，空間相関行列を音源定位の推定結果 $\hat{\mathbf{A}}_k$ から次式のように合成する．

$$\mathbf{Q}_k = \hat{\mathbf{A}}_k\hat{\mathbf{A}}_k^H + \gamma\mathbf{I}. \tag{12}$$

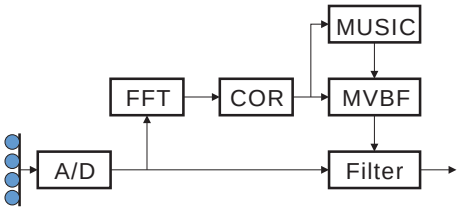


Figure 1: Block diagram of the proposed system.

Table 2: Results of the benchmark test. The processing time for 1 s data is shown.

	Time [s]	Processor
FFT & Correlation	0.86	DSP-B
MUSIC	0.34	Host
MVBF	0.16	Host
Filtering	0.52	DSP-A

この (12) を用いて MVBF2 の係数は，次式のようになる．

$$\mathbf{w}_{MV1} = \frac{\mathbf{Q}_k^{-1}\hat{\mathbf{a}}_{n,k}}{\hat{\mathbf{a}}_{n,k}^H \mathbf{Q}_k^{-1}\hat{\mathbf{a}}_{n,k}}, \tag{13}$$

(12) において，第 1 項 $\hat{\mathbf{A}}_k\hat{\mathbf{A}}_k^H$ は，方向性の音源に対応する項であり，指向性の死角の形成を行う．一方，第 2 項の $\gamma\mathbf{I}$ は，仮想的な背景雑音に相当する．ここで，$\mathbf{I}$ は，単位行列である．γ は，パラメタであり，重み付き最小 2 乗問題の重みの役割をする．γ を大きくとることにより，背景雑音に対する抑制性能が向上する．

MVBF2 の特徴は，早い適応能力である．音源定位の推定結果 $\hat{\mathbf{A}}_k$ は，0.2s 程度の短い区間のデータから推定可能である．MVBF2 では，音源分離フィルタもこの音源定位の結果から合成されるため，MVBF1 では最低でも 1s 程度のデータが必要であったが，MVBF2 では，0.2s 程度ごとにアップデートが可能である．このため，環境の動的変化に対して，追従性能を向上させることが可能である．一方，欠点は，音源分離の性能が，音源定位の推定精度に依存することである．音源定位では，事前に測定した音源位置ベクトルのデータベースから，(3) の空間スペクトルのピークに対応した位置ベクトルが選択されるため，データベースに収録されているデータの分解能以上の精度は望めない．

3 リアルタイムシステム

3.1 ハードウェア

リアルタイムシステムのブロック図を図 1 に示す．マイクロホンへの入力信号は，FFT モジュールで短区間フーリエ変換され，COR モジュールで (6) の空間相関 $\mathbf{R}_k$ が計算される．続いて，求められた $\mathbf{R}_k$ は，MUSIC モジュールへと送られ，音源位置 $\hat{\mathbf{A}}$ が推定される．(3) における周波数平均の範囲は，$[500, 3000]$ Hz とした．MBVF モジュールでは，音源分離のためのフィルタ係数 $\mathbf{w}_{MV1}$ あるいは $\mathbf{w}_{MV2}$ が計算される．MV1 については，ターゲットの音源が休止している場合の空間相関 $\mathbf{K}_k$ が必要となるため，$\mathbf{K}_k$ も COR モジュールで計算され，MVBF モジュールへと送られる．本報告では，$\mathbf{K}_k$ を推定するために，1.0s の入力データを用いた．推定されたフィルタ係数は，フーリエ変換により時間領域に変換され，入力信号はこの時間領域のフィルタにより処理される．

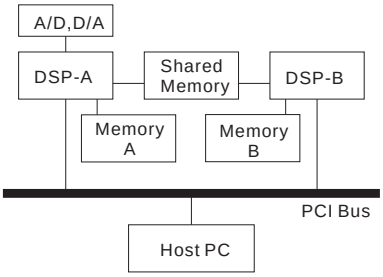


Figure 2: Architecture of the DSP system.

Table 3: Location vector database.

	Direction	Range
60-90 cm range	every 10 °	every 10 cm
100-160 cm range	every 5 °	every 20 cm

リアルタイムシステムは，2 個の DSP (TI, C6701 150MHz) 及びホストの PC (Pentium III 600MHz) により構成される．システムのアーキテクチャを図 2 に示す．計算資源の配分及びベンチマークテストの結果を表 2 に示す． 使用したマイクロホンアレイは，直径 0.5m の円形のものであり，マイクロホンの素子数は $M = 8$ である．マイクロホンアレイは移動ロボット Nomad XR-4000 の上部に搭載されている．

3.2 位置ベクトルデータベース

システムを動作させるためには，位置ベクトルデータベースを作成する必要がある．位置ベクトルデータベースは，音源が取りうるすべての位置から各マイクロホンまでの伝達関数のデータベースである．ただし，波面が平面波として近似できる Far field（本報告の場合は約 $r > 1.5$m）では，距離による位相差の変化はほとんどないため，Near field の場合だけを詳細に測定しておけばよい．本報告では，表 3 に示すように，音源距離 60-160 cm におけるあらゆる方向の伝達関数を，TSP 法を用いて収録した[6]．このデータベースは，音源定位において (5) の $\mathbf{a}_k(r,\theta)$ として用いる．また，音源分離においては， (11) 及び (13) における $\hat{\mathbf{a}}_{n,k}$ 及び，(12) における $\hat{\mathbf{A}}_k$ として用いられる．なお，測定されたデータ及び測定法，測定条件などの詳細な情報は，RWCP のホームページ[7]からダウンロードする事ができる．

Table 4: Source Location

	Source 1		Source 2	
	Angle	Range	Angle	Range
Real	10	80	70	120
Estimated	10	90	70	140

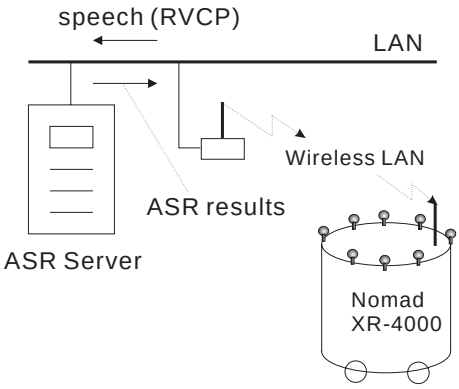


Figure 3: Online ASR system.

3.3 オンライン音声認識システム

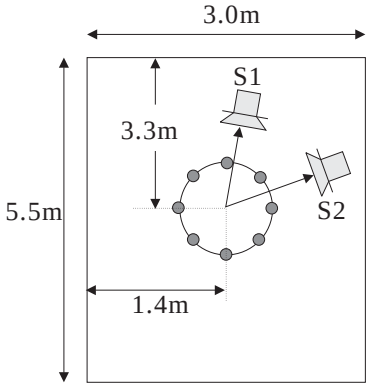
本報告で紹介する音源定位／分離システムは，移動ロボットにおいて，音声認識インターフェイスの前処理系として用いることを前提としている．移動ロボットでは，ロボットに搭載された計算資源が限定されているため，現在，ネットワーク上に分散して配置されたオンライン音声認識サーバを用いることを検討している．図 3 に，その概念図を示す．このシステム（以下 RVCP-niNja と呼ぶ）は，離散 HMM の認識エンジン niNja[8]と音声信号の情報をネットワーク上で交換するためのプロトコル (Remote Voice Control Protocol, RVCP) [9]から構成されている．このシステムの特徴は，RVCP を用いることにより，音声認識エンジン自体をいくつかのモジュールに分割して，ネットワーク上に分散配置できることである．これにより，移動ロボットが複数ネット上に存在するような重い負荷の場合でも，効率よく音声認識が行えるものと考えている．

4 評価実験

4.1 実験条件

評価実験は，通常のオフィス環境で行った．音源とマイクロホン配置を図 4 に示す．また，音源として用いたラウドスピーカの位置を表 4 に示す．音源#1 と#2 からは，それぞれ，音声と音楽を放射した．マイクロホン#1 の位置における，音源#1, 音源#2 及び背景雑音（主に PC のファン）の音圧レベルは，それぞれ，68, 67, 48 dBA であった．

音声認識実験では，比較のため，認識エンジンとして，上述の RVCP-niNja システムの他，オフラインの連続 HMM



(a) Configuration of sound sources

(b) Scene of experiment

Figure 4: Experimental setup.

認識エンジンである HTK [10]も用いた．HTK のシステムでは，IPA の音韻モデル[11]を用いている．

4.2 結果

図 5 は，MUSIC による空間スペクトルである．MUSIC による空間スペクトル推定では，音源数 N を既知として与えてある．この図から，音源の位置を推定することができる．位置推定の結果を表 4 に示す．この結果から，方向については，良好な推定が行えていることが分かる．一方，距離は，実際よりも大きく推定されている．距離に関しては，距離の変化に対するマイクロホン間の位相差の変化が，方向のそれに比べ少なく，分解能が低い．

図 6 は，日本語 492 単語に対する音声認識実験の結果である．音源分離システムとして MV1 を用いた場合は，高い認識率が得られている．一方，MV2 を用いた場合は，認識率が約 20%程度低下している．これは，表 1 で示したように，MV2 が音源定位の精度に依存することが原因である．表 3 で示したように，位置ベクトルデータベースでは，音源距離が 1-1.6m の範囲では，5° 毎に測定してある．これは，音源位置推定が最も正確な場合でも，最大で $\pm$ 2.5° の誤差を含んでいることになる．

図 7 は，MV2 を用いた場合の，妨害音源付近のゲイン

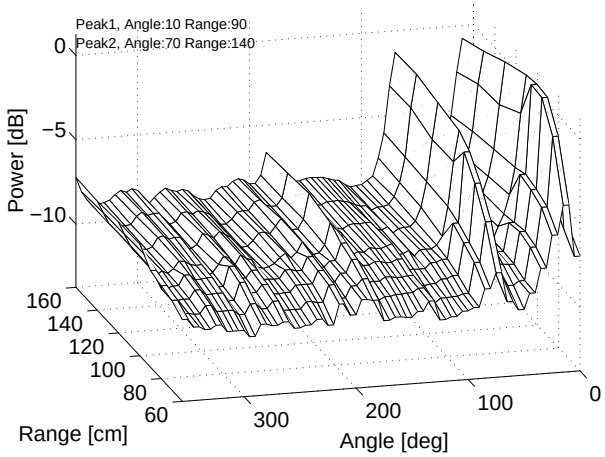


Figure 5: Spatial spectrum obtained by MUSIC.

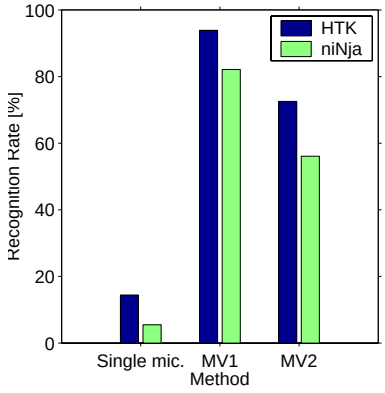


Figure 6: ASR rate.

を示している．この図から，実際の音源位置と推定された音源位置が ± 2.5° 程度ずれた場合，雑音抑制能力は，高い周波数では 10dB 程度に低下してしまうことが分かる．高域での妨害音の抑制性能を高めるためには，位置ベクトルデータベースの分解能を，例えば，現在の 5° 毎から，2.5° 毎などに高める必要がある．

5 開発の経緯及び今後の展望

5.1 第１世代システム

第１世代のシステムは，1997 年頃開発され，これ以降現在に至るまで，移動ロボット Nomad200 システムをベースにしたデモシステムで稼働している．このシステムは，8 素子のマイクロホンアレイと TI 社 DSP C44(50MHz) × 1 からなる小規模なシステムで，遅延和ビームフォーミングによる，単一音源の方向推定と背景雑音抑制を行うことができる[2]．音源推定は，1s に１度の頻度で行い，ビームをスキャンして推定した空間スペクトルから，パワーが最大の点を音源位置として推定する．音源方向を推定した後，その方向にビームを固定することにより，背景雑音を抑制する．雑音抑制の性能は， 約 10-15dB 程度である．このような単純な方法を採用したのは，主に計算資源による制約のためであるが，手法が単純であるため，

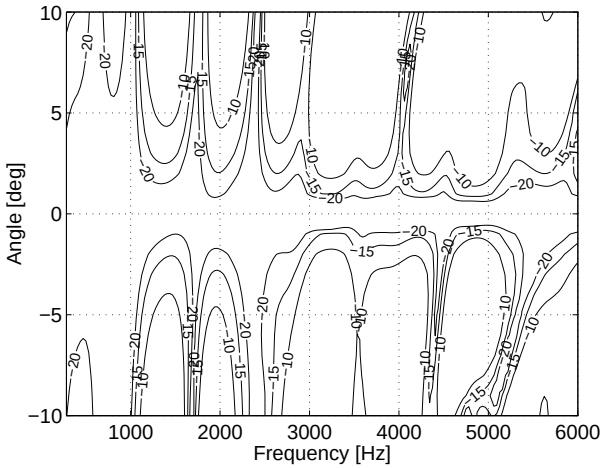


Figure 7: Gain of MV2 in the vicinity (±10°) of the interference direction.

動作は比較的安定していた．音源位置推定を用いているため，ロボットに後ろから呼びかけても，振り向いてから対話を始めるようにすることができる．これにより，話者に対して visual feedback を与えることができ，それなりの効果があったように思う．背景雑音抑制については，背景雑音は主に低域優勢であるが，アレイのサイズが小さいため，低域での抑制効果が少なく，アレイを用いることにより音声認識が目に見えて向上するほどの効果は無かった．

5.2 第２世代システム

現在開発している第２世代のシステムは，計算資源の制約を緩くし，TI DSP C6701 × 2 + Host CPU に拡張した．これにより，複数の音源の定位及び分離が可能となった．第１世代では，ビームを用いて音源定位及び背景雑音抑制を行っていたが，ビームはあまり空間分解能が高くないため，複数の音源の分離などには不向きである．第２世代システムの特徴は，ビームではなく死角 (null) を用いて音源定位及び分離を行う点であり，これにより，複数音源の定位や分離が可能となった．また，第１世代のシステムでは，平面波を仮定し，音源の到来方向だけを扱っていたが，第２世代では，球面波を扱えるようになり，Near field においてマイクロホンアレイを用いる場合の音源信号に対する歪みが低減した[3]．この代償として，計算量が大幅に増大し，電流の消費量も増え，電源の強化を余儀なくされた．また，より高度な学習を行うようになったため，システムのパフォーマンスが学習の結果に大きく依存するようになった．このため，学習が良好に行えている場合は高いパフォーマンスが得られるものの，失敗したときのリスクも増大した．

5.3 第３世代システム

必ずしも，第２世代のシステムの延長上というわけではないが，一応第３世代のシステムとして，最近流行の Blind

Source Separation(BSS) を検討している．第２世代のシステムでは，事前に入手可能な情報は，極力事前測定などで入手しておき，オペレーションの際の学習は，必要最小限にとどめるというスタンスだったが，第３世代として考えている BSS は，この対局にあり，事前学習及び測定はいっさい行わず，オペレーション時の入力信号のみから，システムを自己構築し，音源分離を行う．ロボットのようなアプリケーションでは，アレイの形状などの情報は，既知である場合が多く，完全にブラインド（観測信号のみが既知）であるという制約は必ずしも必要ない場合もあるが，図 3 に示すように，任意の音声インターフェイスがネット上にぶら下がっているような場合は，観測信号のみから，所望の信号を取り出す技術は，有用であるものと考えられる．現在，反射のある環境でブラインド分離を行う手法を開発しており，本稿で紹介した評価実験と同程度の環境で，約 60-70%程度の認識率が得られている[12].

6 まとめ

本稿では，現在開発している，移動ロボットのための音源定位・分離システムを紹介した．このシステムの特徴は，

- 複数の音源の方向及び距離を推定できる
- Near field においてアレイ処理をした場合の信号歪が少ない
- 音源分離フィルタとして MV2 を用いた場合は，0.2s 程度ごとにフィルタ係数をアップデートでき，環境に早く追従できる

などである．今後，フィールドテストなどを重ね，評価及び性能の向上を目指したい．

参考文献

[1] 松井俊浩，麻生英樹，John Fry，浅野太，本村陽一，原功，栗田多喜夫，速水悟，山崎信行，“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.

[2] 浅野太， 速水悟，松井俊浩，“話者方向同定と雑音抑制による音声認識性能の改善，” **日本音響学会誌**, vol. 53, no. 11, pp. 889–894, 1997.

[3] Futoshi Asano, Hideki Asoh, and Toshihiro Matsui, “Sound source localization and separation in near field,” *IEICE Trans. Fundamentals*, vol. E83-A, no. 11, pp. 2286–2294, November 2000.

[4] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propag*, vol. AP-34, no. 3, pp. 276–280, March 1986.

[5] Don H. Johnson and Dan E. Dudgeon, *Array signal processing*, Prentice Hall, Englewood Cliffs NJ, 1993.

[6] Yoiti Suzuki, Futoshi Asano, Hack Yoon Kim, and Toshio Sone, “An optimum computer-generated pulse suitable for the measurement of very long impulse response,” *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1993.

[7] *http://tosa.mri.co.jp/sounddb/.*

[8] K. Itou, S. Hayamizu, K. Tanaka, and H. Tanaka, “System design, data collection and evaluation of a speech dialog system,” *IEICE Trans. INF & SYST.*, vol. E76-D, no. 1, pp. 121–127, Jan. 1993.

[9] Masataka Goto, Katunobu Itou, Tomoyosi Akiba, and Satoru Hayamizu, “Speech completion: New speech interface with on-demand completion assistance,” in *Proc. of HCI International 2001*, 2001.

[10] *http://htk.eng.cam.ac.uk/.*

[11] T. Kawahara, T. Kobayashi, K Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, “Japanese dictation toolkit: Plug-and-play framework for speech recognition r&d,” in *Proc. of ICASSP'99*, March 1999, pp. I–393.

[12] Futoshi Asano, Shiro Ikeda, Michiaki Ogawa, Hideki Asoh, and Nobuhiko Kitawaki, “A combined approach of array processing and independent component analysis for blind separation of acoustic signals,” in *Proc. ICASSP2001*, May 2001, vol. V, pp. MULT–P2.

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-0317-8

車内音声対話の高度化に向けて

Advanced in-car speech communication

武田一哉 早川昭二 磯部俊洋 瀬川修 清水司 村尾浩也 河口信夫 板倉文忠
K.Takeda S.Hayakawa T.Isobe O.Segawa T.Shimizu H.Murao N.Kawaguchi and F.Itakura
名古屋大学　統合音響情報研究拠点
Center for Integrated Acoustic information Research, Nagoya University
takeda@nuee.nagoya-u.ac.jp

Abstract

Providing a safe and confatable interface while driving is one of the most important applications of the spoken language understaning. In order to build an advanced in-car speech dialogue system, there are many problems to be solved, namely, utterances affected by driving operation, robust detection and recognition of speech under heavy noise, understanding highly spontaneous speech. This paper describes the research efforts of the Center for Integrated Acsoutic Information Research (CIAIR, Nagoya University) toward the advanced in-car speech communication.

1　はじめに

音声対話技術により，車内における情報システムへの安全・快適なアクセス手段を提供することに期待が集まっている．しかし，走行中に車内において音声対話を円滑に進めるためには，運転動作と対話との関連，雑音下での音声の切り出しと認識，不完全な文で構成される発話の音声の理解など，様々な課題を解決する必要がある．本稿では，名古屋大学統合音響情報研究拠点において行われている車内音声対話の収集実験の経過とその分析により得られた知見を中心に，話言葉処理の重要な応用分野である車内音声対話の高度化の技術的な課題について論じる．

2　車内音声対話収集実験

車内音声の収集は，特別に設計された実験車[2]により行っている．[4]，[3]．実験車には，16 チャネルの音声，3 チャンネルの画像，自動車運転情報（アクセル，ブレーキ，ハンドル，エンジン回転数，速度）を同期して収録する機能が実現されている．被験者は，同乗する実験オペレータの指示の下，あらかじめ決められた市街地内の道路を実際に車を運転しながら，実験オペレータや Wizard of OZ システム，音声認識を用いた音声対話システムとの対話を行なう．対話の内容の書き起しも行なっている．また対話だけでなく，音声プロンプトを用いた音素バランス文の発声の収録も行なっている．本実験車を用いて，これまでに約 500 名の被験者に関する実験データが得られている．

3　車内音声対話の基本性能

3.1　システムとの対話の収録

音声対話システムと運転者のと対話は，車内音声対話の高度化に必須の研究資料である．そこで，プロジェクトではレストラン検索に関する音声対話システムを車内で動作させ，対話データの収集をおこなっている．被験者は運転しながら料理のジャンルや種類，場所，値段等の検索条件を音声で入力して検索し，希望のレストランを決定する．対話音声収集のための音声対話プロトタイプシステムの構成を以下に示す：

音声認識部　音声認識エンジンとして「日本語大語彙連続音声認識エンジン Julius v3.1」[5] を使用した．音響モデルは「日本語ディクテーションソフトウェア 99 年度版」に含まれる性別非依存 PTM モデル[6]を用いた．音声入力はヘッドセットにより行い，音響的な音声認識性能の劣化を避けている．統計的言語モデルは，予備実験で収録した人対人のレストラン検索対話を書き起こした文と，レストランデータベースに含まれる店名が出現するように人工的に作成した文を合わせた約 2000 文を用いた．これらを Chasen に掛け，CMU Tool Kit により bigram(カットオフ 1) と単語辞書を作成した．単語数は 1423 単語である．

音声合成部　音声合成エンジンとしては波形編集方式の「LINUX 音声合成ライブラリ」を用いた．レストラン

Table 1: 対話のログ

00:00:03-00:00:05	S:	レストラン案内システムです
00:00:05-00:00:07	S:	何か用ですか
00:00:07-00:00:07	U:	
00:00:08-00:00:11	S:	検索条件を入力して下さい
00:00:11-00:00:14	U:	＜ｓ＞ 近く+チカク+16/0/0 で+デ+59/0/0 ＜/s＞
00:00:14-00:00:16	S:	この近くでよろしいですか
00:00:16-00:00:20	U:	＜ ｓ ＞ はい+ハイ+72/0/0 ， +, +76/0/0 一番+イチバン+16/0/0 近い+チカイ+49/0/0 とこ+トコ+2/0/0 で+デ+59/0/0 ＜/s＞
00:00:24-00:00:26	S:	他に条件はありますか
00:00:27-00:00:31	U:	＜ｓ＞ 安い+ヤスイ+49/0/0 ところ+トコロ +22/0/0 ＜/s＞
00:00:31-00:00:41	S:	この近くに和食のうなぎ木屋 中華のかっぱ園菜館 和食の亀彦 が, ありますどのお店にしますか
00:00:41-00:00:44	U:	＜ｓ＞ ん+ン+21/0/0 二番目+ニバンメ+4/ 0/0 の+ノ+68/0/0 ところ+トコロ +22/0/0 ん+ン+21/0/0 ＜/s＞
00:00:44-00:00:47	S:	かっぱ園菜館でよろしいですか
00:00:47-00:00:49	U:	＜ｓ＞ はい+ハイ+72/0/0 ＜/s＞
00:00:54-00:00:59	S:	ご希望のお店は, 予約が可能です. 予約しますか
00:00:59-00:01:02	U:	＜ ｓ ＞ し+シ+45/0/0 ます+マス +71/0/0 ＜/s＞
00:01:02-00:01:07	S:	それでは, 予約しますご希望のお店までナビゲーションいたしますか
00:01:08-00:01:10	U:	＜ ｓ ＞ はい+ハイ+72/0/0 市+シ+28/0/0 の+ノ+68/0/0 ＜/s＞
00:01:10-00:01:14	S:	それでは, ご案内しますので指示に従って下さい

の店名には音声合成の言語解析用辞書では解析できない名前も含まれているため，正しい店名を与えたユーザ辞書を作成し用いた．

対話管理部　対話の状態は 15 種類で，それぞれの状態においてスロット等の条件に従って，遷移先の状態と遷移に伴うアクションが定義されている．

検索部　認識結果からテキストマッチと品詞番号の参照により検索のためのキーワードを 6 種類のスロットに埋めている．検索の結果，候補の店が 3 件を超えているならば，さらに条件を入力させ，3 件以内なら候補を読み上げるようにしている．

対話毎に作成される対話記録 (時間情報付き) の一例を表 1に示す．

被験者に実際に音声対話システムを利用してもらい，対話システムが認識のために自動音声切り出しした音声データを用いてシステムの認識性能を評価した．48 名の被験者について，システムに話し掛けている発声のみ 1153 文を Juliusv3.1 を用いて，対話システムと同じ条件で認識させ，単語認識率を算出した結果，単語正解率 71.5%，単語正解精度 61.7%であった．

3.2　収録対話の分析

対話システムが記録したログを分析し，タスクを達成した 73 対話について，アクセプト率[7]と対話時間の関係を調べた結果を図 1に示す．平均アクセプト率は 74.3%，平均対話時間は 83.6 秒となった．

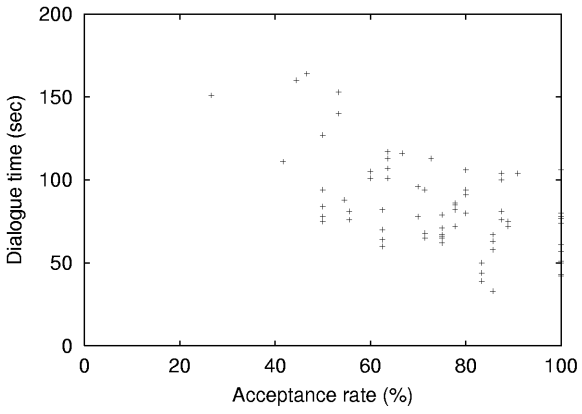


Figure 1: アクセプト率と対話時間の関係

事後アンケートに協力してくれた被験者 55 名に改善すべき点を 6 種類の選択肢から選んでもらった．結果を表 2に示す．まず第一に合成音の聞き取り易さの改善を指摘する被験者が多い．これは以下の原因が考えられる：

- 検索結果であるレストラン名を音声合成部が自然なアクセントで読めない．これはレストラン名は固有名詞が多く，ほとんど音声合成部の言語解析辞書にないため，アクセント情報が付与されないため．
- 初めて聞くレストラン名 (特にカタカナ名) を合成音声で提示されても聞き取れない．(例：トラットリアマルコポーロ，アルページュ)
- 運転中の安全確認に注意が向いているときには，合成音が何を言っているのかわからなくなる．

また認識性能を問題とする被験者は 4 分の 1 であった．これはヘッドセットで音声入力を行っているため，自動車騒音の影響を低く抑えれたためと考えられる．

今回の収集の結果から，音声対話による運転中の車内情報システムの操作性は，合成音声による情報提示の仕方が重要な要素であることが明らかになった．合成音声

Table 2: 改善すべき点

選択肢	人数	割合
正しく話した言葉を認識する	14	25%
システムの反応速度	6	11%
何を発声したらよいは分からない	6	11%
合成音の聞き取り易さ	25	46%
システムの機能が足らない	0	0%

により正確に情報がドライバーに伝われば，未知語発声による音声認識誤り率も削減されると考えられる．

4 車内音声の切り出しと認識

従来の連続音声認識システムの枠組においては，前処理として発話の正確な端点検出が必要であり，常に「発話単位」というものを意識したデコード処理が必要であった．ところが自然発話においては読み上げ文とは異なり，そもそも発話単位そのものが不明確であり，これまで提案されている明示的な端点検出手法を用いた場合，最適な発話単位の分割が必ずしもうまくいかないという問題がある．そこで本節では実環境下で対話音声を認識するアプローチの一つとして，端点検出を行わない連続音声認識手法 [8] を紹介する．

4.1 アルゴリズム

本手法では，音声・非音声の区別なしに入力音声ストリームを数秒程度の一定時間長に機械的に分割した処理ブロック (セグメント) を逐次取り込みながら連続音声認識を行う．一つのセグメントの終端まで認識処理が終った時点で部分単語列を出力しながら認識を進めるため，一定間隔で逐次的に認識結果を出力しながら無限長の連続音声のデコードが可能である．分割されるフレームの箇所は任意であるため，単語区間 (あるいは音素区間) の途中で途切れが起こる可能性が高い．このため，本手法ではセグメント単位の認識がセグメント終端まで達した時点で近傍の信頼性の高い単語境界のフレームを探索し，その時点に遡って再探索を行う．

4.1.1 評価実験

音響分析は，16kHz サンプリング，分析フレーム長 25ms，Hamming 窓，フレームシフト 10ms，特徴量として MFCC 12 次，ΔMFCC 12 次，Δ パワーの計 25 次元を使用した．音響モデルとして IPA ディクテーションツールキット付属の Phonetic Tied-Mixture HMM(性別非依存モデル)[6]を用い，言語モデルは CIAIR 車内音声対話コーパスの書き起こし約 3 万文より学習した語彙数 5232 の単語 bigram および単語 trigram を用いた．

評価データとして言語モデル学習用とは別の合計 4 セッションの対話データを用いた．各セッションにつきドライバーは異なる男性 1 名，ナビゲータは女性 3 名のうち 1 名である．音声の収録には接話マイクを使用している．発話の 1 秒あたりの平均モーラ数はドライバーで約 6.2，ナビゲータで約 5.8 であり，フィラーの出現頻度は 1 発話単位あたりドライバーで約 0.33 個，ナビゲータで約 0.04 個であった[4].

Table 3: 評価に用いた対話データ (D:Driver, N:Navi)

Session	時間 [s]	単語総数	話者
m1051	688	1034	男性 a(D), 女性 a(N)
m1082	702	1442	男性 b(D), 女性 b(N)
m1120	686	1367	男性 c(D), 女性 c(N)
m1122	693	1617	男性 d(D), 女性 c(N)

認識実験はセグメントサイズを 400 フレーム (4 秒) に設定して行った．ドライバーとナビゲータそれぞれの認識率 (%Correct) を表 4に示す.

Table 4: 認識実験結果 (segment size = 400 frames)

Session	Driver(%Corr)	Navi(%Corr)
m1051	48.13	53.71
m1082	44.18	57.09
m1120	48.93	57.19
m1122	52.23	55.59

男女 2 名の発話と環境音が混在する長時間連続音声に対し，発話単位検出を行わない自動書き起こしが可能であることがわかった．ナビゲーターと比較してドライバーの認識率が低いのは，ドライバーのほうがより発話の自由度が高く話者による変動も大きいためと考えられる．全体的に書き起こしの認識性能が十分でないのは，車内音声対話のタスクに対して音響モデル，言語モデルの最適化がなされていないことが要因として考えられる．なお，

同一タスクの対話 20 セッションの中よりドライバー発話のみ選んで，発話単位に切り出した 200 発話 (話者は男女各 10 名) を認識した場合の認識性能 (%Correct) は 58.60 であったことから，音声認識処理を連続に行なうことに起因する性能劣化は小さい．

5 運転動作が発話に与える影響

運転中に音声対話システムを利用する場合，ドライバは，運転行動によって何らかの心的影響を受け，発話は通常の発話と言語的・音響的特徴が異なることが考えられる．そこで本節では，約 200 人の被験者から収集した停車中と運転中における電話番号案内タスク対話のデータを分析し，言語的特徴の一つとしてフィラーの出現に関する特徴について述べる．

被験者は，車内のセンタコンソール上方に設置された提示パネルに基づいてタスクを進行する．パネルは 2 枚

1 組になっている．1 枚目のパネルには店名と住所が記載されている．2 枚目のパネルには詳細住所と業種が記載されており，(番号案内の) オペレータから詳細な情報を求められた際に用いる．207 名の被験者が，停車中（名古屋大学の駐車場内）と運転中（名古屋市内の郊外路）で，それぞれ 2 回ずつタスクを行った．

収集したデータの書き起こしを行い，フィラーの出現密度について調べた．フィラー出現密度は，ドライバの発話中における 1 文節あたりのフィラー数として，次のように定義した．

$$\text{出現密度} = \frac{\text{フィラー数}}{\text{文節数}}$$

アクセル OFF 操作が運転余裕と関係している[10]ことが報告されており，車両操作がフィラー出現密度に影響を及ぼすことが考えられる．そこで，平均速度，アクセル／ブレーキ操作とフィラー出現密度の関係について分析を行った．

図 2 にドライバの発話区間内での平均車速とフィラー出現密度の関係を示す．平均車速が上がるに従いフィラーの出現が増加する傾向が見られるが，統計的な有意差は見られなかった．表 5 にアクセル／ブレーキ操作とフィラー出現密度の関係を示す．今回，操作有無の判断基準として，発話開始 2 秒前から発話終了までにペダルの踏み込み力がある基準値（アクセル 2.0 kg，ブレーキ 2.5 kg）に対して上下に変動したかどうかを用いたところ，操作中には，フィラー出現密度が高かった．今後，アクセル／ブレーキ操作とフィラー出現密度について詳細な分析を進める必要がある．

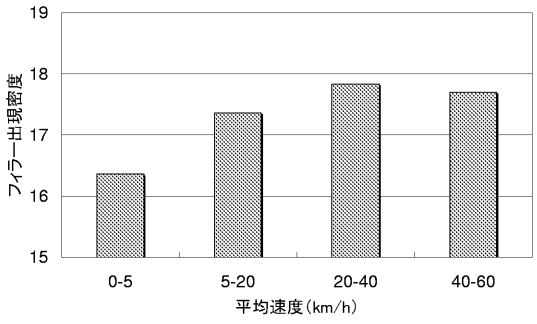


Figure 2: 平均速度毎のフィラー出現密度

6 頑健な意図の抽出

自動車内音声対話システムは次のような要件を満たすことが必要であると考えている．

- ユーザの自由な発話を受理できること運転中のドライバーに発話時の負担をかけないよう，運転中に自然に現れる発話形態 (不完全な発話，多様な言い回しなど) を受理する．
- 自然な応答を行えること画一的でない，多様な表現により応答を行う．

Table 5: 運転動作とフィラー出現密度との関係

アクセル操作	フィラー出現密度
あり	23.9%
なし	17.0%
ブレーキ操作	フィラー出現密度
あり	20.0%
なし	17.1%

これらの要求に応え，自動車内で行われる自由な発話を理解するためには，実際に自動車内で行われた対話データに基づいた発話理解を行うことが有効であると考えられる．これは，人間の行動事例に基づいたアプローチをとることで多様な人間の振る舞いに対応可能な対話システムを実現することができると考えられるためである．そこで本節では，車内音声対話実験において収集した音声対話コーパスに基づく車内音声対話システムについて概説する．本システムは，熟練したオペレータの判断を取り込んだ対話事例データベースを持ち，入力発話の断片や検索結果などをキーとして類似事例を抽出し発話理解，応答生成を行うものである．

6.1 アルゴリズム

図 3にシステムの構成を示す．対話のタスクは自動車内での店舗，施設等の情報検索である．対話事例データベースは熟練した人間 (オペレータ) が行ったユーザとの対話事例を記録したデータベースである．具体的には入力発話内容 (テキスト)，検索式，検索結果，応答発話内容 (テキスト) のセットである．発話内容のテキストは形態素解析されており，重要語 (店名，施設名，食品名など) は意味別にクラス化され単語クラスタグが付与されている．情報データベースは店舗，施設等の情報データベースであり，音声対話コーパスを収集する際に用いたものと同じものを用いる．検索式生成部は入力発話に対し，対話履歴を参照して，その時点の対話の状況に最も近い事例を対話事例データベースから抽出し，その事例で用いられた検索式を現在の状況に合うように修正して出力する．検索実行部は検索式を用いて情報データベースにアクセスし，検索結果を得る．

プロトタイプの動作を，図 4の例を用いて順に示す．(1) テキスト列で与えられた音声認識結果から，キーワードを抽出し，キーワードマッチングにより，対話事例データベース中から最類似事例を抽出する．キーワード選択法

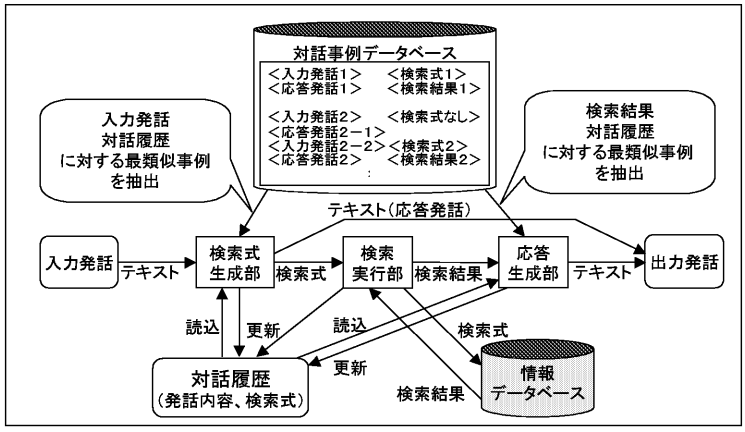


Figure 3: 事例に基づく音声対話

は種々考えられるが，現在は自立語を用いている．(2) 抽出した事例に対応する検索式を対話事例データベースから取り出し，入力発話に合うように修正して検索を行う．(3) 検索結果が得られると，入力キーワード，検索式，検索結果を用いて，対話事例データベース中から最も類似した対話事例を抽出する．(4) 抽出した 対話事例でオペレータが行った応答発話を取り出し，現在の状況に合うように修正して応答発話のテキスト列として出力する．

6.2 評価実験

この対話制御法の有効性を評価するため，対話事例データベースの規模と，正しい検索式が生成される割合との関係を実験的に調査した．対話事例データベースは，模擬的な車内情報システムとして発話を制限するよう訓練されたオペレータが，ドライバーと走行車内にて行った対話データ [1][3] に基づいている．すなわち，オペレータが実際に検索装置を操作しながら対話を行なった結果に基づき，対話内容に対応する検索式と検索結果のデータベース化をおこなった．

結果を図 5に示す[11]．話者数 28 名の対話に現れた事例を用いることで，約 70%の発話に対して正しい検索結果を抽出することができた．また，事例データベースの規模を拡大することで正解率が向上することも確認できた．

一方主たる誤りは以下に起因するものであった．

- 事例不足 (31%) 「ここら辺にスーパーはあるかな」:「スーパー」に関する対話事例がない
- 入力文が曖昧で検索式が作れない (29%) 「さっぱりしたものとか食べたいんだよね」
- 最類似事例抽出方法の問題 (18%) キーワード抽出や，事例との距離関数の不備．

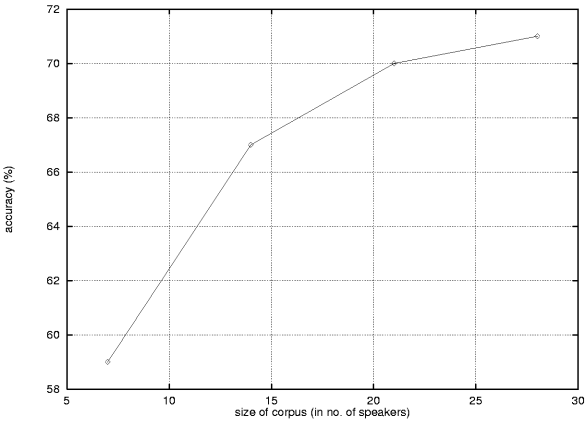


Figure 5: 評価結果

- コーパス書き起こしの誤り (12%) 対話書き起こしと形態素解析辞書との不一致など．
- 未知語 (10%) 「冷やしうどん」「鍋焼うどん」など

これらの結果から，事例データおよび語彙規模を拡大することで，さらに性能の改善が期待される．

7 むすび

本稿では，走行自動車内で安全・快適な情報インターフェスを実現するために，音声対話技術が克服すべき問題をいくつか示すとともに，CIAIR における取り組みを概説した．今後収集データの分析を進めるとともに，収集データの整理と配布，収集データから学習した音響モデルや言語モデルの学習・配布に務めていく．

謝辞

日頃ご議論いただく，名古屋大学統合音響情報研究拠点の教官・研究員・学生諸氏に感謝いたします．本研究

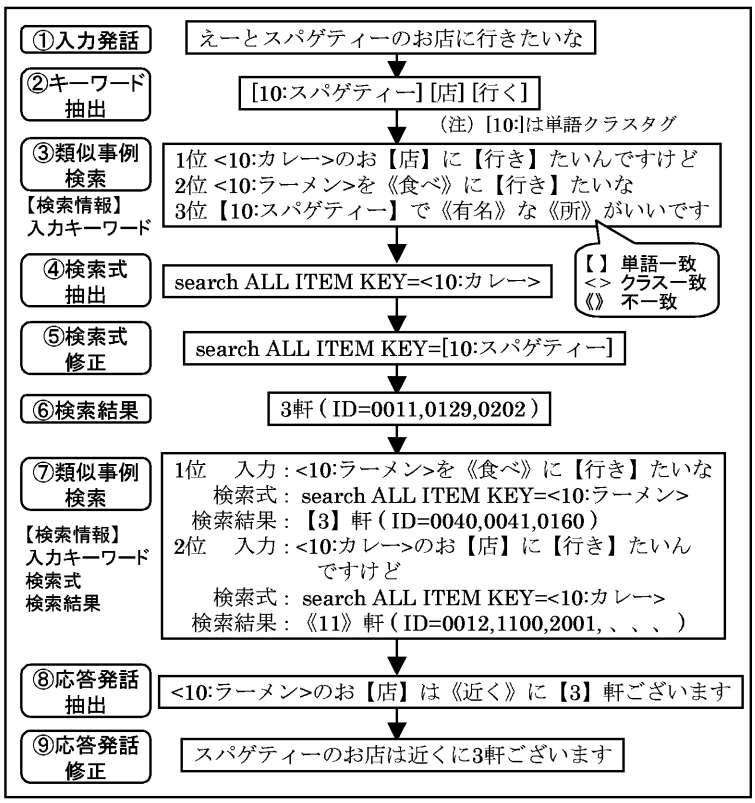


Figure 4: 対話制御アルゴリズム

は文部省科学研究補助金 COE 形成基礎研究費 (課題番号 11CE2005) の補助を受けて行われた.

参考文献

[1] 武田, 板倉: 文部省 COE プログラム統合音響情報研究拠点 (CIAIR) -音声・音響情報処理の多角的研究-, 音学誌 **56**,11, pp.748-751 (2000)

[2] 岩他: 実走行環境下における車内音声対話・音響データ収録装置, 音講論集 **1-Q-29**, pp.189-190 (2000.3)

[3] 河口 他: “実走行車内における音声データベースの構築”, 情処研報 SLP30 pp.57-62. 2000.

[4] 河口 他: “実走行車内音声対話コーパスの設計と特徴”, 情処研報 SLP34 pp.179-184 2000.

[5] 河原 他: “日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価”, 情処研報 SLP31 pp.9-16 2000.

[6] 李他: Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識, 信学論 **J83-DII**,12, pp.2517-2525 (2000)

[7] 山本他: データ入力システムの性能と使用感に関する調査, 信学技報 SP96-77,12, pp.59-66 (1996)

[8] 瀬川, 武田, 板倉: “端点検出を行わない連続音声認識手法”, 情処研報 SLP34 pp.101-106 2000.

[9] 李, 河原, 堂下: “単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識”, 信学論 D-II Vol.J82-D-II No.1 pp.1-9 1999.

[10] 内山他 : 運転状況適応型音声情報提示システム, ヒューマンインタフェースシンポウム 2000 論文集 pp.375-378 (2000.9)

[11] 村尾他: 対話事例を利用した音声対話システムの提案, 音講論集 **1-5-24**, pp.47-48 (2000.3)

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-0317-9

音声認識や音環境理解のための実環境音声・音響データベースの構築

A Design of Acoustic Sound Database Collected for Hands-Free Speech Recognition and Sound Scene Understanding

西浦敬信 (ATR-SLT)[1] Takanobu NISHIURA
中村哲 (ATR-SLT)[1] Satoshi NAKAMURA
比屋根一雄 (MRI)[2] Kazuo HIYANE
飯尾淳 (MRI)[2] Jun IIO
浅野太 (AIST)[3] Futoshi ASANO
山田武志 (Univ. of Tsukuba)[4] Takeshi YAMADA
小林哲則 (Waseda Univ.)[5] Tetsunori KOBAYASHI
金田豊 (Tokyo Denki Univ.)[6] Yutaka KANEDA

Abstract

Recently importance of hands-free speech communication is increasingly recognized. The sound data for open evaluation is necessary for the studies such as sound source localization, sound retrieval, sound recognition and hands-free speech recognition in real acoustic environments. This paper reports on our project aiming the acoustic data collection. There are many kinds of sound scenes in real environments. The sound scene is specified by sound sources and room acoustics. The number of combination of the sound sources, source positions and rooms is huge in real acoustic environments. We assumed that the sound in the environments can be simulated by convolution of the isolated sound sources and impulse responses. As an isolated sound source, a hundred kinds of environment sounds and speech sounds are collected. The impulse responses are collected in various acoustic environments. Additionally we collected sounds from the moving source. In this paper, progress of our sound scene database collection project and application to environment sound recognition and hands-free speech recognition are described.

1 Introduction

Generally, auditory as well as visual information is quite important for human beings to sense surrounding environments. This information is essential for human interaction with the environment. Human beings really sense the surrounding environments accurately integrating both visual and auditory information complementary. For instance, the auditory information plays a more important role for sensing the rear environments. Here, we call the sound environments by the word *sound scene*.

Almost all research on auditory information has been conducted focusing on the individual study of acoustic signal processing, auditory processing, and speech communication. However, the most important point is the close cooperation and integration of these functions to understand the sound scene. To understand a specific sound, the system needs to localize the target sound among multiple sound mixtures in the environment, and focus on the sound.

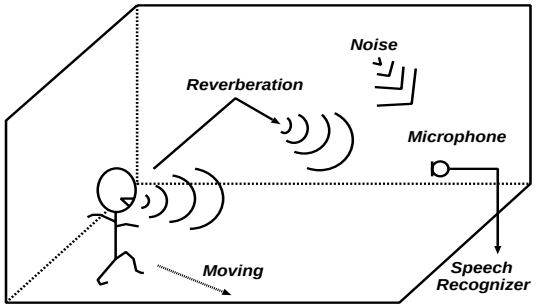


Figure 1: Real environments

The hands-free speech recognition will bring us so natural and friendly man-machine interface that users are not encumbered by microphone equipments and that users can utter from distance while moving. This hands-free speech recognition is actually an urgent technology for the hands-free interface of a car navigation system and a cellular telephone in the car.

If the speaker utters the speech from distance, the accuracy will be seriously degraded by the influences of the noise and reverberation of the room (Figure 1). The speech recognition performance even using a desktop microphone will be also varied if the distance between a mouth and a microphone is changed, and if the speaker turns his face to another direction. The fundamental problems of hands-free recognition already have

[1] ATR 音声言語通信研究所 (ATR Spoken Language Translation Research Laboratories, Kyoto),
{tnishi, nakamura}@slt.atr.co.jp
[2] 三菱総合研究所 (Mitsubishi Research Institute, Tokyo),
{hiya,iiojun}@mri.co.jp
[3] 産業技術総合研究所 (National Institute of Advanced Industrial Science and Technology, Ibaraki),
f.asano@aist.go.jp
[4] 筑波大学 (University of Tsukuba, Ibaraki),
takeshi@is.tsukuba.ac.jp
[5] 早稲田大学 (Waseda University, Tokyo),
koba@tk.elec.waseda.ac.jp
[6] 東京電機大学 (Tokyo Denki University, Tokyo),
kaneda@c.dendai.ac.jp

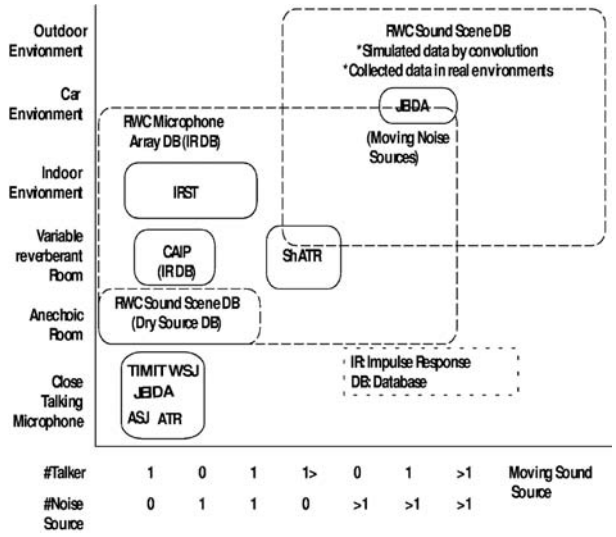


Figure 2: Focus of the RWCP sound scene database from a point of view of sound sources and acoustic environments

lain in the previous speech recognition framework. To these problems, the following technologies are required,

- Robustness to directional noise and omni- directional noise (diffuse noise) in the room.
- Robustness to acoustic reflection and reverberation in the room.
- Localization, tracing and recognition of the speaker among many sound sources including other speakers and noise.

These problems are quite new ones which previous studies haven't been considered. In fact, performance of current LVCSR will be seriously degraded if used in this hands-free context.

To conduct these researches, the collection of sound scene data in real acoustic environments is indispensable. The sound scene database contributes to promote a study of sound scene understanding. Only a few databases were developed for the study of sound mixtures. ShATR [1], reported in 1994, is a database of multi-simultaneous-speakers. Spoken dialogues of five speakers using five headset microphones and one desktop microphone were collected. Video images are also recorded by a camera mounted at the ceiling. However, the ShATR focused only on a study of human perception of mixture of speech utterances in natural surroundings. On the other hand, CAIP and IRST reported databases collected using a microphone array in [2, 3, 4]. These databases are very valuable for the microphone array studies. However, the variety of acoustic environments is very limited for a study of sound scenes in real acoustic environments.

Figure 2 shows the focus of the RWCP sound scene database from the point of view of sound sources and acoustic environments. JEIDA database [5], ATR database [6], and ASJ database [7] are databases collected only for study of speech recognition using a close talking microphone. JEIDA also includes noise data collected in a car while driving on the real road. As indicated in the figure, the RWCP sound scene database aims to collect a variety of sound scenes systematically. The figure also indicates the lack of the database for the study of source localization, sound retrieval, sound recognition and speech recognition for hands-free speech communication and security systems.

In this paper, we describe our sound scene database which is composed of isolated environment sounds and impulse responses in various rooms. Then the results of the some experiments using this database are also described[8, 9].

2 Sound Scene Database

This project is one of the projects supported by RWCP (Real World Computing Partnership). The objective of the project is to provide common standard databases for research concerning real acoustic environments.

It is almost impossible to collect all combinations of the existing sound sources and real acoustic environments. Thus, we started to collect two kinds of sound data. The first data is isolated sounds of environment non-speech sounds and speech sounds. We call the isolated sounds recorded in an anechoic room by the word *dry source* in this paper. The dry source is free from influences of room acoustics. The second data is impulse responses in various acoustic environments. The sound in the environment can be simulated by convolution of the dry sources and the impulse responses. However, there are sounds which is unable to simulate by the convolution such as non point source sounds and moving sound sources. We collected those sounds using a three dimensional microphone array. The microphone array database enables to extract arbitrary sounds by various beam-forming algorithms.

The data is collected in an anechoic room, a variable reverberant room, office environments, where many sound sources exist. Various kinds of sound sources including speech are also collected as target sounds.

3 Data Collection

3.1 Dry Source Database

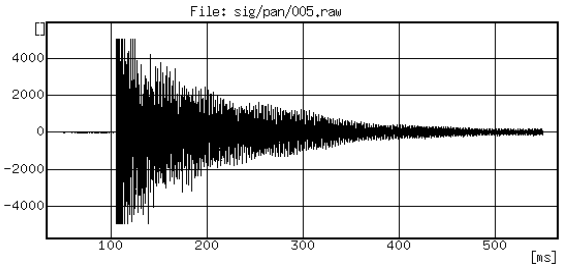
Dry source is the sound recorded in an anechoic room which is free from room acoustics. The environment sound can be simulated by convolution of the dry source and an impulse response if the transmission channel is linear and stationary. We collected three kinds of environment sounds shown in table 1. The first class is collision sounds of wood, plastic and ceramics. The second class and the third class are composed of sounds occurred when human beings operate on things like spray, saw, claps, coins, books, pipes, telephones, toys, etc. The sounds of the second class are the sounds whose source materials can not be easily associated. Whereas the source materials of the third class sounds can be easily associated uniquely.

We recorded around 100 samples for about 90 kinds of sounds sufficient enough for statistical model train-

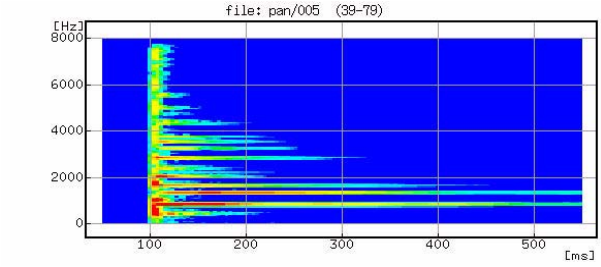
Table 1: Dry Source sound

	Category	#samples	Sound source
Collision Sound	Wood	1187	wood boards, wood stick
	Metal	1000	metal boards, metal stick
	Plastic	550	plastic boards, plastic stick
	Ceramic	800	glasses, china
Action Sound	article dropping	200	dropping article in box
	gas jetting	200	spray, pump
	rubbing	500	sawing, sanding
	bursting and breaking	200	breaking stick, air cap
	clapping sound	829	hand clap, slamming clip
Characteristic Sound	small metal articles	1072	small bell, coin
	paper	400	dropping book, tearing paper
	musical instruments	1079	drum, whistle, bugle
	electronic sound	705	phone,toy
	mechanical	1000	spring, stapler

ing. The recording is conducted in an anechoic room by B&K 4134 microphone and DAT recorder in 48kHz 16bit sampling. SNRs of the data are around 40-50dB.



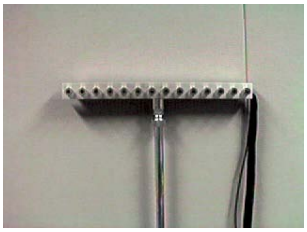
(a) Graph image of waveform



(b) Graph image of spectrogram

(c) Photograph of the sound source

Figure 3: Sample data of non-speech sound dry source



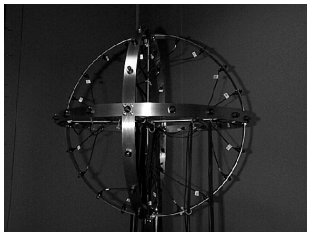


Figure 4: 14ch linear and 54ch spherical microphone arrays

Table 2: Recording conditions for impulse responses

A/D, D/A	Pavec MD-8000mk2 64ch 24bit
Microphone	54ch Spherical array 14ch Linear array 16ch Circle array
Source	Diatone DS-7 loud speaker B&K Type 4128 Head-Torso
Source Sounds	Time stretched pulse Balanced words(216) Balanced sentences: TIMIT SX(40), ATR(50)

Figure 3 shows the signal waveform and image data of the spectrogram waveform with the "pan" sample of the dry source (beating a handheld pan with a metal stick) as sample data of non-speech sound dry source.

3.2 Impulse Response Database

We collected impulse responses at different locations in different rooms. The sounds are recorded in an anechoic room, a variable reverberant room and offices using 3 kinds of microphone arrays by the Diatone DS-7 loud speaker and B&K Type4128 Head-Torso. Reverberation times of the rooms are 9 variations from 0.0 to 1.3 seconds. Table 2 shows recording conditions of impulse responses. Figure 4 shows a 14ch linear microphone array and a 54ch spherical microphone array used in the data collection.

Figure 5 shows a variable reverberant room whose re-

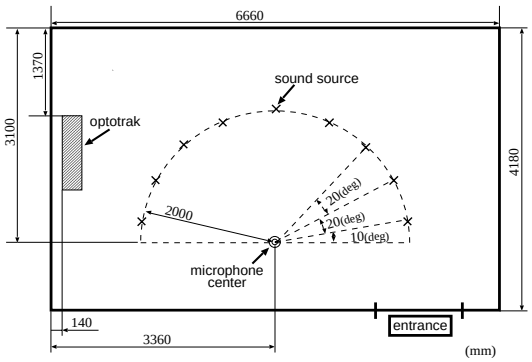
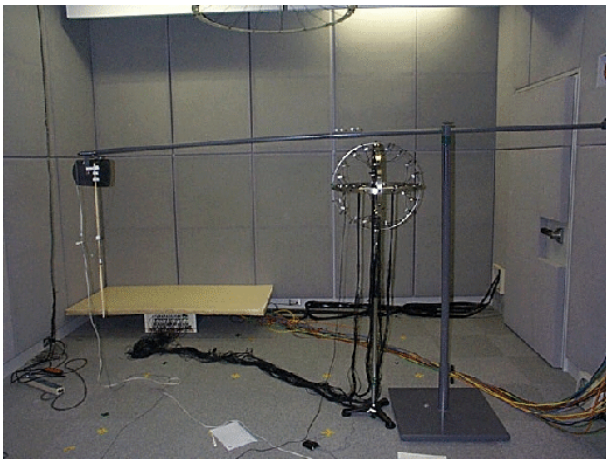
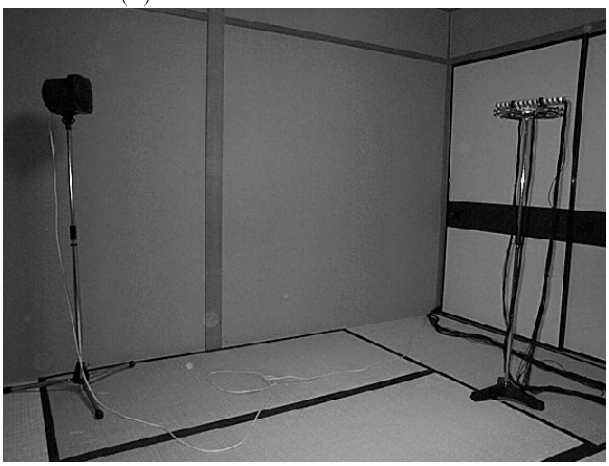


Figure 5: Data collection in a variable reverberation room

verberation time can be adjusted from 0.3 to 1.3 seconds by changing reflection walls. The impulse responses are measured from different angles from the sound source and a microphone. Also, Figure 6 shows the setup of these recordings and Figure 7 shows the impulse response waveforms which are measured in reverberant time 0.0, 0.3, and 1.3 sec. environments.



(a) Variable reverberation room



(b) Tatami - floored room

Figure 6: A setup of recording

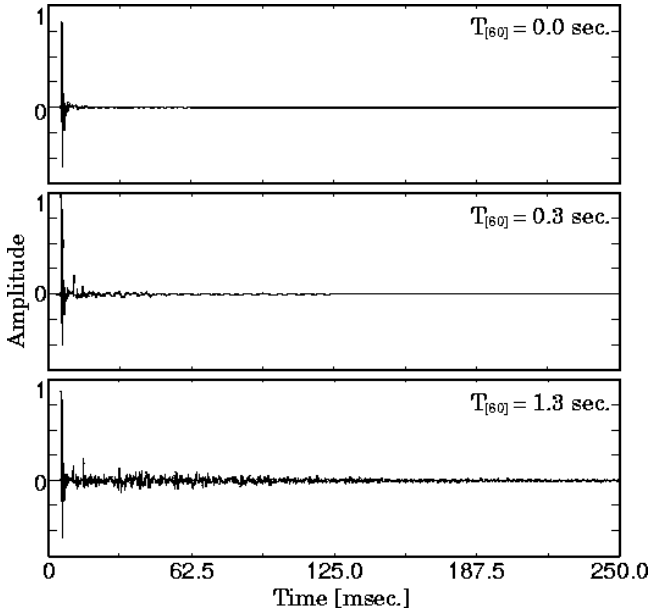


Figure 7: Impulse responses

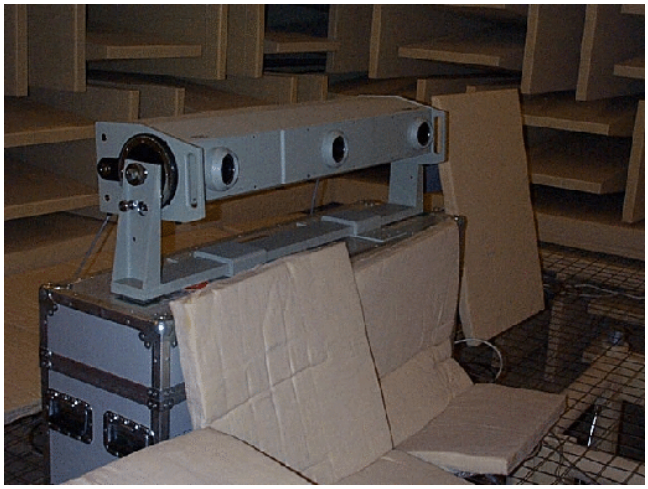


Figure 8: OPTOTRAK

3.3 Moving Sound Source

The sound in the real room can be simulated by convolution only if the transmission channel is linear and stationary. However, speakers may move while uttering in the real situation. We collected a moving sound source with respective position (x,y,z) simultaneously by OPTOTRAK. Figure 8 shows the OPTOTRAK. The OPTOTRAK is an infrared optical position sensing system with very high position resolution whose RMS resolution is 0.1mm. Phonetically balanced words and sentences (TIMIT SX sentence set and ATR balanced sentence set) are played through a loud speaker attached to moving sound system. Figure 9 and Figure 10 show the moving sound system we developed and an example moving sound source position trajectory for a sentence utterance. Also, Figure 11 shows the waveforms of captured signal in anechoic and reverberant environments.

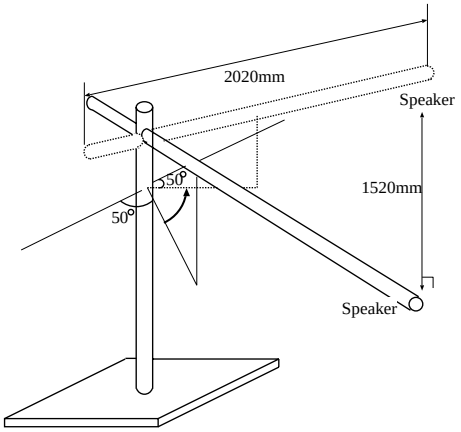
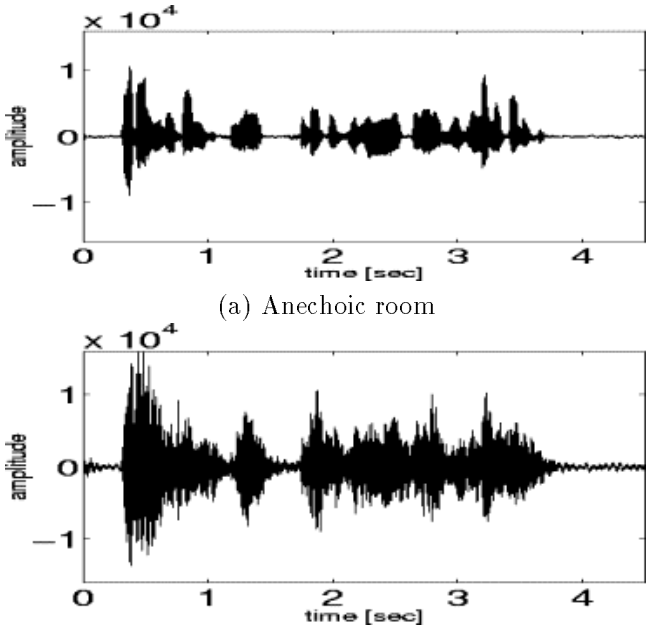


Figure 9: Equipment used for moving sound data collection



(a) Anechoic room

(b) Reverberant room

Figure 11: Waveforms of captured signal

4 Sound Source Identification Using a Microphone Array

We conducted a study of statistical sound source identification and speech recognition using a microphone array making use of a RWCP database. Sound source identification of "speech" and "non-speech" and speech recognition were carried out by the convolution of RWCP-DB impulse responses and a dry speech source of ATR-DB and a dry non-speech source of RWCP-DB.

4.1 Statistical sound source identification based on GMMs

Until now, a speech model alone was usually used for speech/non-speech segmentation [10] or identification. However, a single speech model has problems in

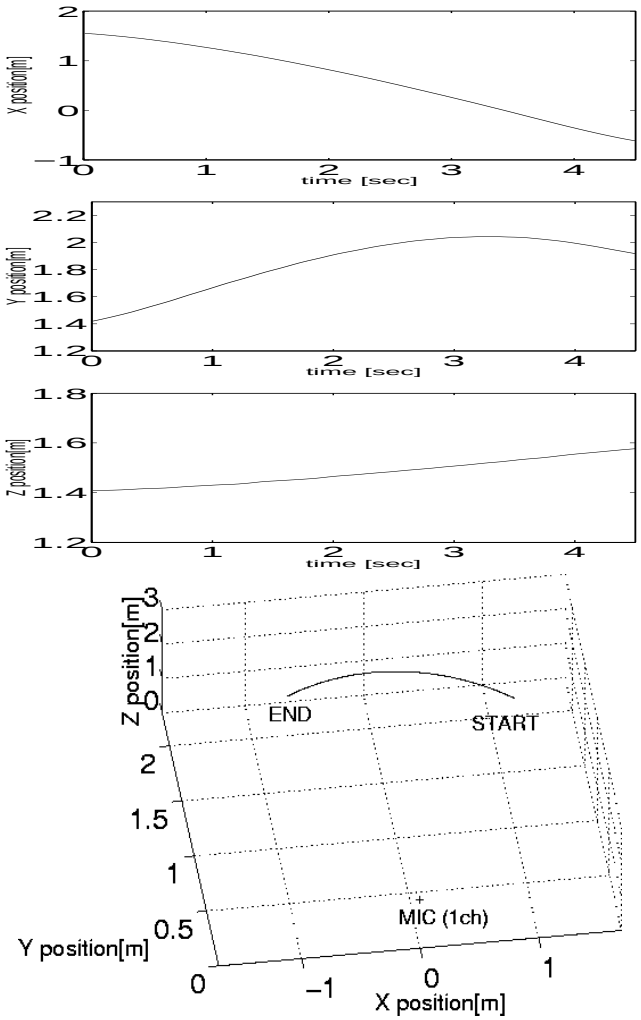


Figure 10: Position trajectory for the moving sound

that it not only requires a threshold to identify between "speech" and "non-speech", but also degrades the identification performance in noisy reverberant environments. To overcome these problems, we propose a new speech/non-speech identification algorithm that uses statistical speech and environmental sound GMMs (Gaussian Mixture Models). The multiple sound signals are identified by Equation (1).

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}} P(S(w)|\lambda_s, \lambda_n), \tag{1}$$

where $S(w)$ is the captured signal (frequency domain), λ_s represents the statistical speech model, and λ_n represents the statistical environmental sound model. The signals are identified as "speech" or "non-speech" by estimating the maximum likelihood in Equation (1).

4.2 Evaluation experiments

We first evaluate the basic performance of statistical sound modeling though environmental sound recognition. Then, the sound source identification performance is evaluated in distant-talking robust speech recognition.

4.3 Preliminary experiment: environmental sound recognition

Environmental sound recognition experiments are carried out with 20 samples for 92 kinds of environmental sounds and a single transducer in a clean environment. The feature vectors are MFCC, ΔMFCC, and Δpower. As a result, the environmental sound recognition performance is an average rate of 88.7% in multiple occurrence environments of the same sounds. This result confirms that the statistical modeling is very effective not only for speech recognition but also for environmental sound recognition.

4.4 Experimental conditions

The sound source identification performance is evaluated with known multiple sound source positions. Figure 12 shows the experimental environment. The desired signal comes from the front direction and white Gaussian noise comes from the right direction. The distance between the sound source and the microphone array is two meters. In this situation, the statistical sound source identification performance and ASR performance are evaluated subject to variations in the SNR (Signal to Noise Ratio) and the environment.

Table 3 shows experimental conditions for statistical sound source identification. We evaluate the statistical sound source identification performance using a single transducer and a microphone array, subject to SNR of -5dB, ∼, 30dB, and clean, and the reverberation times are $T_{[60]} = 0.0$, 0.3, and 1.3 sec. We also evaluate the ASR performance with the experimental conditions for ASR which are shown in Table 3.

In this paper, we evaluate the statistical sound source identification performance with 616 sounds consisting of speech (216 words × 2 subjects (1 female and 1 male)) and environmental sounds (92 sounds × 2 sets). The ASR performance is also evaluated with speech (216 words × 2 subjects). Equation (2) shows a definition of the sound source identification rate (SIR).

$$\mathrm{SIR} = \frac{\sum_{n=0}^{N} I_{cor}[n]}{N}, \quad I_{cor}[n] = \begin{cases} 1 & \widehat{Q[n]} = Q[n] \\ 0 & \widehat{Q[n]} \neq Q[n], \end{cases} \tag{2}$$

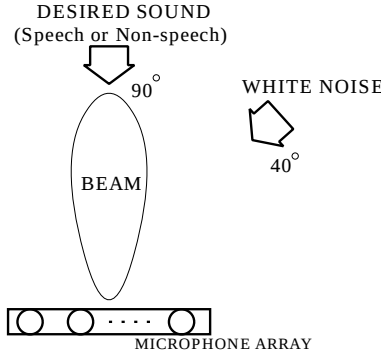


Figure 12: Experimental environment.

Table 3: Experimental conditions for sound sources identification

Frame length	32 msec. (Hamming window)
Frame interval	8 msec.
Feature vector	MFCC (16 orders, 4 mixtures), ΔMFCC (16 orders, 4 mixtures), Δpower (1 order, 2 mixtures)
Number of models	Speech: 1 model Non-speech: 1 model
Speech DB model training	ATR speech DB SetA [6] 150 words × 16 subjects (8 female and 8 male)
Non-speech DB model training	RWCP-DB 92 sounds × 20 sets
Test data (Open)	Speech: 216 words × 2 subjects (1 female and 1 male) Non-speech: 92 sounds × 2 sets

Table 4: Experimental conditions for ASR

Frame length	25 msec. (Hamming window)
Frame interval	10 msec.
Feature vector	MFCC, ΔMFCC, Δpower (+ CMS [11])
Test data (open)	216 words × 2 subjects

where $Q[n]$ is the correct answer, $\widehat{Q[n]}$ is the sound source identification result, and N is the number of all sounds. The ASR performance is also evaluated by the word recognition rate (WRR).

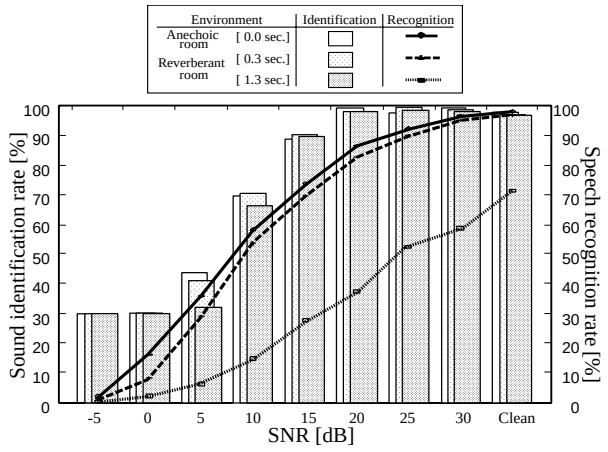
4.5 Experimental results

Figure 13 show experimental results using a single transducer and a microphone array that steers the directivity to the known desired sound source position (Delay-and-sum beamformer [12]). In these figures, the bar graphs represent sound source identification rates (SIR), and the line graphs represent word recognition rates (WRR).

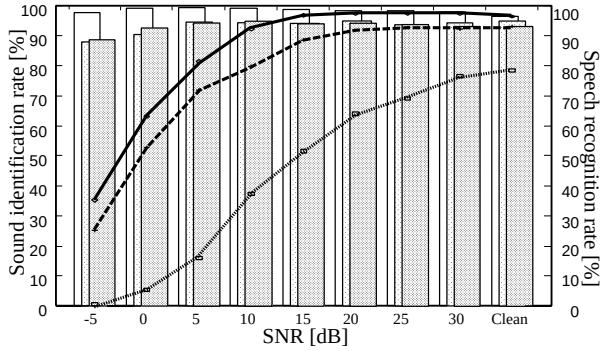
First, we focus the bar graphs in Figure 13. In these figures, by comparing the results using the single transducer and using the microphone array steering, we can confirm that the microphone array steering results give a higher sound source identification performance than the single transducer results especially in lower SNR environments. We therefore confirm that the proposed algorithm can achieve a higher sound source identification performance by using the microphone array steering.

Second, we describe the robustness against reverberation on the sound source identification. In Figure 13(b), the sound source identification performance using the microphone array steering is almost the same in each reverberant environment while the performance tends to decline slightly in the lower SNR and higher reverberant environments. With these results, we confirm that the proposed algorithm can distinguish "speech" or "non-speech" accurately in higher reverberant environments.

Next, we compare the proposed method with a conventional method using only speech GMM. Statistical sound source identification was carried out with the conventional method by distinguishing "speech" or "non-speech" using a threshold. Figure 14 shows the threshold estimation. As shown in the figure, we calculate accu-



(a) Single transducer



(b) Microphone array

Figure 13: Experimental results.

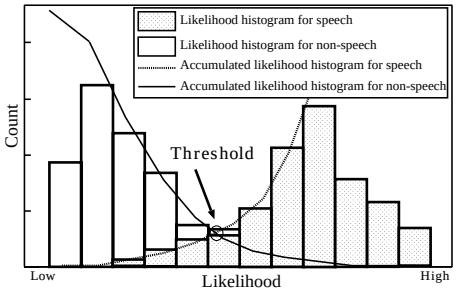


Figure 14: Threshold estimation by a conventional method.

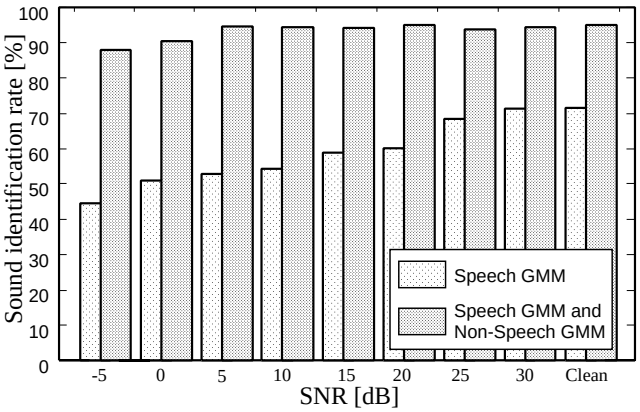


Figure 15: Comparison of sound source identification performance levels using the proposed method and conventional method.

mulated likelihood histograms with training data. Then, we estimate the threshold for the conventional method by finding the equal probability point with the accumulated likelihood histograms. Figure 15 shows results of the proposed method and the conventional method with microphone array steering in $T_{[60]}$ = 0.3sec. environments. The sound source identification rate (SIR) is only about 70% with the conventional method in the higher SNR environments, although the identification performance improves where the SNR is higher. However, the sound source identification rate is more than 90% with the proposed method not only in the higher SNR environments but also in the lower SNR environments, The performance of the conventional method using only speech GMM depends a lot on the threshold. However, the proposed method using speech and environmental sound GMMs can distinguish "speech" or "non-speech" accurately because its uses the difference of two GMM's likelihoods.

We also evaluate the relationship of the number of Gaussian mixtures for feature vectors MFCC and ΔMFCC and the sound source identification rate. Figure 16 shows the results. In the figure, we can confirm that the sound source identification performance is almost the same with more than four mixtures, while the performance degrades with less than two mixtures. Therefore, the proposed method may be able to distinguish "speech" or "non-speech" even if speech and environmental sound GMMs consist of few mixtures.

Finally, we focus on the line graphs showing the word recognition rates (WRR) in Figure 13(a)(b). In these figures, by comparing the results using the single transducer and using the microphone array steering, we can confirm that the microphone array steering results give a higher ASR performance especially in lower SNR environments than the single transducer results. As an example, we explain the ASR performance in the SNR = 10 dB environment. In the $T_{[60]}$ = 0.0 sec. and SNR = 10 dB environment, WRR is 58.3% for the single transducer. However, WRR improves from 58.3% to 92.6% using the microphone array. In addition, in the $T_{[60]}$ = 1.3 sec. and SNR = 20 dB environment, WRR is 37.1% for the single transducer. However, WRR improves from 37.1% to 64.3% using the microphone array. This confirms that the proposed algorithm using the microphone array results in a higher ASR performance than that using the single transducer not only in anechoic environments but also in reverberant environments.

According to the above evaluation experiments, we confirm that the talker can be localized accurately by sound source identification using statistical speech and environmental sound GMMs and microphone array steering among known multiple sound sources. We also confirm that the talker's speech can be recognized robustly with the microphone array in noisy reverberant environments.

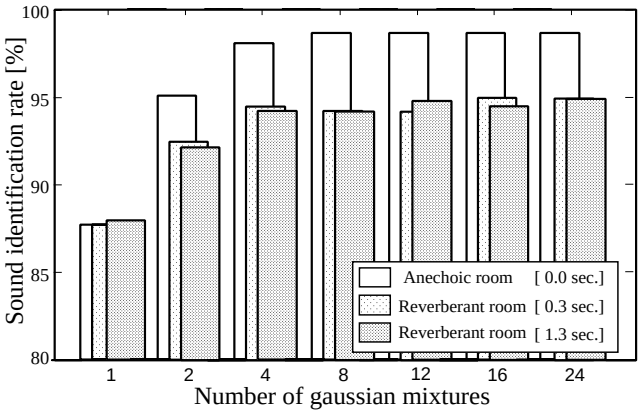


Figure 16: Relationship of the number of Gaussian mixtures for GMMs and sound source identification rate at SNR = 30 dB.

5 Conclusion

This paper describes sound scene data collection indispensable for studies of sound understanding including sound source localization, sound retrieval, sound recognition and speech recognition in real acoustic environments. We collected a dry source database and an impulse response database. Furthermore, we collected a moving sound data with the respective sound source position, since moving sound source can not be simulated by the convolution. Then we tried to identify "speech" or "non-speech" based on GMM and tried to recognize the distant talking speech in noisy reverberant environments.

The collected data is scheduled to be distributed freely for research purposes by three DVD-ROMs containing the acoustic dry sound source data, impulse responses, and sound position information. The information regarding to this database is summarized in the following URL:

http://tosa.mri.co.jp/sounddb/indexe.htm

The page also introduces our schedule of distribution and a way to get the DVDs. The URL includes not only the database specification but also theories and methods of measurement of an impulse response by TSP, estimation of reverberation time, sound source localization, and convolution. The application researches using the RWCP database are also described.

References

[1] M. Crawford, G. J. Brown, M. Cook, and P. Green, "Design, collection and analysis of multi-simultaneous-speaker corpus," *Proc. the Institute of Acoustics, Vol.16, Part 5*, pp.183–190, 1994.

[2] Q. Lin, C. Che, and J. French, "Description of the caip speech corpus," *Proc. ICSLP*, 1994.

[3] E. Jan, P. Svaizer, and J. Flanagan, "A database for microphone array experimentation," *Proc. Eurospeech*, 1995.

[4] D. GIuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of different microphone array configurations for hands-free speech recognition in noisy and reverberant environment," *Proc. Eurospeech*, 1997.

[5] S. Itahashi, "Recent speech database projects in japan," *Proc. ICSLP*, 1990.

[6] K. Takeda, Y. Sagisaka, S. Katagiri, and H. Kuwabara, "A japanese speech database for various kinds of research purposes," *Proc. ICSLP*, 1988.

[7] T. Kobayashi, S. Itahashi, and T. Takezawa, "Asj continuous speech corpus for research," *Journal of Acoustical Society of Japan*, pp. 48. 12. pp.888–893, 1992.

[8] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition," *Proc. Eurospeech99*, 1999.

[9] S. Nakamura, K. Hiyane, F. Asano, T.Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," *Proc. ELREC2000*, 2000.

[10] R. Singh, M.L. Seltzer, B. Raj, and R.M. Stern, "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination," *Proc. ICASSP2001*, May 2001.

[11] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, Vol. 55, pp. 1304–1312, 1974.

[12] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, Vol. 78, No. 5, pp. 1508–1518, Nov. 1985.

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-0317-10

早稲田大学ヒューマノイドプロジェクト
Humanoid Research in Waseda University

橋本周司
Shuji Hashimoto

早稲田大学ヒューマノイド研究所
Humanoid Robotics Institute, Waseda University
shuji@shalab.phys.waseda.ac.jp

Abstract

One of the ideal robots is a cohabitant robot that can work together with human partners in the human living environment. Such robot is a most suitable example to study a new relationship between human and machine. In these ten years, not a small number of project named “Humanoid” have started to develop human-like machines. Waseda University has a long history on the research of humanoid robot since the late-professor Kato started the project in 1970. This article introduces my considerations on the next generation human-machine interface and the past and current activities of the humanoid robotics research group in our university.

1. はじめに

人間中心の「人に優しいシステム」,「誰でも使えるシステム」,「察しの良いシステム」,「我々の創造力を最大限活かすシステム」を目指すならば，論理的ばかりでなく感性的な入出力も可能なインタフェースは必要不可欠である．従来，人間の論理的な活動としての知性を規範として研究されてきた情報処理が，感性という人間の情緒的側面も対象にするようになってきた．人間の心の動きは，五感により検知された外界情報により大きな影響を受け，これが感性的な反応として現れる．感性的情報は，主観的かつ多義的であり個人や状況への依存性が強く，多くの点で従来の情報処理の対象であった論理的な知識情報とは異なる．しかしながら我々人間の行動は，知識情報ばかりでなく感性情報によるところも多い．

一方，最近，国内外でヒューマノイドという名のロボット研究プロジェクトが，相次いでスタートとしている．それぞれのプロジェクトの内容は必ずしも同じでなく，アプローチ手法も異なっているが，最新の情報処理技術と機械技術を総合して，人間型ロボットを作ろうというところは，共通している．また，ペット型ロボットの開発も盛んであり，市販されるものも出て来た．このような新しいタイプのロボットは人間との共生をテーマとして，工場から我々の生活空間へと活動範囲を広げることを目指している．つまり，これまで第２次産業を中心として普及してきたロボットは，２１世紀に向けて第１次および第３次産業分野に展開・普及してゆくであろうことは大方の予想するところとなっている．特に高齢化社会の到来を目前にして，家庭内での家事の補助，身体の不自由な人や病人の介護などの支援，あるいは高齢者や子供の遊び相手などを目的としたパーソナルユースのロボットの登場には高い期待が寄せられており，ロボットにも思いやりや察しの良さなど感性的な能力が期待されている．

人間共存ロボットは，人間のために作られた環境において不特定の使用者と密着して作業するため，それに適した形態と機能を持つばかりでなく，特別な使用訓練を必要としない安全で柔軟なインタフェースを備えることが要求される．また，人間をサービス対象にするロボットは，インタフェースのマルチメディア化など，従来の自動作業機械というよりも情報機械としての側面を強く持つことになる．

このような情報機械システムの研究は，ソフトウエア主体の従来の人工知能に対して，実世界で行動する新しいタイプの人工知能への発展も期待される．本稿では，次世代インタフェースについての私見を述べ，早稲田大学ヒューマノイドプロジェクトの現状と目標を最近の成果と併せて紹介する[1][2][3].

2. メディア技術の発展とロボットの発展

2.1　メディア技術と人間

メディア技術は，従来からの人間社会の実環境をより柔軟に作り替えると同時に，ネットワークの中に仮想的な新しい環境を創り，我々に第２，第３の活動の場を提供する．我々は原初には環境のセンシングに用いていた五感を現在はコミュニケーションの受信装置として使用している（図１）．送信はほとんどの場合，筋肉系による環境の物理的な操作により行われる．したがって，身体的なインタフェース技術は，メディア技術によって実現される情報的な仮想環境においても，メディア技術の重要な要素となる．

また，人間中心の技術はリアリティのある技術あると言うことができる．ＶＲの中心課題となる物理レベルのリアリティは，外界とシステムのインタフェースを透明化して違和感を消去する．ＡＩの中心課題である論理レベルのリアリティは人間の知的な違和感を消去する．次世代技術ではこれらに加えて，特に感性のレベルのリアリティが重要となる（表１）．これが環境の”居心地”を決定すると同時に，その環境におかれた人間の積極性と創造性を刺激するからである．これまで情報技術が深く踏み込んでこなかった人間の感性の取り扱いが次世代技術の新しい課題となるのである．

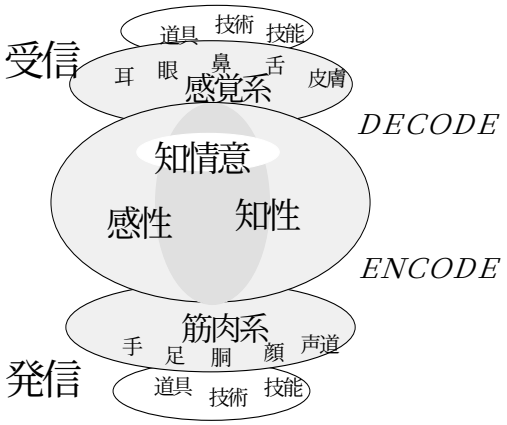


図１　人間のコミュニケーション方式

表１　リアリティの階層

階層	チャンネル	支配則	分野	ﾘｱﾘﾃｨの尺度
波形レベル	光，音，力	物理法則	人工現実感	説明可能性 因果的無矛盾
意味レベル	言語，ｼﾝﾎﾞﾙ 図形，数式	論理，文法	人工知能	証明可能性 論理的無矛盾
感性レベル	音楽，絵画 表情，仕草	主観，共有性 快不快	感性情報処理	共鳴可能性 合成的無矛盾

2.2　オーディオ・ビジュアルを超えて

我々の感性情報には，視覚，聴覚など感覚に依存する部分と，依存しない部分がある．現在のメディア技術は主として，オーディオ・ビジュアルを中心としており，最近触覚．力覚関連のメディア技術が開発されつつある．これらは，計算機のディジタル処理をベースとするエレクトロニクスとメカトロニクスの応用であり，通信ネットワークとの整合性も高い．今のところ，電子メディア技術は嗅覚と味覚には及んでおらず，香りや味の情報は視覚などの情報により引き出される２次的なものとなっているが，より豊かな感性的な情報環境を造るためには，これらも考慮される必要があろう．論理的なレベルではオーディオ・ビジュアル情報で十分であろうが，感性レベルでは触覚・嗅覚・味覚の情報は無視できないものである．感性情報工学で

は，ディジタル化し易い物理的なメディアに加えて，化学的なメディアも視野に入れておく必要があると考えられる．それは，情報ネットワークの中の仮想環境と実環境に引き裂かれることのないメディア環境を実現し，電子情報と体感の融合をはかるためにも重要である．

2.3　人間共生ロボット

人間のために作られた生活環境において人間と共に行動するには，ロボットがそれに適した形態と機構を持つばかりでなく，人間との間に高度な意志の疎通を可能とするインタフェースが不可欠となる．これを実現する一つの方法は，より人間に近い思考形態と行動形態をロボットに持たせることである．つまり，ロボットは，行動空間ばかりでなく情報空間も人間と共有する必要がある．

行動空間の共有に関しては，ロボットの機構を人間環境に整合させるために出来るだけ人間や家庭で飼われる動物と同じ形態にするという機構上の課題の他に，人間との接触を前提とした安全性と動作の柔軟性の確保という制御上の課題がある．情報空間の共有に関しては，キーボードやスイッチなどによらず，音声，画像，触覚などのチャンネルによるマルチモーダルインタフェースの開発と，質問して使用者の意図を汲み取る自発的コミュニケーションを実現する能動的な思考能力の開発が課題である．　人間同士の対話の場合は，ジェスチャや顔の表情によって，言語化し難い情報が伝わる．したがって，ロボットがこれらを理解し，同じように自己表現できることは，人間－ロボットの円滑な協調に有益であると考えられる．したがって，次世代ロボットは，従来の自動作業機械というよりも情報機械としての側面を強く持つことになる．

3. 早稲田大学ヒューマノイド

3.1 プロジェクトの歴史

生活空間の中で人間と共生するロボットの一つの理想形は人間そのものであり，人間と機械との新しい関係を考える上でも，ロボットは格好の題材である．早稲田大学理工学部では，学科横断プロジェクトとして，故加藤一郎教授を中心に３０年程前から断続的に人間型ロボットの開発研究が行われてきた．この間ほぼ１０年毎に個別要素技術を統合した人間型ロボットを製作してきた．具体的なトータルシステムの第１号は，音声対話により指示を受けて，室内の特定物体を発見し２足歩行により接近し，両腕で把握・運搬するロボットＷＡＢＯＴ－１（1973 年）である（図２－ａ）．また，第２号のＷＡＢＯＴ－２（1984 年）は筑波科学博の政府館に出品されたが，音声で会話をすると同時に，市販の楽譜を認識し歌声に合わせて鍵盤楽器を１０本の指と両足によって演奏する音楽ロボットであった（図２－ｂ）．

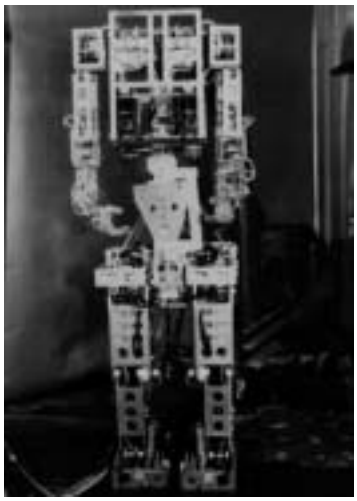


図 2-a　ＷＡＢＯＴ－１

図 2-b　ＷＡＢＯＴ－２

現在の早稲田大学ヒューマノイドプロジェクトは，これらの発展形として，より人間に近い振る舞いをし，人間との真の共生を目指すロボットの開発を目指して，機械工学科，電気電子情報工学科，情報学科，応用物理学科の研究室が合同で推進してきたものであり，昨年度から早稲田大学ヒューマノイド研究所という組織になった．

本プロジェクトは，１９９２年にスタートし，視覚，頭脳，機械，対話に関する要素技術の研究を行なうと共に，１９９５年には案内ロボット Hadaly-1（図３），１９９７年には共同作業ロボット Hadaly-2（図４）および２足歩行ロボット WABIAN（図５）を製作した．

図３Hadaly－1

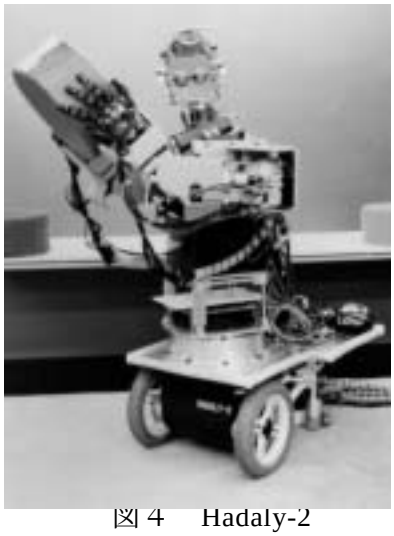


図４　Hadaly-2

Hadaly-1 は，移動はできないが音声系と視覚系および簡単なジェスチャによって人間と対話し，校内案内を行なうロボットである．利用者の方を向いて話をして目的の方向を腕で指し示すなど，通常のディスプレイによる案内に比べて直感的に判りやすいインタフェースを人間型ロボットを用いて実現した．

Hadaly-2 は車輪走行であるが，掴む柔軟な手と腕を持ち，眼球も動く頭部を持っている．情報処理能力も高度であり，自分の位置を周りの景色から判断することができ，室内に入ってきた人間を視覚と聴覚で探し，握手をしてパンフレットを渡したり，顔を見てジェスチャを交えて対話しながら積み木遊び等の共同作業を行うことができる．ロボット本体は全高１７０ｃｍ，総重量１５０ｋｇ，機械系の自由度は，眼球部２ｘ２，首部２，胴体１，腕部７ｘ２，ハンド１３ｘ２，車輪２の合計４９自由度で，柔らかい関節を持つことにより柔軟な動作と安全性を確保している．

WABIAN は人間と同程度のサイズ（約１６６ｃｍ，重量１０７ｋｇ）で，動的なバランスを取った２足歩行が可能である．頭部には，Hadaly-2 と同じく人間を追跡する視覚系が搭載され，簡単なジェスチャ認識も行う．また，柔らかい上半身を音楽に合わてスイングしてダンスを踊ることができる．さらに，通信回線に接続することにより，遠隔地からの制御も可能である．機械系の自由度配置は，眼球部２ｘ２，首部２，腕部７ｘ２，ハンド部３ｘ２，胴体３，２足歩行を行う下肢機構部には，足関節，膝関節，そして股関節にピッチ軸回りの自由度を，足底部に受動的自由度（各 4 自由度 X，Z 軸方向，ピッチ，ロール軸回り）を持ち，合計で 43 自由度である．股関節については，位置エネルギーを利用した効率の良い歩行を可能とするために，人間の駆動方式と同様な拮抗駆動方式を採用し，幅広い範囲で関節の剛性を変えることもできるようになっている．

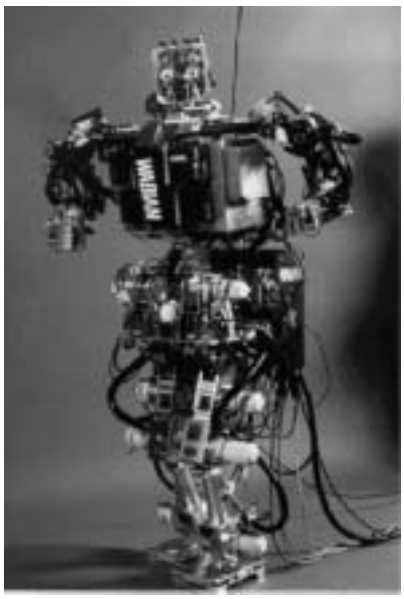


図５　WABIAN

3.2　ヒューマノイドプロジェクトの目標

我々のヒューマノイドプロジェクトでは，作業機械としてばかりでなく情報機械としてのロボットの新しい側面に注目している．我々が想定している次世代ロボットの応用分野は，共同作業ロボット，家庭内作業ロボット，高齢者・障害者のための介護ロボット，話相手ロボット，マルチメディア情報端末ロボット，機器管理サポートロボット，ロボットアクター，など生産現場からアミューズメントまで多岐にわたるが，主たる研究課題は，いずれも人間型機能をベースとした人間と機械のインタフェースに係わるものである．

従来のロボットは，作業機械としての人間を模倣あるいは人間の作業能力を強化するものであった．ヒューマノイドプロジェクトの第１の目標である人間との共同作業はこの延長上にあるが，工場での作業ロボットに比べて，人間との相互作用がはるかに密に行われる．定型的なコマンドに従って作業を行うのではなく，相手の人間に応じた柔軟な行動の変更が必要なためである．ロボットが作業環境である室内の壁や配置物などの状況を判断することはもちろん重要であるが，その中で動き回る人間を発見しその意思を理解することも共同作業においては不可欠である．特に，家庭内での共同作業においては，一般ユーザにロボット操作に関するリテラシーは期待できない．したがって，普通の人間と共同作業をしているような感覚でロボットを操作できるインタフェースとコミュニケーション能力がロボットに必要とされる．

ヒューマノイドプロジェクトでは第２の目標として，情報端末としてのロボットの利用を検討している．現在のコンピュータ端末は，グラフィックスとポインティングデバイスによるＧＵＩが主流となっているが，我々はロボットそのものを端末とすることを考えている．ネットワークにロボットを接続して，画像，音声，表情，身振り，触覚など人間のあらゆる知覚チャンネルを使って，情報を引き出したり，計算機を使ったりするのである．そこでは，ロボットの手を握ること，嬉しそうに話すこと，画像を見せること，などが計算機システムへの入力であり，ロボットの仕草，ロボットの表情，ロボットが我々の肩をたたくこと，などが計算機の出力である．最近の自動車には多くのコンピュータが組み込まれており，運転者は意識せずにそれらを使っている．マルチメディア情報端末としてのロボットは，ＧＵＩとは異なった計算機の利用形態を提供し，計算機をリアルワールドで体感し”ドライブ”することを可能にする．また，この延長として，ネットワークを介した遠隔制御やテレイグジスタンスとは一味異なった，自律性を持った実体エージェントとも言うべき応用も考えられる．遠隔地の人間に依頼するの同様の手続きでロボットに作業をさせるのである．

次に，センサ集合体としてのロボットの利用が考えられる．ロボットは多くのセンサを装備した動く情報収集システムであるから，人間型でなくとも環境を自律的に動き回ることにより情報収集を能動的に行なうことができる．また，人間型であれば人間のシミュレーションが可能である．例えば，ロボットに衣服を着せて着心地を確かめる．ロボットは腕や首を動かしたり，歩き回ってデザインに不都合は無いかを確認するのである．このようにマルチモーダルかつ能動的センサ系としてロボットを使って，人間が使用する製品や生活環境の評価を行うことは現在でもある程度可能である．ちょっとした床の段差にも難渋する２足歩行系，細かなパターンは識別

できない視覚系，あるいは少しの雑音があっても誤りを犯す音声認識系を持つ現在のロボットが活動するのに不自由のない環境は，年老いて身体機能の衰えた人間にとっても快適なはずである．

最後に，人間の生活空間に存在するロボットの用途の一つに「癒し機械」が考えられる．現在，開発されているペット型ロボットの多くは実際の作業を行なうには不充分であるが，ユーザとのインタラクションによって感性的な満足感を与えるにはある程度の有効性がある．また，現在のＴＶディスプレイを用いたコンピュータゲームに比べて，表現手段としての身体を持ったゲーム機の可能性は非常に大きい．このようなロボットが情報家電や家庭内 LAN と結びついて，エンターテイメント性を持ちながら機器操作のアシスト機能を果たす日が来るのはそう遠くないと思われる．

4. プロジェクトの現状

本プロジェクトは，現在，本部キャンパスの近くに新たに建設された理工総研ハイテクリサーチセンターおよび大久保キャンパスに研究スペースを持ち，６つの研究グループで約５０名の大学院生が１５名の専任および客員教員とともにそれぞれの課題に取り組んでいる．研究は，早稲田大学および文部省などの公的資金ばかりでなくコンソーシアム参加企業からの支援によって支えられている．また，国内ばかりでなく海外のいくつかの研究機関との協力関係もあり，公式，非公式の研究会も開催している．2001 年１１月にはヒューマノイドロボット国際シンポジウム（HUMANOIDS2001）を開催する予定である．

現在は，個別技術開発のフェーズであり，約３０の個別研究と数年後の統合システムへ向けた準備を行なっている．以下に最近の成果をいくつか紹介する．

１） 能動的視覚系

筆者のグループはロボットビジョンと感性インタフェースをテーマとしているいるが，ロボットビジョンの研究では，ロボットが環境内での自分の位置と方向をシーンの画像処理により特定するためのモデルベースビジョンシステムと環境モデルの自動作製を行うためのアクティブビジョンシステムを製作し，小型移動ロボットにより自発的にランドマークを設定して必要な画像を収集し，その結果を用いて環境地図の自動作成と自己位置認識を行なうことに成功した（図６）．

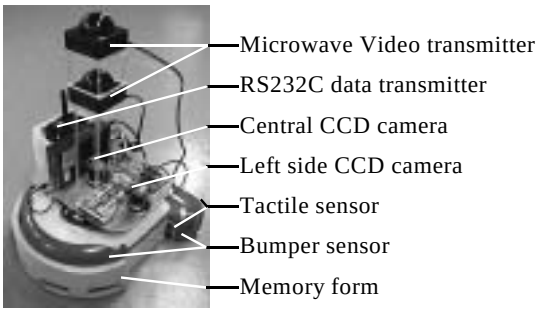


図６　能動的視覚ロボット

２） バーチャルヒューマノイド

成田誠之助教授のグループは，サイバーロボティックスを主なテーマとしているが，コンピュータ内に作成した仮想空間内のバーチャルロボットを，ネットワーク経由で遠隔操作し，ハイテクリサーチセンター内の共同研究室において，3D ステレオ表示および通信時間の遅延がある中での動作確認を行った（図７）．また，外部からインターネットを通じてバーチャルロボットの遠隔操作の実験も行なっている．

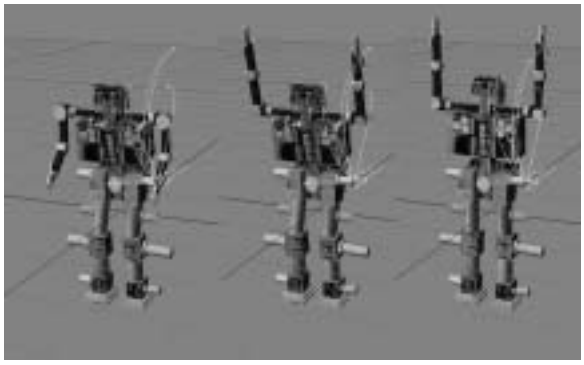


図７　バーチャルロボット

３）人間共存ロボット

菅野重樹教授のグループは，拘束条件に応じて作業性と安全性を維持するための腕部・体幹部運動制御則と双腕表皮カバー設計論を提案し，それらを人間共存ロボット WENDY

に実装して，作業性と安全性が両立できることを確認した（図８）．さらに，WENDY のサブシステム間のデータ共有を可能とする統合 OS を開発し，それを利用して家庭内作業の一例である調理（卵を割る，野菜を切る）のデモンストレーションを行なった．

４）　情動表出ロボット

高西淳夫教授のグループでは，人間とロボットの情緒豊かなコミュニケーションを目指して，首部までを含めた頭部ロボットＷＥ‐３ＲＩＶを開発している．これは，４自由度の首と４自由度の眼球で指標の動きに追従するばかりでなく，触覚，聴覚，視覚，嗅覚センサを備えて，外部環境情報によって，心理状態を動的に変化させて，眉，唇，瞼の動きや顔色の変化によって感情を表出することができる（図９）．

図８　人間共存ロボット　WENDY

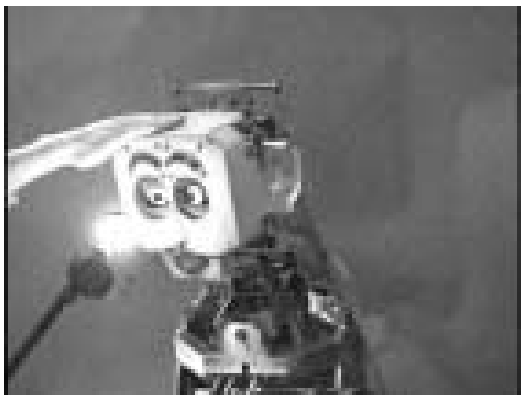


図９　頭部ロボットＷＥ‐３ＲＩＶ

５）　対話ロボット

白井克彦教授とともに対話系を主に担当する小林哲則教授のグループでは，ジェスチャ認識と顔方向認識による利用者の指示動作と注視方向の理解，指示語を含む音声言語の理解などにより，話題対象についての共有感を作りながら対話のできるロボット ROBITA を開発した（図１０）．また，うなずきや表情などの動作をロボットに実装することによって，従来ジェスチャやアイコンタクトだけでは伝えきれなかった内部状態をユーザに伝えることができるようになった．

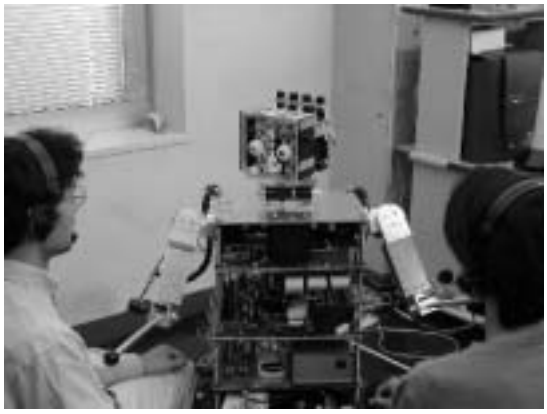


図１０　対話ロボット　ROBITA

６）力覚インタフェース

筆者のグループでは，ロボットの機能の一部を使った力覚インタフェースとして，いくつかの試みも行っている．これらは皆，新しいコミュニケーションチャンネルとして，力あるいは接触感覚を用いようとするものである．（図１１，図１２）

図１１　力覚インタフェース１

(A) 触覚・力覚情報を用いた遠隔地とのコミュニケーション

(B) 電話回線を通じた遠離握手システム

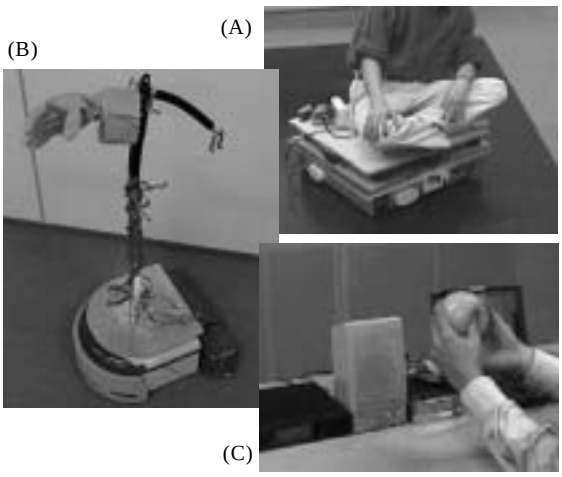


図１２　力覚インタフェース２

(A) 重心移動インターフェイスと全方向移動ロボット
(B) 人間協調移動ロボットのための力覚インタフェース
(C) 握り動作による音楽生成システム

以上の他にも，人間と音楽パフォーマンスを行なうロボットの開発や感性的な歩行やＳ字歩行が可能な２足ロボットなど，多くの成果が出つつある．これらの多くは早稲田大学ヒューマノイド研究所のホームページに公開されているので，参照されたい．

（http://www.humanoid.waseda.ac.jp）.

5. おわりに

感性工学にかかわる最近の話題と早稲田大学におけるヒューマノイド研究について述べた．人間の感性を意識したメディア技術が重要なのは，単に”心地よさ”に関係するためだけではない．感性メディア技術には，さらに積極的な役割がある．それは人間のメンタルな部分を適度に刺激して，創造性や積極性を引き出すためにも重要である．

ロボットはこのような用途においても大きな意味がある．ロボットが工場から出て家庭にやってくる状況は、まずはペットロボットから実現が始まっているが、我々の当面の目標は、「あまり役に立たなくてもよいが、側にいて欲しいロボット」、「察し良くマルチモーダルな対話ができるロボット」、「我々の動作や仕草のほとんどが理解でき真似られるロボット」の実現である．

我々のプロジェクトの特徴は，機械系と情報系の研究者が分野を超えて密に共同研究を行なっているというところにあり，他にあまり例を見ないものである．現在進行中の第３期プロジェクトでは，ヒューマンインタフェースも含めた大きな問題として人間共存ロボットを捉え，広く社会との連携を密にして議論を深めてゆきたいと考えている．

故加藤教授らが１９７０年代に２足歩行の人間型ロボットを開発した頃には，「面倒な２足歩行が本当に意味があるのか．車輪や他の解決手段があるではないか.」などと言われたと聞いている．実用面での可能性はほとんど見えない状態であり，むしろ人間型ロボットを作ることによって人間を研究するという科学としての意味合いが強かった．３０年後の現在は人間生活における科学技術の役割が大きく変わりつつあり，効率重視から居心地重視の人間中心主義の重要性が言われている．また，情報処理も論理的なデータばかりでなく人間の感性に係わる問題を扱い始めている．これまでの科学技術は，時間的，空間的あるいは知的な人間の限界を乗り越えることを目指してきたことを考えると，まさにスーパーマンからヒューマンへの転換期にある．

ヒューマノイド研究も科学的興味ばかりでなく産業的な意義を持ち得る時代である．より人間の生活に密着した働きをするロボットを考えると技術的な問題ばかりでなく，生き物のような機械が人間に与える心理的なストレスや事故責任に関する法整備などの社会的な問題を解決する必要がある．我々技術者もこれらについて考えると同時に，技術の現状をできるだけ広く公開して，より良い解を探すための議論の環境を準備することが重要である．

参考文献

（詳しい文献は以下を参照下さい．）

1] 橋本 “感性情報処理の諸相”，映像情報メディア学会誌，52 巻，1 号，pp.41-45，1998

2] 橋本，他，“ヒューマノイドー人間形高度情報処理ロボットー”，情報処理，38 巻，11 号，pp.959－969，1997

3] 早稲田大学ヒューマノイドプロジェクト編，“人間型ロボットの話”，日刊工業新聞社，1999

〒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.)

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