人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-0318

ＡＩチャレンジ研究会 (第１８回)

Proceedings of the 18th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2003 年 11 月 13 日　　場　所　京都大学工学部　8 号館 中会議室

Kyoto University, Nov. 13, 2003

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

共催 社団法人日本ロボット学会 ロボット聴覚研究専門委員会
Robotics Society of Japan, Special Interest Group on Robot Audition

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-0318-1 (11/13)

ROBISUKE: 新世代の対話ロボット

ROBISUKE: A New-Generation Conversation Robot

小林哲則
Tetsunori KOBAYASHI
早稲田大学
Waseda University
koba@tk.elec.waseda.ac.jp

Abstract

ROBISUKE, the 3rd generation robot, can estimate the mental states of conversational partner using the intonation of the utterances and the head gestures even if the partner does not express his idea explicitly with words. Even if ROBISUKE happens to meet un-known expressions, he can extract the meanings out of the partner through the conversation. After this learning process, he can behave appropriately according to the expressions. These functions are very important to realize rhythmical conversations and avoid the monotonous conversations. In this paper, we introduce these functions of ROBISUKE.

1 はじめに

人間共生型ロボットの実現が期待される中で，その必須要素技術としてのロボット対話に関する研究が活発化している。多くの研究者が，様々な切り口で，ロボットにおける音声対話インタフェースの問題に取組んでいる [橋本, 1997a][Imai, 2001][岩沢, 2002][Nisimura, 2002] [Nakadai, 2003]。

筆者は，これら対話ロボットを，備える機能と実現される対話の質との関係から，おおよそ 3 世代に分類している。第 1 世代は，音声によってロボットに対する指令を伝えることを主な目的としたもの。第 2 世代は，ロボットの身体表現を言語表現と協調させる機能を持つもの。第 3 世代は，これらに加え，知覚情報処理による対話状況把握機能を持って，さらに円滑なコミュニケーションを実現するものである。

70 年代から 80 年代にかけて開発された，WABOT, WABOT-2 は第 1 世代の代表的対話ロボットである [Fujisawa, 1974] [白井, 1985]。当時，音声認識装置自体がめずらしい中で，音声でコマンドを入力できる画期的なシステムであった。しかしながら，音声認識装置とロボットとの関係についてみると，両者はほぼ独立のものといって差支えなく，ロボットが会話するということ固有の問題には踏み込んでいなかった。このため,「一緒に対話をしている」という対話相手との一体感を感じられるものではなかった。安価な対話ロボットは，現在のものでもこの世代に分類されるものが多い。

1990 年代の半ばに開発された，Hadaly，Hadaly2 などは第 2 世代の代表である [Hashimoto, 1997b][Hashimoto, 2002]。アイコンタクトや，空間表現動作，象徴的動作を交えた会話は，Hadaly で扱われたテーマである。この世代のロボットが身体表現に用いたものは，腕や頭部動作，視線制御程度のものに限られ，表現は大雑把なものであったが，対話相手との一体感は格段に改善した。最近研究が進む会話ロボットの多くは，この世代に分類される。

ROBITA は，第 3 世代初期の対話ロボットである。第 2 世代で開発された機能に加えて，対話相手の顔向きに代表される会話の状況に関する視覚的理解能力を持つことで，グループ会話を実現することができた [松坂, 2001]。グループ会話とは，複数の参加者が対等の関係で対話を行う状況をいい，誰が（全員に対し話しかけることも含めて）誰に話しかけることも許される画期的なロボットである。（多人数から一人選んだ上で，一対一の対話をするものとは本質的に異なることに注意を要する。）

ここで紹介する ROBISUKE は，ROBITA の後継であり，特にパラ言語の理解能力を向上させることと，対話内容を充実させることを目標としている。パラ言語とは，言語情報の伝達行為に付随して生じる言語以外の情報であって，言語情報の円滑な伝達を支えるために機能する情報をいう。パラ言語によって，対話調整的な情報や，言語情報では表しきれない発話者の心情などが運ばれ，豊

かなコミュニケーションの成立基盤となる。対話内容の充実に関しては，対話相手が使う新たな言い回しに対する対応規則の獲得や，発話内容に関する話題の獲得を行う。これにより，深みのある対話が実現可能となる。

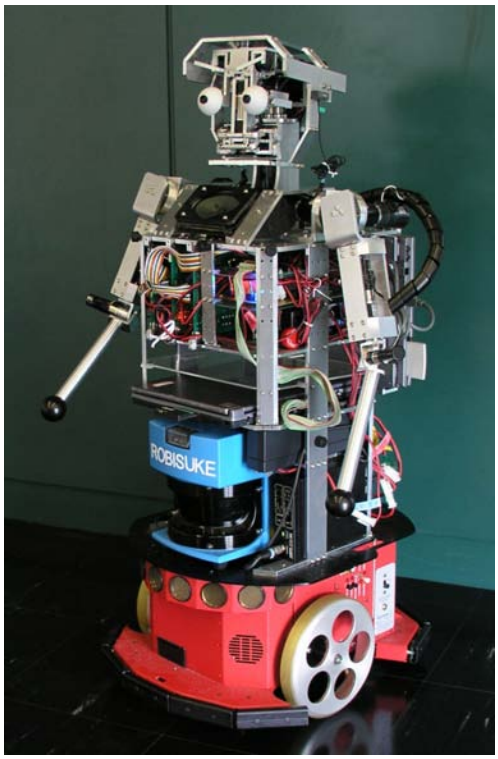


Figure 1: ROBISUKE

2　パラ言語コミュニケーション

会話とは，そもそも言語情報の伝達だけによって実現されるものではない。音声言語の伝達行為に付随して声質，表情，身振りなどに発話者の態度，状態などが現れ，これが伝達・受理されることによってはじめて円滑な会話が成立する。言語情報の伝達行為に付随して生じる言語以外の情報であって，言語情報の伝達を支えるために機能する情報は，パラ言語情報と呼ばれる。パラ言語情報とは，マルチモーダルな性格を持つ。即ち，これを扱う対話システムは，音声チャネルでのパラ言語の表出・受理機能が必要とされるだけでなく，視覚チャネルでの表出・受理機能が必要となる。ロボットの身体は表出機構として，視覚システムはその受理機構として相応しく，ロボットはそもそも対話システムとしてうってつけのシステムといえる。会話におけるパラ言語の主な役割は，発話の番の交代に関連した対話調整的機能と，言語内容に付加的意味を添える機能などがある。

2.1　身体表現によるパラ言語情報の表出

人間型ロボットを用いて音声会話システムを実現することの利点は，第一に透過性の向上にある。機械内部がどのような処理状態にあるかを利用者にフィードバックし透過性を高めることは，インタフェースとしての必須の条件であるが，元来音声チャネルはこの目的に不向きであり，視覚チャネルを援用することが必要となる。ROBISUKE では，人間の会話時における顔を模した表情によって透過性を実現する。例えば相手の目を見つめることによって，聞く準備ができていることを表現する。また相手の発話を理解できなかったことを，怪訝な表情をすることで伝える。これらの行為は前身の ROBITA で既に実装されたものであるが [Tojo, 2000]，利用者に対する自然な発話要求になっており，会話を滞りなく進めることに役立つ。人間の表情を模したことは重要な意味を持つ。表情による情報は通常意識下で伝えられるものと考えられるが，このことを考慮すると，人間－機械の音声対話システムを実現する場合，機械側も人間と同じような手段を用いて情報を送らなければ，人間の意識下での情報処理機構に訴えることは期待できないからである。

2.2　視覚情報処理によるパラ言語情報の受理

前節では，生成の観点から透過性の問題について述べた。同様に人間も透過性向上のためのシグナルを身体表現によって送っているのであるから，ロボットにもこれを理解する機能を実装する必要がある。我々は，発話の受理状態を対話相手にフィードバックする動作として，うなずき，首振り，かしげの 3 動作を選定し，これらの認識を行うシステムを実現した。頭部画像領域を上下左右に 4 分割し，それぞれの部分の上下方向，左右方向のオプティカルフローを算出することよってできる 8 次元のベクトルを特徴量として，頭部動作のスポッティングを行う HMM ベースのシステムを実現した。ここでは特に，ロボット自身が動くことによる画像の乱れへの対処が必要となる。この問題を，MLLR によるモデル適応によって解決した [江尻, 2003]。これらにより，発話が相手に伝わったかどうか，発言が支持されたかどうかを確認しながら会話を進めることが可能になっている。

Table 1 は，自由対話 60 分を含む，ジェスチャデータ 204 分に現れる合計 2148 個の頭部ジェスチャの認識結果である。

Table 1: 頭部ジェスチャ認識の実験結果

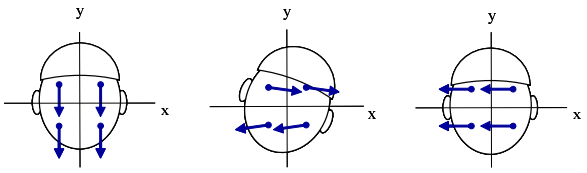
	認識結果			
	うなずき	かしげ	首振り	脱落誤り
うなずき	**1144**	6	0	233
かしげ	9	**322**	4	189
首振り	2	2	**290**	18
挿入誤り	347	157	42	

(a) 入力画像

(b) 色モデルによる抽出

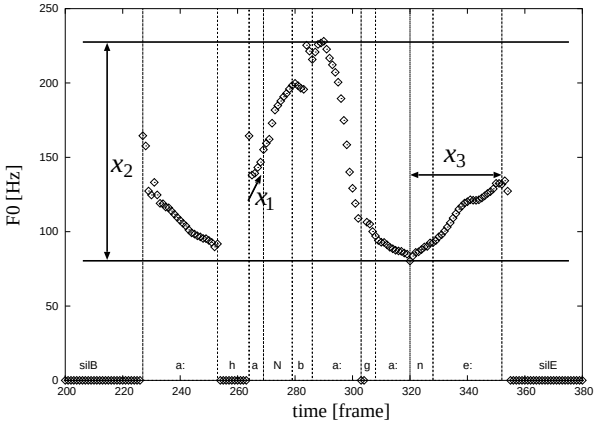
(c) 首領域の除去

Figure 2: 頭部の抽出

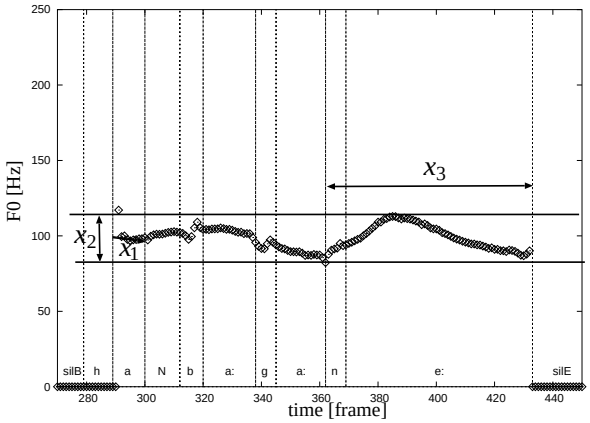


(a) うなずき (b) かしげ (c) 首振り

Figure 3: 各ジェスチャの特徴的なフロー



(a) 肯定的



(b) 否定的

Figure 4: 特徴量抽出の例

2.3 韻律処理によるパラ言語情報の受理

発話者は韻律を用いて心情を表現することがある。韻律に現れるパラ言語として代表的なものとしては，相手の発話に対する，賛同の程度が挙げられる。

我々が行っている昼食相談タスク（お昼をどこでとるかを相談者が提案者に相談するタスク）における予備実験によれば，例えば「カレーなんてどうですか」という提案に対し，相談者が明示的に言語情報で受理／拒否を伝えることは少ない。むしろ,「カレーか」と復唱するだけのことが圧倒的に多い。しかし，ほとんどの場合，この復唱の抑揚の中に，提案に対する相談者の受理／拒否の態度が込められており，それを受けて提案者はかなりの確率で提案を修正すべきかどうかを判断できる。

Figure 4 は,「ハンバーガーね」を肯定的（提案を受理する立場）で発話したときと，否定的（提案を拒否する立場）で発話したときの韻律パターンを示している。図に見られるように，肯定的な場合は F0 パターンのダイナミックレンジが広く，否定的な場合は，狭い。

我々は，F0 ダイナミックレンジの他に，最終モーラの継続長，語頭の F0 の傾きの 3 つ組を特徴量として用い，確率モデルとして GMM を用いて認識器を構成した。Figure 5 は，20 人分計 2000 発話のデータに対し，発話が肯定的か，否定的かを識別した結果である。発話内容には,「○○か」「○○か」「○○」のような復唱の他,「そうだね」「いいんじゃない」の各々を肯定的および否定的に発話したものが含まれている。

識別結果は混合数 16 時の 82.9%が最大である. これは同じ発話を人間が聞いて判断したときの認識率とほぼ同等である。また，Table 2 は，同じタスクにおける人間同士の判断の一致率 κ と機械と人間の判断の一致率を比べたものであるが，各々に大差はなく，韻律からの肯定／否定の判断をほぼ人間に感覚に等しい精度で実現できていることがわかる。

2.4 音声対話システム

ROBISUKE は，上のようなパラ言語の生成・理解機能を備えており，これらを使って対話を進めることができる [藤江, 2003]。現在の対話タスクは，先に述べた昼食相談であり，ROBISUKE は利用者からの相談を受けて適切なレストランを提案する (Figure 6)。

ROBISUKE は，まず，韻律情報から得られる認識結果

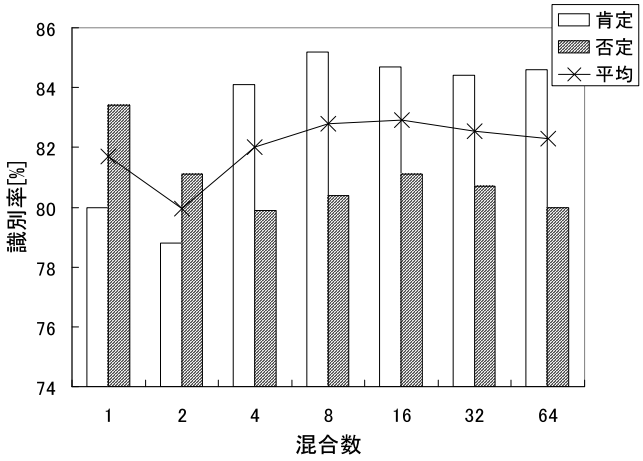


Figure 5: 韻律による肯定的／否定的発話の識別結果

Table 2: Cohen's κ の計算結果

	最小	最大	平均
M	**0.52**	**0.80**	**0.66**
A	0.52	0.85	0.72
B	0.52	0.79	0.62
C	0.65	0.80	0.72
D	0.65	0.85	0.76
E	0.61	0.79	0.72

と, 頭部ジェスチャの認識結果をもとに，発話者の状態を Table 3 のように推測する。例えば, 頭部ジェスチャが「うなずき」で韻律情報による態度が「肯定」の場合は「強い肯定」と解釈するが, 韻律情報による態度が「否定」の場合はユーザは判断に迷って考えていると解釈する．次に，

Table 3: 認識結果の統合
(* は対話中ほとんど現れない組合わせを表す)

		頭部ジェスチャ		
		うなずき	かしげ	首振り
韻律	肯定	強い肯定	弱い肯定	思案*
	否定	思案*	否定	強い否定

ここで推定された話者（相談者）の心的状態に応じて，次のような指針に従って行動する。

- 強い肯定：提案はユーザに受け入れられたと判断し，システムは自信を持って詳細な提案や情報を提供する．
- 弱い肯定：提案はユーザに大部分受け入れられたと判断し, システムは詳細な提案や情報を提供する．
- 思案：決定は見送られ, システムは「うん」や「何?」などを発話しながらうなずき, ユーザにもう一度応答を促す．
- 否定または強い否定：提案は却下されたと判断し, システムは代替案を提供する．

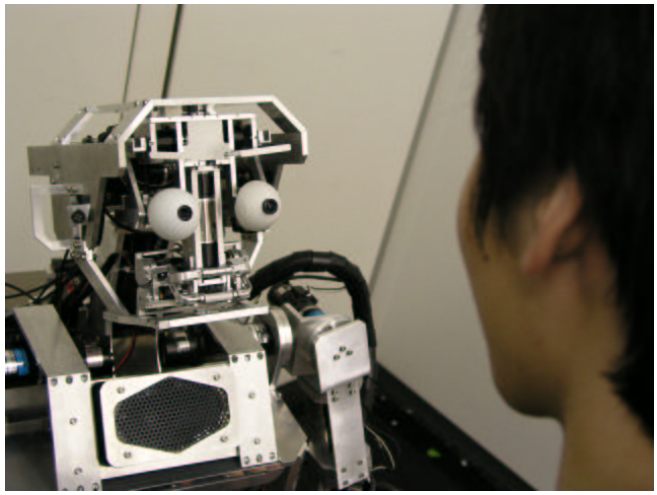


Figure 6: 対話する ROBISUKE

このように振る舞うことで，言語的な意味での明示的な応答がなくとも，ROBISUKE は適切に提案内容を決めることができ，円滑に対話を進めることができる。

対話例を，Figure 7 に示す．システムは, 提案に対してユーザの応答が否定的な場合は代案を, 肯定的な場合はより具体的な提案を行っている。

U: お昼ご飯なんだけど,
　どこかいいところ無いかな?
R: カレーなんてどう?
U: カレーかー **(強い否定)**
R: それじゃあ, ハンバーガーなんてどうかな
U: あーハンバーガーね **(強い肯定)**
R: ハンバーガーなら,
　近くにマクドナルドがあるよ

Figure 7: 対話例
U:はユーザの発話, R:は ROBISUKE の発話

3　対話内容の充実：新出表現に対する対応

一般に，許される表現が狭ければ，会話システムとしては魅力のないものとなる。そこで，利用とともに，表現に関する制約を除くための枠組みを検討している [細川, 2003]。

先の昼食相談のタスクで，システムが受け付けるべき最も重要な表現は，どのような店を相談者が望んでいるかを表す店の検索条件である。しかし，このような内容の発話は，かなり変化に富んでおり，それらを予め登録しておくことは不可能である。相談者は,「安い店が良い」「ボ

リューム満点の店が良い」と，明示的・直接的に希望を言うばかりではなく,「雨降っているね（歩きたくないから，近いところがいい）」「調子悪いんだよね（さっぱりしたものがいい／栄養をとりたい）」などと，状況を言うことで，間接的に条件を示すことも多い。このため，このような場面を想定して，新出の発案依頼表現への対応方法を検討している。

開発したシステムでは，予め，店の特徴を現す直接検索キーワードが設定されている。例えば,「近い」「安い」「(食べ物を出すまでの時間が) 早い」「ボリュームがある料理」「さっぱりとした料理」など，直感的に店を選ぶ条件として指定することが予想される単語である。システムのバックエンドのＤＢは，これらの条件で店を選択することができる。

また，システムは間接検索表現と直接検索キーワードの関係表を用意する。これは検索を要求する個々の表現に対し，各々のキーワードに対する関係の深さを数値で表したもので，例えば「体調が悪い」という間接検索表現には,「近い」「さっぱりとした料理」などの検索キーワードに高い値が与えられる。

ここでシステムが知識獲得として行うことは，新出の間接検索表現に対し，この関係表を作り以降これを保守することである。

システムは，直接検索キーワードで発案要求された場合はそのキーワードで，既知の間接検索表現で発案要求された場合には，その表現の関係表が与える検索キーワードで紹介すべき店を検索する。

システムは，未知の発案表現に出会うと,「それってどういうこと」と聞き返すことで，その表現と既知の情報とを結びつけるなんらかの手掛かりを要求する。例えば,「急いでいるんだ」という表現によって間接的に「近くで早く食事をだす店を提案すること」を要求する例の場合,「急いでいる」ということによって，何を要求したいのかの一端,(例えば,「近い」,「早い」のどちらか，あるいはそれらに結びつく既知の表現）を相談者から直接聞きだす。

次にシステムは，前段で得られたキーワードから，他の各キーワードとの関係を算出し，新出表現に対する直接検索キーワードとの関係表の初期値を作成する。この値は，既出の表現において，そのキーワードと他のキーワードとがどのような相関関係を持っていたかによって決める。この関係表は例えば次のようにして利用される。先の例で「近く」という語が「急いでいる」の表現に対し関連する検索キーワードとしてユーザから陽に与えられたとしよう。この場合，この語だけで検索しても提案すべき店を絞り込めないことが考えられる。このような場合,関連の深い他の検索キーワードがわかれば，それを付加的に用いて店を絞り込むことが可能である。例えば,「近い（時間をかけずに行って来られる）」と関連するものとして,「早い（時間をかけずに食べられる）」が得られたなら，これを追加のキーワードとして店の絞込みに役立てることが可能である。また，これらの関連検索キーワードは，最初の提案が拒否された場合の第 2 案を探す場合にも利用できる。

上で得た関連表の値は，陽に与えられる手掛かりから,要求表現と関連の深いものとして予想によって決めたものであって，実際に関連が深いかどうかは明らかでない。例えば,「体調が悪い」ということを理由に「近く」の店を要求した場合には，必ずしも「早い」ことは求められていない。このような場合，対話を通じてこうした差異を検知し，それを修正する方向に関係表の値を更新する。

以上のような機能を実装することで，ROBISUKE は,以下のような会話を実現できる。

```
U1: お昼ご飯なんだけど,
    どこかいいところ無いかな。
R : どんなところがいいですか。
U1: ちょっと体調が悪いんだよね。
R : それってどういうこと。
U1: さっぱりしたものが食べたいんだ。
R : それなら,「そばの実」はどう?
U1: いいね，そばの実。
R : それじゃきまり。いっていらっしゃい。
   ++++++++++
U2: お昼いいところ無いかな。
R : どんなところがいいですか。
U2: 体調が悪いんだ。
R : さっぱりしたものでよければ,
    「そばの実」はどう?
```

Figure 8: 表現の獲得を含む対話例

4 まとめ

会話ロボット ROBISUKE を紹介した。

ROBISUKE は，対話相手が明示的に意思表示をしない場合でも，発話の韻律パターンや頭部ジェスチャから，対話相手の心情を推測し，この情報を用いて効率的に対話を進めることができる。また，初めて聞く表現でも，それが意味するところを対話相手から聞きだすことができ,聞き出したあとは，この表現に対し適切に振舞うことができる。これらの機能によって，リズムある対話が実現され，従来の対話ロボットとは全く印象の異なる対話が実現することができた。

本稿では，言語コミュニケーションにおけるパラ言語の扱いに焦点を当てたが，感性的なコミュニケーションを指向して，より細やかな表情を表出できるロボットの開発

も進められている [Kismet, 2000][Miwa, 2002]。次の段階では，このような技術との融合が望まれる。

参考文献

[橋本, 1997a] 橋本周司 他，“ヒューマノイドー人間型高度情報処理ロボット,” 情報処理，Vol.38，No.11，pp.956-969, 1997.

[Imai, 2001] M.Imai, T.Ono, H.Ishiguro, “Physical Relation and Expression: Joint Attention for Human-Robot Interaction,” RO-MAN2001, pp.512-517, 2001.

[岩沢, 2002] 岩沢 透, 大中 慎一, 藤田 善弘, “状況検知を利用したロボット用音声認識インタフェースの一手法とその評価,” 第 16 回人工知能学会 AI チャレンジ研究会,Nov. 2002.

[Nisimura, 2002] R. Nisimura, T.Uchida, A.Lee, H.Saruwatari, K.Shikano, Y.Matsumoto, “ASKA: Receptionist Robot with Speech Dialogue System, ” Proceedings of IEEE/RSJ IROS2002, pp.1314–1317, Sep. 2002.

[Nakadai, 2003] K. Nakadai, H.G.Okuno, H.Kitano, “Robot Recognizes Three Simultaneous Speech By Active Audition,” Proc. IEEE-RAS ICRA-2003), May 2003.

[Fujisawa, 1974] H. Fujisawa, K. Shirai, “An Algorithm for Spoken Sentence Recognition and Its Application to the Speech Input-Output System,” IEEE Trans. On Systems, Man and Cybernetics, 1, SMC-4, 5, 1974.

[白井, 1985] 白井克彦，他，“ロボットとの柔軟な対話を目的とした音声入出力システムー WABOT-2 における会話系,” 日本ロボット学会誌，Vo.3，No.4，pp.104-113, 1985.

[Hashimoto, 1997b] S. Hashimoto, et al., “Humanoid Robot — Development of an Information Assistant Robot Hadaly—,” 6th IEEE International Workshop on Robot and Communication, 1997.

[Hashimoto, 2002] S.Hashimoto, et al., “Humanoid Robots in Waseda University — Hadaly2 and WABIAN —, Autonomous Robots,” Vol.12，No.1，pp.25-38, Jan. 2002.

[松坂, 2001] 松坂要佐, 東條剛史, 小林哲則, “グループ会話に参与する対話ロボット,” の構築,” 電子情報通信学会論文誌, Vol.J84-D-II, No.6, pp.898–908, 2001.

[Tojo, 2000] T.Tojo, Y.Matsusaka, T.Ishii, T.Kobayashi, “A conversational robot utilizing facial and body expressions,” in Proceedings of 2000 IEEE SMC2000, vol. 2, pp. 858-863, 2000.

[Lieske, 1997] C.Lieske, J.Bos, M.Emele, B.Gambäck, CJ Rupp, “Giving prosody a meaning,” in Proceedings of ISCA EUROSPEECH’97, vol. 3, pp. 1431–1434, 1997.

[Kawato, 2000] S.Kawato and J.Ohya, “Real-time detection of nodding and head-shaking by directly detecting and tracking the ‘between-eyes’,” in Proceedings of Fourth IEEE international conference on automatic face and gesture recognition, pp. 40–45, 2000.

[Kapoor, 2002] Ashish Kapoor and Rosalind W. Picard, “A real-time head nod and shake detector,” Tech. Rep. 544, MIT Media Laboratory Affective Computing Group, 2002.

[Kobayashi, 1997] Hiroshi Kobayashi and Fumio Hara, “Facial interaction between animated 3d face robot and human beings,” in Proceedings of 1997 IEEE SMC97, vol. 4, pp. 3732–3737, 1997.

[Cohen, 1960] J. Cohen, “A coefficient of agreement for nominal scales,” Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960.

[江尻, 2003] 江尻康, 中島慶, 藤江真也, 小林哲則, “動作中の対話ロボットにおける頭部ジェスチャ認識,” 電子情報通信学会, PRMU, Nov. 2003.

[藤江, 2003] 藤江真也, 江尻康, 菊池英明, 小林哲則, “パラ言語の理解能力を有する対話ロボット,” 情報処理学会研究技術報告, SLP-48, pp.13- 20, Oct., 2003.

[細川, 2003] 細川健一郎, 藤江真也, 小林哲則, “検索・提案型対話システムのためのユーザとのインタラクションによる適応的意図理解” 人工知能学会, SLUD, Nov. 2003.

[Kismet, 2000] C.Breazeal, “Sociable Machines: Expressive Social Exchange Between Humans and Robots,” Sc.D. dissertation, Dept.EECS, MIT, 2000.

[Miwa, 2002] H.Miwa, T.Okuchi, H.Takanobu, A.Takanishi, “Development of a New Human-like Head Robot WE-4,” IROS2002, Vol. , pp.2443-2448, 2002.

[小林, 2003] 小林哲則, “ 会話ロボットの実現に向けて,” 電子情報通信学会, ヒューマンコミュニケーション基礎研究会, April 2003.

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-0318-2 (11/13)

発達ロボティクスからみたロボット聴覚研究

Auditory Capacity for Epigenetic Robotics

小嶋 秀樹 ・ 矢野 博之
Hideki Kozima ・ Hiroyuki Yano
通信総合研究所*
Communications Research Laboratory
{xkozima, yano}@crl.go.jp

Abstract

Human cognitive capabilities, including the one for audition, develop over time from one's birth or conception. From this epigenetic stance, we are building robots, *Infanoid*s, that develop their cognitive capabilities through the physical and social interaction with the environment. This paper describes what auditory capacity we need in epigenetic robotics and what outcomes we expect from this approach, especially by investigating innately perceivable value of sound, acquired meaning of sound, and their roles in the emergence of inter-personal communication.

1 はじめに

ロボット聴覚の研究が大きな発展をみせている．音源定位・話者分離・聴覚による情景理解などが実環境・実時間で可能になりつつある．また，これら要素技術によって，いままで理想化された音環境あるいはテキストベースで研究されてきた自然言語処理技術（音韻分析から語用論的処理まで）が実世界に根を下ろそうとしている．

このような流れのなかで,「聴覚」はロボットあるいは人間にどのような情報をもたらすのかを改めて議論してみたい．とくにコミュニケーション能力の発達という視点から，人間が発する音（声に限らず，手や道具をつかって発する音も含める）がどのように知覚され，その経験がどのように認知発達（とくにコミュニケーション発達）につながっていくのかを考察したい．

まず，つぎの第２節では，この議論の背景となる発達ロボティクス（Epigenetic Robotics）について，著者らが開発しているロボット Infanoid・Keepon を例として解説する．つづく第３節では，生まれたばかりの乳児にとって，音（とくに人間の発する音）はどのような価値をもつものとして知覚されるのか，また経験をとおしてその音はどのような意味をもつようになっていくのかを考える．第４節では，このような音の価値や意味を他者と共有することで，コミュニケーションがいかに発現するのかを考え，これら議論をふまえて，ロボット聴覚研究に新たな方向性を示唆したい．

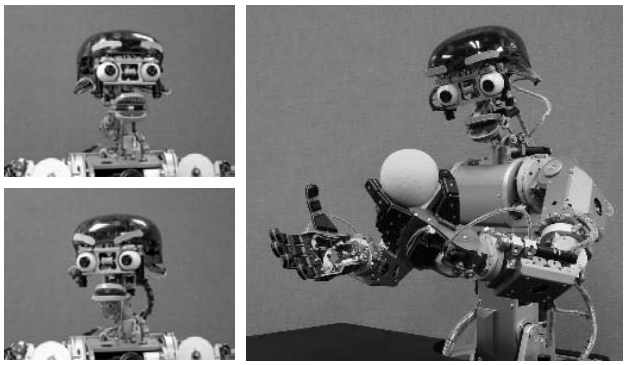


Figure 1: *Infanoid* and some facial expressions

2 発達ロボティクス (Epigenetic Robotics)[1]

著者らは，子どもと養育者のあいだのコミュニケーションの発達を手がかりとして，それと同じようにロボットを発達させることをめざした「Infanoid プロジェクト」を進めている．ロボット上にコミュニケーション能力の発達を再現することと，そのロボットを使って子どもの発達を観察すること——これら相補的なアプローチを行き来することで，コミュニケーション能力のなりたちを解き明かし，人間とロボットのあいだのコミュニケーション，そして共生の可能性を探索していく．

* 通信総合研究所 けいはんな情報通信融合研究センター 社会的インタラクショングループ（〒 619-0289 京都府相楽郡精華町光台 3-5）

[1] 発達ロボティクスに関する国際会議のひとつに Epigenetic Robotics (http://www.epigenetic-robotics.org) があげられる．

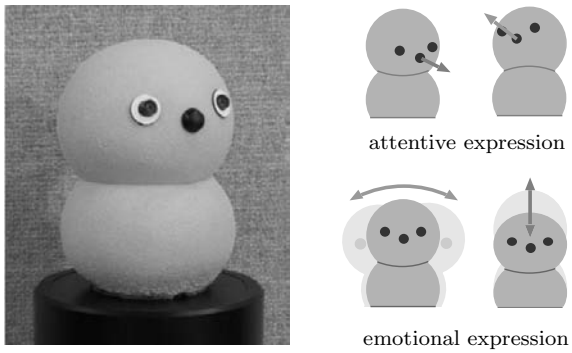


Figure 2: *Keepon* and its expressive functions

2.1 研究プラットフォーム：Infanoid と Keepon

発達ロボティクスの研究プラットフォームとして開発しているのが「子ども型ロボット Infanoid」[Kozima, 2002] である（Figure 1，右）．自由度 29 の上半身ヒューマノイドであり，3〜4 歳児とほぼ同じ大きさ（高さ 48cm）をもつ．手には 5 本の指があり，指さしをしたりオモチャなどをつかむことができる．頭部には左右の眼球があり，上下左右にすばやい視線移動（サッカード）とスムーズな対象追跡が可能である．それぞれの眼球には周辺視と中心視のための２つのビデオカメラ——左右で合計４つ——が装着され，クラスタ PC での画像処理によって人間の顔やオモチャなどの検出・測距・追跡が可能となっている．また，眉毛や上下の唇を動かすことで，さまざまな表情をつくりだす（Figure 1，左）．耳にあたる左右のマイクロフォンから人間の声を聞きとり，その韻律情報や音韻情報を抽出すること，また，それらを音声合成装置に入力することで，いわゆるオウム返しができるようになっている．

もうひとつの研究プラットフォーム「Keepon」[2] は，乳児との身体的コミュニケーションを指向した「ぬいぐるみロボット」である（Figure 2，左）．黄色いダンゴ型（高さ 12cm・直径 8cm）の身体に，左右の眼（ビデオカメラ）と鼻（マイクロフォン）をもち，Infanoid とほぼ同じ視聴覚機能を発揮する．このシンプルな身体にできる動作は，① 注意の表出——顔の方向を上下左右に動かし，視線をある対象に定位させることと，② 情動の表出——身体を左右に揺らしたり上下に伸縮させることで，楽しさや興奮などを表現することの２つに絞り込まれている（Figure 2，右）．頭や腹はシリコンゴムで一体成形されているため，これらの動作は全身をやわらかく変形させることで行なわれる．

2.2 注意と感情のやりとり

Infanoid と Keepon は，視線や身ぶりをつかった注意のやりとりをとおして，人間とのあいだに身体動作の時間的同期・空間的定位（焦点化）をつくりだし，互いの身体的な経験（とくにその情動的な側面）を共有していくことで，共感的なコミュニケーション関係をつくりあげていくこと [小嶋, 2001] をねらってデザインされている．

Figure 3: Eye-contact and joint attention

このような関係構築のベースとなる注意と感情のつながりには，アイコンタクトと共同注意が大きな役割を果たしている．そこで Infanoid プロジェクトでは，まずアイコンタクトと共同注意の能力をロボット上に実現し，それを出発点として注意と活動のつながりを発現させることを試みる．アイコンタクトとは，互いの眼あるいは顔を同時に見ることであり，インタラクションに時間的な同期を与える．Infanoid や Keepon では，人間の正面顔を検出し，その顔に視線を（あるいは頭や手も）向けることで実現されている（Figure 3，左）．共同注意とは，互いに同じ対象を同時に見ることであり，インタラクションに空間的な焦点化を与える．Infanoid や Keepon では，人間の顔の位置と向きを捉え，その顔の向きに沿って対象（オモチャなど）を探しだし，そこに視線などを向けることで実現されている（Figure 3，右）．

2.3 子どもとのインタラクション

アイコンタクトや共同注意の能力をもったロボットに，子どもたちがどのように関わろうとするのかを観察し，彼らがこれらロボットをどのような存在と捉え，どのようにコミュニケーション関係をつくりあげていくのかを調べている[3] [小嶋, 2003]．Infanoid にはおもに幼児期の，Keepon にはおもに乳児期の子どもたちが，何の予備知識も課題設定もない状況で対面する（Figure 4）．このインタラクションの観察から，子どもからロボットへの関わりかたにつぎのような変化傾向がみられた．

① 動く「モノ」として観察する．Infanoid には，関わり方を決めかねて困惑（neophobia）を見せた．

② 応答する「システム」として探索する．オモチャを見せる・ロボットに触れるなどして，ロボットの応答パターンを引き出す．

③ 心をもった「他者」として社会的なやりとり（物を見

[2] Keepon のハードウェアと基本ソフトウェアは小嶋が開発し，行動システムを仲川こころ（CRL）との共同で開発している．

[3] Infanoid を使ったインタラクション観察は，川合伸幸（名大）・矢野喜夫（京教大）・小杉大輔（京大）・仲川こころ（CRL）・小嶋の共同で実施し，Keepon を使ったものは，小杉大輔・村井千寿子（京大）・仲川こころ・小嶋の共同で実施した．（敬称略）

Figure 4: How children interact with the robots

せる・物を手渡す・ことばで質問するなど）や，向社会的な関わり（誉める・物の名前や扱いかたを教えるなど）をみせる．

0 歳児は ① のみ，1 歳児は ② まで，2 歳以上の子どもは ③ まで，時間経過とともに関わり方を深めていった．

この観察から，子どもからみたロボットという存在がつぎのように変化していることが示唆される．①「モノ」から「システム」への変化は，姿勢や表情の変化，発声などから，ロボットが注意や感情をもった自律的な〈主体〉であることに気づくことによって達成される．②「システム」から「他者」への変化は，ロボットの注意や感情が子ども自身の行為に随伴していることへの気づきがトリガとなり，主体的なロボットとの関係をつくりあげることによって実現する．ロボットの主体性と関係性を捉えることで，子どもたちは心の帰属を受けとめるられる「他者」としてロボットを捉えるようになる．

このようなインタラクション観察から，その身体的・音声的なやりとりを支え，コミュニケーション関係を構築していくための土台となる認知機能が明らかになりつつある．次節からは，ロボットおよびヒト乳児におけるコミュニケーションの成り立ちという視点から，音がどのように知覚されるべきかを考察していく．

3　音の価値・音の意味

ロボット聴覚研究の多くは，音を客観的なデータ——数値データ（１次元から数次元）の時系列——として扱うことから出発する．しかし人間の聴覚は，たとえ新生児のような聴覚経験に乏しい個体であっても，何らかの主観的な価値——快/不快といった情動あるいは接近/回避といった反応とのむすびつき——をもつものとして音を知覚し，それを認識のベースとしている．ここでは「共感覚」という現象を手がかりに，音のもつ主観的な価値とは何か，それは聴覚発達をどのように方向づけるのかを考える．

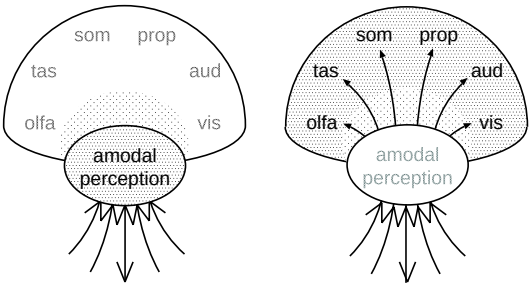
3.1　共感覚と amodal な知覚

共感覚（synesthesia）とは，ある感覚モダリティでの刺激が別の感覚モダリティでの知覚を不随意的に引き起こす現象である [Cytowic, 1995]．たとえば，ある楽器の音を聞くことで，（実際には刺激として存在しない）何らかの色彩を知覚することは，共感覚の代表例といえる．成人ではごく一部（おそらく数千人にひとり）の人が共感覚をもつ．女性と左利きに多い．近親者での共起頻度が高いことから，遺伝的な要因が関与していると考えられる．先にあげた「音から色へ」だけでなく，嗅覚・味覚・触覚を含めたいずれの感覚モダリティの間にも共感覚は生じうるが，一般に共感覚は一方向的（たとえば聴覚→視覚）となる．共感覚は随意的に促進あるいは抑制できない現象である．成人の共感覚は一般に時間経過とともに変化することはない．

共感覚はごく一部の人にみられる非典型的（atypical）な知覚スタイルであるが，共感覚ほど強力ではない感覚モダリティ間の干渉はふつうの人にも起りうる．たとえば，照明の色によって聴覚の鋭敏さが変化すること，おなじ大きさと質量の黒箱と白箱では黒箱を重く感じることなどが知られている [山田, 2000]．また，共感覚メタファ [楠見, 1995]（「黄色い声」「うるさい模様」といった表現）は，このような普遍的な干渉現象が言語文化に沈殿したものとも考えられる．

従来は，脳の機能局在の考えから，皮質領野間のクロストークとして共感覚が説明されてきたが，最近の赤ちゃん研究からの示唆をうけ，共感覚への新しい見方が出てきた．それは，ふつうの乳児でも生後 4ヵ月頃までのあいだ共感覚的な知覚スタイルをもつという仮説 [Baron-Cohen 1996] である．この仮説によって，乳児のもつ生得的コンピテンス，たとえば新生児模倣 [Meltzoff, 1977]（他者の口開けや舌出しといった顔動作を模倣できること）や，形の異なるオシャブリをしゃぶっただけで視覚的にそれらを区別できること [Meltzoff, 1979]，刺激のリズムあるいは強さを聴覚と視覚のあいだでマッチングできること [Spelke, 1987] などが説明できる．また，この仮説では，生後 4ヵ月頃までの乳児は amodal（未分化）な知覚スタイルをもち，これが視覚・聴覚・触覚などのモダリティに分化していくと考える．この「未分化から分化へ」という考えは，従来の Piaget 流の考え——経験をとおした感覚統合による認識——とは方向性がまったく異なる [Gibson, 2000]．そして，ごく一部の人にみられる共感覚は，この分化プロセスに何らかのバイアスが加わった結果，amodal な知覚の一部分がそのまま残されたものだと考えられる．

乳児の amodal な知覚は，ランダムに近い結合をもつ皮質上で生じているという考えもあるが，著者らは，リズム・テクスチャ・強さ・方向といった次元によって皮質下で統一的に知覚されていると考えている．「リズム」とは



(a) 0 to 4 months (b) 5 months and on

Figure 5: Amodal vs conventional perception

Figure 6: Ramachandran's figures

刺激の強弱パターンであり,「テクスチャ」とはより細かなパターン（音色・肌理など）である.「強さ」は刺激の大きさや移動速度であり,「方向」とは自分からみた刺激の位置や移動方向である．このような次元によって対象を捉え，ダイレクトに情動（快/不快など）や身体動作（定位・リーチングあるいは回避動作など）に結びつけることで[4]，乳児のもつ生得的コンピテンスを説明できると考えている．

3.2 音の価値

前項で説明した amodal な知覚が，ふつうの成人でも，通常の知覚プロセス（感覚モダリティごとに分化した情報処理）と並行して，意識下のうちに作動していると考えてみよう．つまり，乳児が発達するにつれて，モダリティごとの情報処理システム（上位）が皮質上に構築されるが，それまで皮質下で作動していた amodal な情報処理システム（下位）も作動しつづける[5] と考える．すべての感覚入力情報は，下位システムで共感覚的な「拡張」を受けてから上位システムに送られる（Figure 5）．このモデルによって，先にあげた感覚モダリティ間の干渉や共感覚メタファをうまく説明できるだけでなく，単一モダリティでは情報の一部が欠けていたり曖昧であったりする場合に，意識下で「穴埋め（filling-in）」されることも説明できるだろう．

音の知覚についても，このモデルを想定することで，いろいろな現象を説明できる．Ramachandran [2003] が報告した現象から見てみよう．被験者に２つのよく似た閉領域図形（Figure 6）を見せる．一方は角が丸い曲線で描かれ，もう一方は直線状のとがりをもっている.「“bouba” はどちらで “kiki” はどちらか」と質問すると 98%の被験者が「丸い方が “bouba” で，とがった方が “kiki”」と答える．どちらも無意味単語であるが，被験者たちはその音（聴覚刺激）と図形（視覚刺激）との感覚モダリティを超えたマッチングを行なっている．このような「類像性（iconicity）」は有意味単語にも潜んでいる．とくに通時的な変化がすくない基本語彙（たとえば「甘い」「雷」「ぶつかる」「笑う」）には，音と意味の類像性が感じられるものが多い[6] また，単語の脚韻（語尾の音韻パターン）からその意味の「イメージ」を予測できることが，英語とタイ語について報告されている [Ross, 2001]．たとえば英語の場合，“-ump” は半円あるいは半球状（*bump*, *hump*, *lump*, *mump(s)* など）を，“-og” は湿った状態（*bog*, *fog*, *frog*, *sog(gy)* など）をイメージさせるという．

3.3 音の意味

リズム・テクスチャ・強さ・方向といった次元によって，聞きとった音の amodal な価値を経験することで，どのような効果が得られるのだろうか．まず，その音のもとになっている対象・事象をひとつの「まとまり」として知覚することができる．つまり，いま聞いている音が，いま見ている光景とどのように関係しているのか，どの音がどの光と関連しているのかを，ある特定のリズム・テクスチャ・強さ・方向をもった amodal な知覚単位として捉えられるというものである．また皮質上にモダリティ分化した情報処理システムが構築されてからは，この amodal な知覚単位から各感覚モダリティの注意を制御し，ひとつの「まとまり」を意識化することが可能となっていく．

もうひとつの効果は，知覚した対象・事象と自分の身体動作（厳密には内部受容感覚）とをつなげられることである．たとえば，何かが自分に向かって飛んできたとき，無意識のうちに定位あるいは回避姿勢をとることはそのよい例であろう．また，他者の身体動作を，リズム・テクスチャ・強さ・方向として捉え，それを自分の身体動作として「再現」することも考えられる．これは完全な模倣ではないが，身体間のマッピング——視覚イメージとして捉えた他者の身体動作を自分の運動イメージに変換す

[4] 身体動作に付随する内部受容感覚（proprioception）も amodal な知覚の対象となりうる．自分の身体動作のリズム・強さ・方向などを，外部からの刺激（たとえば視覚的に捉えた他者の身体動作）とマッチングさせることができる．

[5] これは subsumption アーキテクチャ[Brooks, 1991] の一種と考えられる．amodal な情報処理システムの出現は，個体発生的に先行するだけでなく，おそらく系統発生的にも先行するだろう．

[6] シェークスピアはサクソン語に由来する英語基本語彙を多用し，心を揺さぶる文体をもつと言われる．ノルマンコンクエスト以降，フランス語から抽象的な語彙（-tion で終わる単語など）が導入されたが，これらはあまり類像的でないという．

ること――に頼らなくても，ある程度類似した動作を再現できる．環境からの制約（アフォーダンスの配置など）を共有し，試行錯誤を重ねることで，他者の身体動作の結果を再現すること（emulation）[Tomasello, 1999] にもつながっていくだろう．このような身体動作のやりとりは，乳児のコミュニケーション発達に欠かせない．

このように音の「意味」とは，音のもと（多くの場合は人間の活動）から「まとまり」を捉えること，そして「まとまり」への自分の応答（身体の定位など）を準備することといえる．次節では，このような音知覚から対人コミュニケーションがどのように発現するのか，その道すじを提示してみたい．

4 結論――コミュニケーションへ

コミュニケーションの本質は，相手の心の状態にアクセスすることである．とはいっても相手の心は目に見えない．見えるのは相手の身体とその動き[7] だけである．相手の身体動作から心の状態を読みとるには，相手がどのように環境を知覚しているのか，そして切りだした環境にどのように働きかけようとしているのかを捉え，それをいわば疑似体験することが必要となる．

乳児と養育者のコミュニケーションがいかに始まる（回りだす）のかは興味深い問題である．互いの身体やその動きを amodal な「まとまり」として知覚し，互いに同調しあうことで，ダイナミックなシステムが形成される．このシステムは，あくまで amodal な「まとまり」を媒介としているため，さまざま身体動作に対応でき，またさまざまな人工物（あるいは外乱）の導入にも対応できる．このようなシステムが作動することで，乳児と養育者は知覚・注意・感情をすりあわせながら，さまざまな事象を経験し共有していく [Trevarthen, 2001]．最初の段階では，養育者が乳児のもつ amodal な知覚-行為パターンを読みとり，利用することで，このシステムの作動が維持されるだろう．やがて，乳児のほうも，養育者の応答パターンをゆっくりと学習していく．乳児の感覚モダリティが分化してゆくにつれて，養育者からの働きかけもレパートリが広がってゆく．養育者は乳児の欲求や情動を積極的に読みとり，それに応じてやることで，非対称ながらも外見的には意図や感情をやりとりする社会的なインタラクションが発現する．やがて乳児も養育者のこのような応答を予測できるようになり，その予測を反映した行動ができるようになっていく．こうして，意図や感情のやりとりは双方向のものに発展し，真の意味で社会的なインタラクションへと入っていく．

このようなコミュニケーション発達をロボット上に実現するためには，いままでとは質的に異なる聴覚システムが必要となるだろう．具体的にいえば，リズム・テクスチャ・強さ・方向といった次元によって音を知覚するシステム，そしてこの amodal な聴覚情報を，視覚・触覚・内部受器感覚などからの amodal な知覚情報と関連づけ，情報の選別や拡張を行なうシステムなどが必要となる．このような研究を動機づけるのは，役に立つものをつくる工学的なスタンスだけでなく，人間の認知発達を解き明かし，より深く人間を理解することへの魅力かもしれない．これからも工学と人間理解のバランスをとりつつ研究を進めていきたい．

[7] 「身体の動き」とは，筋肉運動（発語を含む）と一部の分泌系活動（顔色の変化や発汗など）である．

参考文献

[Baron-Cohen, 1996]] Baron-Cohen, S.: Is there a normal phase of synaesthesia in development?, *Psyche*, Vol.2, No.27, 1996.

[Brooks, 1991]] Brooks, R. A.: Intelligence without representation, *Artificial Intelligence Journal*, Vol.47, pp.139-159, 1991.

[Cytowic, 1995] Cytowic, R. E.: Synesthesia: Phenomenology and neuropsychology, *Psyche*, Vol.2, No.10, 1995.

[Gibson, 2000] Gibson, E. J. and Pick, A. D.: *An Ecological Approah to Perceptual Learning and Development*, Oxford Univ. Press, 2000.

[Kozima, 2002] Kozima, H.: Infanoid: A babybot that explores the social environment. Dautenhahn, K. *et al.* (eds), *Socially intelligent agent*, Kluwer Academic Publishers, pp.157–164, 2002.

[小嶋, 2001] 小嶋 秀樹・高田 明：社会的相互行為への発達的アプローチ：社会のなかで発達するロボットの可能性，人工知能学会誌，Vol.16, pp.812–818, 2001.

[小嶋, 2003] 小嶋 秀樹：赤ちゃんロボットからみたコミュニケーションのなりたち，発達，Vol.24, No.95, pp.52-60, 2003.

[楠見, 1995] 楠見 孝：比喩の処理過程と意味構造，風間書房，1995.

[Meltzoff, 1977] Meltzoff, A. N. and Moore, M. K.: Imitation of facial and manual gestures by human neonates, *Science*, Vol. 198, pp. 75–78, 1977.

[Meltzoff, 1979] Meltzoff, A. N. and Borton, R. W.: Intermodal matching by human neonates, *Nature*, Vol.282, pp.403–404, 1979.

[Ramachandran, 2003] Ramachandran, V. S. and Hubbard, E. M.: Hearing colors, tasting shapes, *Scientific American*, Vol.288, No.5, pp.52-59, 2003.

[Ross, 2001] Ross, P.: Image schematic rhyme in Thai: Perspectives from first language acquisition, *Proceedings of the International Workshop on the Relation between Cognitive and Linguistic Development (Bangkok, Thailand)*, 2001.

[Spelke, 1987] Spelke, E. S.: The development of intermodal perception, In Cohen, L. B. and Salapatek, P. (eds), *Handbook of Infant Perception*, Academic Press, 1987.

[Tomasello, 1999] Tomasello, M.: *The Cultural Origins of Human Cognition*, Harvard Univ. Press, 1999.

[Trevarthen, 2001] Trevarthen, C.: Intrinsic motives for companionship in understanding: Their origin, development, and significance for infant mental health. *Infant Mental Health Journal*, 22, 95–131, 2001.

[山田, 2000] 山田 尚勇：日本語をどう書くか— 入力法および表記法のヒューマン・インタフェース学入門：I. ヒューマン・インタフェースと脳の科学，中京大学情報科学部テクニカルレポート 1999-2-02，2000.

QRIO SDR-4XII の音声インタラクション
Verbal Interaction Implemented in QRIO SDR-4XII

下村 秀樹
Hideki Shimomura

ソニー(株) ライフダイナミクス研究所準備室
Life Dynamics Laboratory Preparatory Office, Sony Corporation
simomura@pdp.crl.sony.co.jp

Abstract

This paper describes verbal interaction implemented in QRIO SDR-4XII, which is an entertainment humanoid robot. The interaction is designed based on following four policies: 1)utilizing user identification, 2)natural interaction sequence in real environment, 3)effective combination of current dialogue technologies, 4)development of entertainment verbal interaction functions using LVCSR. In this paper, the implemented verbal interaction and its design policies are described in detail. The basic hardware/software architecture of QRIO, on which the interaction is implemented, is also introduced.

1. はじめに

最近，人間とのインタラクションを指向したロボットが次々と開発され，その上でさまざまな研究が行われている．概観すると，インタラクションの基本原理・機構の解明を狙った基礎研究から，ロボットの持つエンタテイメント性を重視し，家庭での使用をかなり強く意識した実用的検討まで幅広い試みが見られる［Ohnaka01, Kanda02］．その中でも特にヒューマノイドロボットにおいては，人間に似たその形状から言語による音声インタラクションが自然と期待され，その機能が検討されている．

音声でのインタラクションには，音声認識，音声合成，言語処理，対話技術（非言語インタラクションを含む）が重要かつ必須であるのは言うまでもない．さらに家庭など実環境での使用を考えるならば，環境認識，音の方向や種類の認識，その情報に基づく移動なども重要である．例えば，後ろを向いているロボットに呼びかけ，近くに移動させて，会話を行うところまでインタラクションの解釈を広げれば，これらの必然性は明らかであろう．

しかし客観的に現状の技術を見るとき，それらを統合して全体として極めて自然な音声インタラクションを実現するのは困難である．例えば，実世界での認識・意味理解・音声対話，そして家庭での長期間使用に耐え得るインタラクションの設計と実装など多様な課題が存在する．この段階では，その要素技術への深い取組みの必要性は当然であるが，ロボットでの音声インタラクションの全体的な評価を行ったり方針を定めたりするために，技術を統合・実装して検証を繰り返すことも重要である．特に，長期間使用するためのインタラクションの課題，様々な実環境でどう使えるかなどの点は，要素技術の評価だけでは説得力がない．

我々はエンタテイメントロボット QRIO SDR-4XII（以下 QRIO）を開発し，その上に前記のような意図も込めて音声インタラクション機能を実装した．以下本稿では，QRIO に実装した音声インタラクションを紹介する．まず 2 節で，QRIO での音声インタラクションに関する我々の設計方針を述べる．次に，3 節でハードウェア，ソフトウェアのアーキテクチャを説明し，続く 4 節で QRIO での音声インタラクションの実装，会話の具体例を報告する．

2. QRIO での音声インタラクションの設計方針

我々は，QRIO の音声インタクラクションを実現するにあたり，家庭での使用を前提として次の点を重視した．

- 個人識別機能の有効利用
- 実環境での自然なインタラクションシーケンス
- 既存の対話技術の有効活用と連携

・大語彙連続音声認識（LVCSR）を利用したエンタテイメント対話の開発

次に，それぞれを説明する．

2.1 個人識別機能

ロボットのインタラクションは，その相手が誰であるかによって変化すべきである．例えば，その人物と以前に何を話したのかを記憶しておき，その記憶に基づいて適切な話題を選べるとしよう．この能力は，ユーザにある種の知性を感じさせるだけでなく,「自分がロボットに知られている」という喜びを与えることができる．そのためには，まずロボットが人物を学習し，必要に応じて同定して，インタラクションを実施する機構を設ける必要がある．また，学習はオンラインで（ロボットの通常動作中に),音声インタラクションの一部として実行されることが望ましいだろう．

個人を識別する情報としては,画像処理による顔認識，音声による話者認識，その人物の名前（音声認識），その他特殊なセンサを利用した指紋認識などが考えられる．どのような認識器を使うとしても，オンライン学習機能を持つ認識器と，その認識結果を相互に関連付ける連想記憶機構を用いることで，個人の記憶・識別が実現できる．

2.2 実環境でのインタラクションシーケンス

ロボットを実環境，例えば家庭環境に置いたとき，そもそもロボットと会話できる状況を作り出すことが難しい課題となる．例えば，ロボットが人間を見ていないときに，呼びかけ，人間の存在を認識してもらい，さらに会話できる距離まで移動してきてもらうことを考えよう．そのためには，音源方向の検出，ある程度の距離での顔の検出が必要である．また，家庭内の環境（障害物の存在等）を想定した目的地への歩行も必須の技術となる．

他の例として，ある人と会話をしているときに別の人から割り込まれた後にも元の会話に戻れる，といったシーケンスを想定しよう．そのためには，音の方向検出の他，誰と会話していたかの記憶（見えなくなった場所に存在した人や物の記憶)，前に行った会話状況の記憶，行動への復帰処理などが求められる．このためには，それらを可能にする行動制御アーキテクチャが必要である．

2.3 既存の対話技術の活用と連携

ヒューマノイド型である QRIO は，人間の自然言語を用いたインタラクションが自然に期待される．しかし，すべての場面で完全な意味理解と適切な応答を行うことは不可能ある．その前提に立ち，何が人間をロボットとのインタラクションに引き付けるであろうかということを，既存の技術の上で検討すること重視した．具体的には，1)長期間使用できる対話の枠組みの検討，2)大量のコンテンツによるエンタテイメント性の強化，3)対話機能間の独立な実装と緩やかな連携，を意識して検討を進めた．

1)に対しては，ネットワークを通して新たな情報を供給し続ける枠組みがまず考えられる[1]．それ以外には，例えば会話してユーザから得た情報を記憶し，それを後の対話の中に織り交ぜて，会話内容の連結を計ることも有効であろう．2)については，ロボットの会話内容をデザインしたい人がそのイメージをできるだけ簡易に実現できる枠組み（編集環境）を用意し，大量の会話コンテンツを実装できることが重要と考えた．さらに，対話全体を統一するような一般的枠組みを既定することは難しいと考え，様々な方式での対話をあまり相互関係を考慮せずに実装する方針を持った．これが 3)に該当するが，一方で，ある程度の文脈に沿った会話が成立しやすなるよう，複数の対話の話題を関連付けて遷移する仕組みを考慮すべきであろう．

2.4 LVCSR を使った音声エンタテイメント機能

一般の対話システムでは，LVCSR を有効に活用する方法が提案されているとはいえない．それは，大語彙を認識できたとしても，認識間違いを含む大語彙の認識結果をうまく扱う言語処理・意味処理技術にまだ課題が残っているからである．しかし，エンタテイメントとして視点を変えたときには,適用可能な領域があると考える．

例えば，ユーザが言ったことを LVCSR で認識し，そのまま繰り返す，あるいは少し変形して言い返すだけでも，場面によっては会話をスムーズに進行させる効果を持つことがある．また LVCSR を使った伝言機能も，録音音声ではなく認識結果をロボットの声で別の人間に伝えるようにすれば，ロボットが家庭内に存在する面白さが現れてくる可能性もある．

これらの検討は研究というよりも工夫という範疇になるかもしれないが，エンタテイメントの機能として検討する価値があると考えた．

3. QRIO のソフトウェア/ハードウェアアーキテクチャ

本節では，2 節で述べた音声インタラクションを実装するプラットフォームである QRIO について説明する．具体的には，ハードウェアと，インタラクションを含む行動全般を支える行動制御アーキテクチャを述べる．

---

[1] 今回はネットワークの機能は実装しなかった．

図１　QRIO SDR-4XⅡ

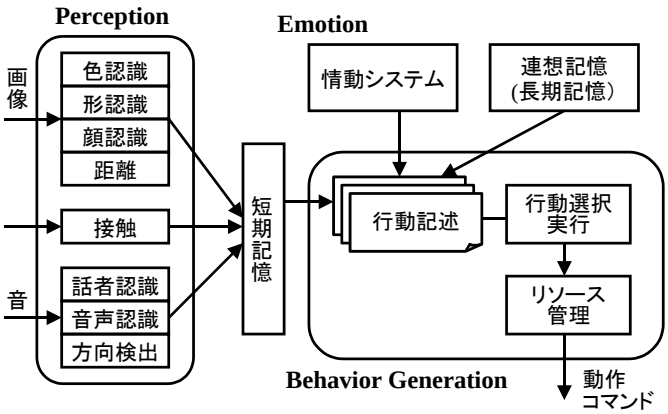


図２　QRIO の行動制御アーキテクチャ(EGO Architecture)

3.1　QRIO の概要

QRIO は，身長約 60cm，体重約 6Kg，全身で 38 の自由度（うち両手の指 10 自由度）を持つ 2 足歩行ロボットである．外観を図 1 に示す．主なセンシング機器として，ステレオカメラ，マルチマイクロフォン(7 つ)，肩や頭部に人間とのインタラクションを主目的としたタッチセンサを装備している．

運動機能としては実時間適応歩行生成技術による不整地路面の歩行や，実時間歩容生成技術による外力適用などを実現している．転倒しても自力で起き上がることができる．またカメラからの画像を使っての顔認識，距離計測などをソフトウェアで実現している．画像処理による環境認識と運動機能と組み合わせた障害物回避歩行も行える [Sabe02]．インタラクションに関連の深い音声関連の機能では，マルチマイクロフォンを用いた音源方向同定，話者識別，LVCSR，新規単語（未知語）獲得，音声合成等を実装している．

3.2　EGO アーキテクチャ

QRIO はこの基本構成の上に，EGO アーキテクチャ（Emotionally GrOunded Architecture）と名付けた自律行動制御アーキテクチャを採用している[Fujita01]．音声インタラクションもこのアーキテクチャの上で実現されている．EGO アーキテクチャを図 2 に示す．

EGO アーキテクチャは，Behavior Based Architecture［Arkin98］を発展させたものと捉えることができ，ロボット内部の情動システム（内部欲求，感情）と外界からの刺激（センサ，認識結果）を総合判断して，最も自分を満足させる期待値の高い行動を選択・実行するというアイデアを基本とする．行動はある粒度でシステム内に複数存在し，他の行動との競合の結果，選択されたものが発現することになる．

このアーキテクチャの上で我々は，音声インタラクションに関連する行動も，通常の行動の一つとして位置付けた．もちろん音声インタラクション中に身体の様々な部位を動かすことも可能なので，他の行動と音声インタラクション行動の差は，「音声入出力を多く使う」こと以外に基本的にはない．

EGO アーキテクチャ上で音声インタラクション行動が自律的に発生するためには，「人とインタラクションをする意思」に関連する内部欲求が必要である．我々は，その目的で情動システムを拡張し，特別な内部状態変数を用意した．例えば「ロボットが何か質問する」という行動に対しては「情報に対する空腹感」という量を定義し，その値が満たされていないと質問行動が多くなる，といった制御を試みている．

図 2 の短期記憶には，センサや認識から入力された情報が，時空間の関連性を考慮して統合され，ある期間（短期間）保存される．この機能は，ある行動が別の行動に割り込まれたとしても元の行動に復帰するための外界情報（記憶）を提供する．行動選択実行部も，行動の中断と途中からの再開を行える構成となっており，アーキテクチャ全体が柔軟な行動切替えをサポートしている [Hoshino03].

4.　音声インタラクションに関連する機能とその実装

本節では，QRIO に実装した音声インタラクション関連の機能のうち，特徴的なものを説明する．

4.1　人物学習・同定［Shimomura02］

3 節でも触れたが，我々は画像処理での顔検出・顔認識を開発した．また音声処理では LVCSR のほか，話者

識別，任意の音韻系列を新規単語（未知語）として獲得する技術も開発した［Lucke01］．顔認識，話者識別，新規単語獲得は，オンラインでの学習機能を持っている．ここでは，これらと連想記憶を組み合わせた QRIO の人物学習の枠組みを説明する．

図 3 に人物学習に関連するモジュールの関係を示す．音声認識は対話中の人間の音声を認識するだけでなく，新規単語獲得機能によって，事前登録されていない名前を獲得する（既定義パターンに埋め込まれた音韻系列を新規単語として取り扱う）．話者認識，顔認識は，それぞれ音声，画像に対してこれまで登録されたどの人物のものと最も近いかを判別し，ID を返す．ここではそれぞれ SpeakerID，FaceID とする．

新しい人物の学習は，対話行動の一種である人物学習行動が制御しており，初めて人間に会ったときか，既知の誰かであると仮定して実行した同定処理が失敗したときに発動する．このとき，話者認識と顔認識のモジュールに学習要求を発行して，現在聞こえている声，見えている顔を学習する．またユーザの言った名前を，新規単語として獲得する．その後，ここで発行された新しい SpeakerID と FaceID，獲得した名前を連想記憶に格納する．人物学習の対話例を紹介する（R:ロボット,H:人間）．

R:こんにちは　*（顔を見てFaceID は知っている状態）*
H:こんにちは　*（SpeakerID 取得）*
R:あれ，自信ないなあ，誰？　*（ID 競合して自信がないとき）*
H:私の名前はヤマモトです
R:ヤマモトさんですね？
H:そうだよ　*（聞いたことない名前なので新規人物と判断）*

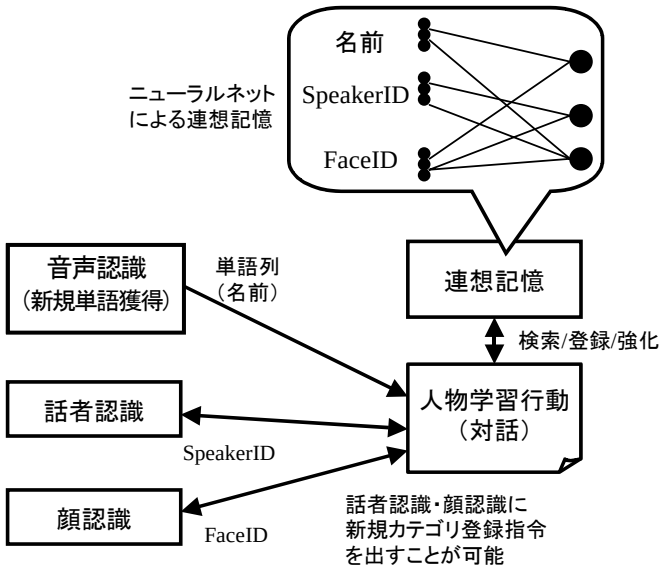


図 3　人物学習に関連するモジュール

R:はじめまして．元気ですか？
U:うん元気だよ　*（顔や声の登録を実施中）*
（その後，連想記憶に登録する）

学習した人物は，それ以降にその声（話者認識），その顔（顔認識），その名前（新規単語を含む音声認識）が観測されれば，連想記憶によって想起される．実環境では認識器の出力が矛盾する場合もあるので，複数の認識結果の出力からある人物を仮定し，必要に応じて名前を対話で確認するという流れで，個人同定を実施する．

4.2　記憶に基づく対話［Aoyama03］

ユーザに長期間会話をしてもらえるための機能のトライアルとして，記憶を用いた対話を実装した．このコンセプトは比較的単純で，ユーザに対してロボットが質問し，回答から獲得した情報を記憶しておき，後にそのユーザ自身あるいは別のユーザに話す，というものである．前述した個人同定機能によって，各個人に別々の対話を提供することができる．

現在の実装では，記憶する情報は事前定義し，フレーム形式で表現している．情報の種類は，できるだけ個人依存の情報であること，また回答が日々変化する可能性の高いものが望ましい，との基準で選定した．実装した情報項目を図 4 に示す．情報獲得の対話例は次のとおり．

R:ゆきこさんのことで教えて欲しいことがあるんだけど聞いてもいいかな
H:いいよ
R:ゆきこさんの好きな食べ物は何なのかな？
H:りんごだよ
R:ゆきこさんの好きなものはりんごでいいのかな？
H:そうだよ
R:教えてくれてありがとう
（自分の記憶に格納し別の場面で自発的に話す）

4.3　シナリオ対話

自己紹介や機能説明，あるいは日替わりの話題提供など，インタラクティブ性の低い会話を，コンテンツ作成者が記述，実行できる対話の枠組みも用意した．これを「シナリオ対話」と呼ぶ．シナリオ対話では，単にロボットが一方的にしゃべるだけではなく，ユー

好きな色	誕生日	食べたいもの
嫌いな色	職業	欲しいもの
好きなスポーツ	好きな食べ物	友達の名前
嫌いなスポーツ	嫌いな食べ物	今の気分
好きな動物	好きな花	今日の天気
嫌いな動物	嫌いな花	

図 4　記憶に基づく対話の記憶項目

ザの応答を受け付ける場所を指定することができる．応答によってシナリオを分岐させることもできる．したがって，記述の枠組みとしては，状態遷移モデルで自由な対話が記述できる．しかし，実用的には分岐を多くしすぎるとコンテンツ管理が困難になるため，分岐を抑え，長めのストーリをロボットが一方的にしゃべる形に利用している．記述には専用の開発環境を使っており，発話とモーションを同時に再生するなどの指定もできる．対話例を次に示す．

R:そういえば，旅行は好きなんだっけ？
H:旅行？ *(想定外なので応答内容は無視して先に進む)*
R:私は旅行するのも趣味のひとつなんだよ．といっても，自分で好きなところには行けないから想像しているだけなんだけどね．
(この後ロボット主導で対話が継続する)

この例では理解できない発話を無視して，先に会話を進めている．このような一方的な方法でも，ある程度はインタラクションが成立することもある．しかし，無視するだけでなく，理解できなくても何かしらの応答をした方がインタラクションとしては適切なことも多いと考え，後に説明する「フレーズ駆動対話」の技術もここに組み込んでいる．フレーズ駆動対話を利用するかどうかは，コンテンツの設計者が決める．

また，会話内容に一貫性があるように見せるため，シナリオのいくつかを，記憶に基づく対話の話題に関連する内容にしている．そして例えば，記憶に基づく対話で食べ物の話をした後，食べ物のシナリオ対話が発現するように関連性を持たせている．

4.4 LVCSR を利用した対話機能

ロボットが人間のすべての発話を理解して対応することはできないという前提に立ち，ルールベースで当り障りのない相槌的な応答や，ユーザの言葉を再利用した応答を生成する機能を実装した．これを「フレーズ駆動対話」と呼んでいる．技術的には Eliza や Parry［Wilks99］で用いられたものと同様であるが，我々はこれを LVCSR の現実的な利用法のひとつと位置付けた．応答生成のルール記述では，正規表現を使った文マッチング，応答文へのユーザ発話の部分流用などが指定できる．ただし，この枠組みを単純に運用しただけでは，内容理解と全くかけ離れた非常に表層的な応答の繰り返しになるおそれがある．我々はこれをシナリオ対話という強いコンテクストの中に埋め込むことで，応答の不自然さを目立たせず，かつシナリオ対話の問題点（広い範囲のユーザ発話には対処できない）をある程度解消しようと考えた．対話例を次に示す．

R:たかしさん，今日は楽しかった？
H:別に，普通かな
R:別に普通かな，*なんだ (相手の発話を流用)*
(この後会話が継続する)

また，LVCSR を使ったエンタテイメント性のあるアプリケーションの別の例として，伝言機能を実装した．内容は，LVCSR で認識した文を指定したユーザに後で伝えるだけである（個人同定機能を使う）．しかし，発話が音声合成で行われることから，単に録音再生するのに比べると，ロボットという個性を持った実体がユーザ間を仲介してくれるという設定の面白さがある．また，直接的な意図ではないが，微妙な認識間違いがかもし出す面白さなど，興味深い機能になっている．

なお，LVCSR はそれを必要とする機能が起動している間だけ稼動させることで，計算量削減を図っている．

4.5 音源方向検出とインタラクションシーケンス

音源方向検出は，QRIO に内蔵されているマイク（主にはそのうちの 4 つ）を利用し，入力される音声信号の時間差分に基づいて行っている．この音源方向検出は，人間の音声だけでなく拍手に対しても適用した．拍手は雑音環境下でも比較的特徴を捉えやすく，実環境での性能が期待できる．この機能によって，後ろや横を向いているロボットに呼びかける（拍手をする），向き直ったロボットに「こっちに来て」といって移動させる（このとき障害物をよけながらやってくる），近くに来たら個人同定を行い，会話を始める，といったインタラクションシーケンスがよりロバストに行えるようになった．

また，ある人物を見ているときに，別の人物の呼びかけによってそちらと会話を始めるような自然な行動も可能である．図 5 に，呼びかけが起こった際の短期記憶の変化を示す[2]．各画面コピーの中央がロボット，そこから出ている V 字の点線が視野を意味する．呼びかけられたことでその方向に視野が移動する（首を向ける）様子と，首を動かしても以前見えていた場所の記憶が保存されていることがわかる．

5. おわりに

本稿では QRIO SDR-4XⅡの音声インタラクションに関連し，そのコンセプトと実装した機能を紹介した．具体的には，実環境での長期的な音声インタラクションを

[2] これは，実際の開発でも利用している短期記憶のビューアの画面である．

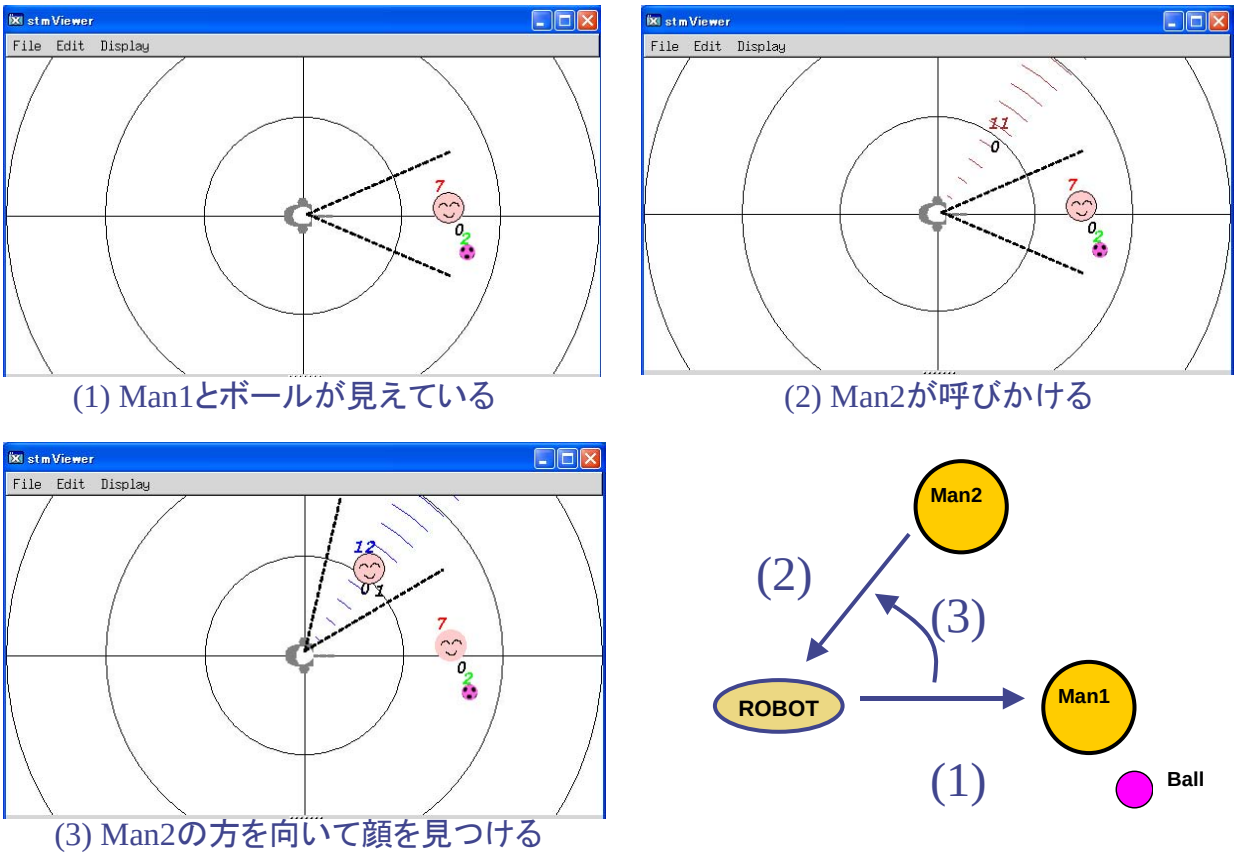


図 5　呼びかけに伴う短期記憶の変化(実際の開発環境での出力)

意識し，個人識別を基本として既存技術といくつかのアイデアを組み合わせた多様な音声インタラクション機能を検討した．現状では，ロボットに呼びかけ，近づいてきてもらい，個人を同定した後，いくつかのエンタテイメント対話が楽しめる，という一連の音声インタラクションシーケンスが実現されている．

QRIO では，音声対話が実環境で困難であることを承知のうえで，あえて現状で利用可能な技術を統合し，音声インタラクションに取り組んだ．もちろん，まだ機能は不足しており，多くの課題が残っている．実際に試用した知見をフィードバックしながら，よりよい音声インタラクション，特に，エンタテイメントインタラクションの検討を進めたい．

謝辞

QRIO の共同開発を行ったソニー(株)エンターテイメントロボットカンパニーの皆様に感謝いたします．

参考文献

[Ohnaka01] 大中他: 人とのインタラクション機能を持つパーソナルロボット PaPeRo の紹介，情報処理学会音声言語情報処理研究会, 37-7 (2001)

[Kanda02] 神田他: 人間と相互作用する自律型ロボット Robovie の評価，日本ロボット学会誌, Vol.20, No.3 (2002)

[Sabe02] 佐部他: ロボットによるステレオ画像を用いた障害物回避と歩行計画，第 8 回画像センシングシンポジウム (2002)

[Lucke01] H. Lucke et al.: Automatic Word Acquisition from Continuous Speech, EUROSPEECH2001 (2001)

[Fujita01] M. Fujita, et al., An Autonomous Robot that Eats Information via Interaction with Humans and Environments, IEEE International Workshop on Robot and Human Interactive Communication (2001)

[Arkin98] R. C. Arkin: Behavior-Based Robotics, The MIT Press (1998)

[Hoshino03] 星野他: パーソナルロボットにおける行動モジュールを用いた行動制御アーキテクチャ，日本ロボット学会第 21 回学術講演会 (2003)

[Shimomura02] 下村他: エンタテイメントロボットと音声対話，人工知能学会言語・音声理解と対話処理研究会, SIG-SLUD-A202-04 (2002)

[Wilks99] Y. Wilks (edit): Machine Conversations, Kluwer Academic Publishers (1999)

[Aoyama03] 青山他: ユーザ固有の情報を獲得・再利用するロボットでの音声対話，人工知能学会言語・音声理解と対話処理研究会, SIG-SLUD-A301-06 (2003)

音響と画像の情報統合を用いた話者追跡と音源分離

Speech event tracking and separation based on the audio and video information fusion

浅野太 [1]，麻生英樹 [1]，原功 [1]，吉村隆 [1]，緒方淳 [1]，市村直幸 [1]，本村陽一 [1]，後藤真孝 [1]，山本潔 [2]

Futoshi Asanoo[1], Hideki Asoho[1], Isao Harao[1], Takashi Yoshimurao[1], Jun Ogatao[1],
Naoyuki Ichimurao[1], Yoichi Motomurao[1], Masataka Goto[1] and Kiyoshi Yamamoto[2]

[1] 産業技術総合研究所，[2] 筑波大学

[1] **National Institute of Advanced Industrial Science and Technology (AIST),** [2]**Tsukuba University**

f.asano@aist.go.jp

Abstract

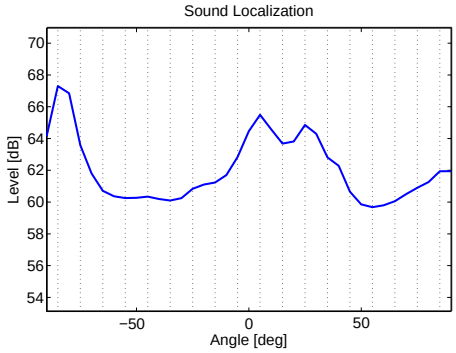
In this paper, a method of detecting and separating speech events in a multiple-sound-source condition using audio and video information is proposed. For detecting speech events, sound localization using a microphone array and human tracking by stereo vision is combined by a Bayesian network. From the inference results of the Bayesian network, the information on the time and location of speech events can be known in a multiple-sound-source condition. Based on the detected speech event information, a maximum likelihood adaptive beamformer is constructed and the speech signal is separated from the background noise and interferences.

1 Introduction

実環境において，音声認識を用いる場合，発話区間の検出が重要である．また，音声認識の前処理として，マイクロホンアレイシステムや他の雑音除去手法 (例えばスペクトルサブトラクション) を用いる場合も，発話区間が検出されていると，性能が格段に向上する場合がある．

発話区間の検出法として，Voice Activity Detector(例えば，[1]) が挙げられるが，例えば，テレビからの音のように，雑音源も音声である場合は，用いることができない．そこで，我々は，マイクロホンアレイからの音響情報に加え，カメラからの画像情報を用いることで，話者から発せられる発話イベントのタイミングと空間的位置を推定する手法を提案している [2, 3]．さらに，この発話区間検出法と適応ビームフォーミングによる音源分離，音声認識におけるモデル適応技術などを組み合わせて，実環境でロバストな音声インターフェースの構築を進めている．本報告では，この音声インターフェースを紹介し，音声認識による評価実験について述べる．

(a) Spatial Spectrum



(b) Video Tracking

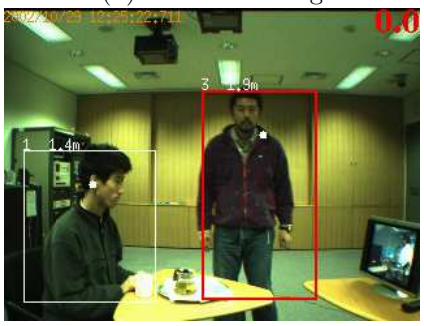


Figure 1: An example of the audio information (spatial spectrum) and the video information (human tracking).

2 音響及び画像情報の抽出

ここでは，マイクロホンアレイ入力から音源位置の推定を行う音源定位と，画像入力から人物位置を推定する人物追跡の手法を簡単に述べる．

2.1 マイクロホンアレイによる音源定位

音源定位には，サブスペース法の一種である MUSIC 法 [4] を，固有値の重みを用いて，広帯域に拡張した方法を用いる [2]．

まず，入力ベクトルを

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

と定義する．ここで，$X_m(\omega,t)$ は，第 m 番目のマイクロホンへの入力信号の短区間フーリエ変換である．これから，次式の空間相関行列を求める．

$$\mathbf{R}(\omega) = E[\mathbf{x}(\omega,t)\mathbf{x}^H(\omega,t)]. \tag{2}$$

これを次式のように固有値展開する．

$$\mathbf{R}(\omega) = \mathbf{E}(\omega)\mathbf{\Lambda}(\omega)\mathbf{E}^{-1}(\omega). \tag{3}$$

ここで，$\mathbf{E}(\omega) = [\mathbf{e}_1(\omega),\cdots,\mathbf{e}_M(\omega)]$ は固有ベクトル $\mathbf{e}_m(\omega)$ からなる固有ベクトル行列，$\mathbf{\Lambda}(\omega) = \mathrm{diag}(\lambda_1(\omega),\cdots,\lambda_M(\omega))$ は固有値 $\lambda_m(\omega)$ からなる固有値行列である．小さい方から $M-N$ 個の固有値に対する固有ベクトルを用いて，狭帯域の場合の MUSIC 空間スペクトルは，次式のように求まる．

$$P(\theta,\omega) = \frac{|\mathbf{g}(\theta,\omega)|^2}{\sum_{m=N+1}^{M}|\mathbf{e}_m^H\mathbf{g}(\theta,\omega)|^2}, \tag{4}$$

ここで，M 及び N は，マイクロホン数及び音源数を表す．$\mathbf{g}(\theta,\omega)$ は，仮想音源 (方向 θ) の位置ベクトルであり，仮想音源位置から各マイクロホンまでの直接音の伝達関数を，その要素に持つ．この仮想音源位置を，探索対象となる空間全体にスキャンすることにより，空間スペクトルを求める．[2] では，この狭帯域空間スペクトル $P(\theta,\omega)$ を，次式のように重みつき平均することで，広帯域の空間スペクトルを算出する方法を提案している．

$$\bar{P}(\theta) = \sum_{\omega=\omega_l}^{\omega_h}\bar{\lambda}(\omega)P(\theta,\omega), \tag{5}$$

ここで重みは，次式で定義されるように，大きい方から N 個の固有値となっており，これは，方向性信号のエネルギーの和を表す．

$$\bar{\lambda}(\omega) = \sum_{n=1}^{N}\lambda_n \tag{6}$$

ただし，固有値は，大きい順にソートされているとする．これにより，SNR の高い周波数帯域に大きな重みが付く．$[\omega_l,\omega_h]$ は，周波数範囲を示す．

図 1(a) は，上述の方法により推定された空間スペクトルである．このスペクトル上で，ピークを検出することにより，音源位置の推定を行うことができる．

2.2 画像による人物追跡

人物追跡を行う手法は，多数提案されている．提案法では，画像上で人物位置をピクセル値で返す手法であれば，いかなる手法でも用いることができる．本報告では，実装の簡単な背景差分を用いている [5]．この方法は，事前に人のいない背景画像を取得しておき，この背景画像と，現在の画像との差分を” 人” として認識する．

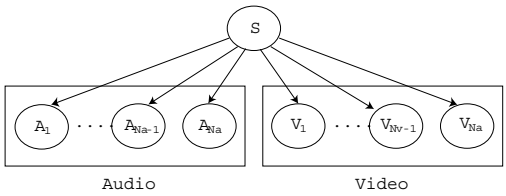


Figure 2: Bayesian network for fusing audio and video information.

3 音響と画像の情報統合

3.1 基本的なコンセプト

前節で述べたように，音源定位の情報から、音源の位置を知ることができる．また，画像による人物追跡の情報から，人物の位置を知ることができる．これらの情報を統合することにより，発話する人物を検出する．

具体的には，音響の観測空間のうち，ある特定の領域において音響イベント (音が鳴る) が発生したかどうかを観測するセンサを，音源定位の情報から，仮想的に作ることができる．画像についても，同様に，画像イベント (人が存在する) が発生したかどうかを検出する仮想センサを作る．これらの仮想センサからの情報を用いて，特定の領域で，音響イベントと画像イベントの共起を，発話イベント (人が話している) と仮定し，検出する．

3.2 情報統合に用いる Bayesian Network

音響イベントと画像イベントの共起を検出する手法として，本研究では，Bayesian Network [6, 7] を用いる．

図 2 に，本研究で用いる，Bayesian Network のトポロジーを示す．入力ノードは、音響及び画像に分かれており，それぞれ，N_a 個 (音響) 及び N_v 個 (音響) のノードを持つ．これらの入力ノードが，音響及び画像の観測空間の特定領域を観測する仮想センサの役割を果たしている．本研究では，入力ノードは，音響イベント及び画像イベントが発生したかどうかに対応した，$\{0,1\}$ の 2 値 (0:発生しない，1:発生した) の状態をとることとする．

一方，出力ノードは，発話イベントの状態に対応する．S を出力ノードを表す記号とし，次のような N_s+1 の状態をとるものとする : $S = \{S_1,\cdots,S_{N_s},NoEvent\}$．状態 $\{S_1,\cdots,S_{N_s}\}$ は，次のような空間位置: $\{S_1,\cdots,S_{N_s}\} = \{-30^\circ,\cdots,+30^\circ\}$ に対応し，例えば，$S=-30^\circ$ の場合は，-30° の位置で，発話イベントが発生したことを示す．一方，$S = NoEvent$ の場合は，観測空間内で，発話イベントが起きなかったことを示す．

3.3 特徴ベクトル

図 3 は，本研究で用いられる音響情報，画像情報とこれに対応する Bayesian Network を示したものである．

音響の観測空間は，$N_a = 17$ の領域 (–90° 〜+90° , 10°

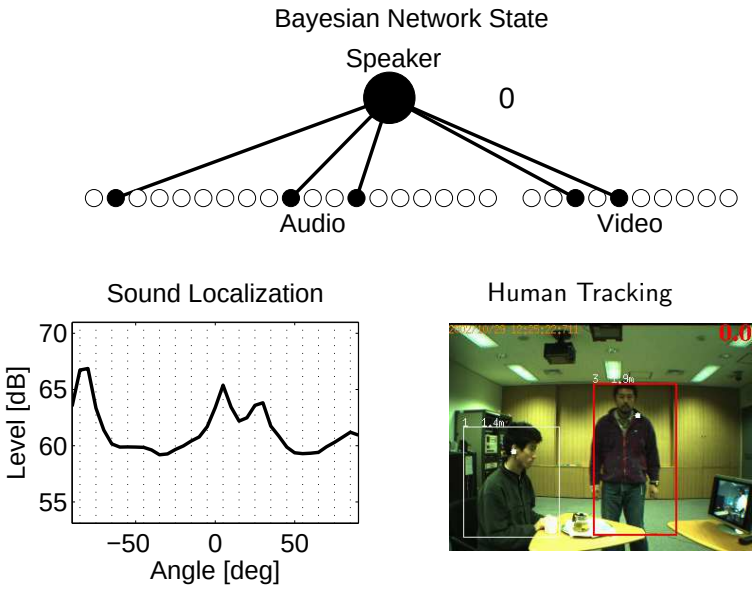


Figure 3: An example of a state of Bayesian network.

毎) に分割され，これらの領域が，音響の入力ノードに対応する．図 3 の左下に示される空間スペクトル中の縦線が，分割された領域を示す．各領域について，空間スペクトルのピークの有無を検出し，ピークが存在すれば”1”，存在しなければ”0” の状態とする．この検出結果から，特徴ベクトル $\mathbf{a}(t) = \{A_1(t), \cdots, \cdots, A_{N_a}(t)\}$ を構成する．ここで，$A_i(t))$ は，時刻 t における第 i 番目のノードの状態 $\{0, 1\}$ を表す．図 4(a) に音響特徴ベクトルの例を示す．この図の縦のスライスが，時刻 t での音響特徴ベクトル $\mathbf{a}(t)$ となる．

画像情報についても同様に，観測空間は，$N_v = 10$ の領域 (1 - 480 pixels, 48 pixels 毎) に分割され，画像の入力ノードに対応づけされている．この各領域において，人物の存在を検出し，存在すれば”1”, 存在しなければ”0” の状態を与え，特徴ベクトル $\mathbf{v}(t) = \{V_1(t), \cdots, \cdots, V_{N_v}(t)\}$ を構成する．図 4(b) に画像特徴ベクトルの例を示す．

3.4 Bayesian Network による推論

続いて，観測値から特徴ベクトルがあたえられた場合，発話ノード S の状態を推定する手法について述べる．S の状態の推定は，特徴ベクトルが与えられた場合の S の条件付確率 $P(S|A_1, \cdots, A_{N_a}, V_1, \cdots, V_{N_v})$ を推定することにより行う．本研究では，S の状態が与えられた場合，音響ノード A_i の状態と，画像ノードの状態 V_j は，条件付独立であると仮定する．これにより，S の条件付確率 $P(S|A_1, \cdots, A_{N_a}, V_1, \cdots, V_{N_v})$ は，次式のように分解することができる．

$$P(S|A_1, \cdots, A_{N_a}, V_1, \cdots, V_{N_v})$$

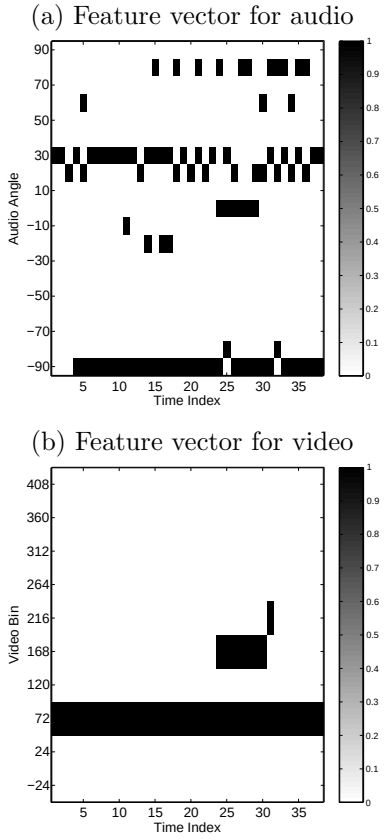


Figure 4: Feature vectors for audio and video.

$$= P(S)\prod_{i=1}^{N_a} P(A_i|S)\prod_{j=1}^{N_v} P(V_j|S)/Z, \qquad (7)$$

ここで,

$$Z = \int_S P(S)\prod_{i=1}^{N_a} P(A_i|S)\prod_{j=1}^{N_v} P(V_j|S)dS. \qquad (8)$$

発話状態の推定は，式 (7) の値を，各時刻ごとに入力する特徴ベクトルと，事前に学習しておいた条件付き確率値とから，評価することにより，行う．ただし，事前確率は，$P(S) = 1$ と仮定する．

3.5 Bayesian Network の学習

条件付確率 $P(A_i|S)$ と $P(V_j|S)$ は，学習用サンプルデータから，事前に学習しておく．学習は，教師付学習であるため，特徴ベクトル $\mathbf{a}(t)$ と $\mathbf{v}(t)$ とに対応した出力ノードの状態 S を教師としてあたえ，条件付確率値を求める．

この条件付確率値の表は，CPT(Conditional Probability Table) と呼ばれ，音響の観測空間と，画像の観測空間との対応を表す働きを持つ．この対応関係は，本来，一対一のものであり，綿密なキャリブレーションにより，この対応関係を正確に求めることも考えられるが，本研究で用いる CPT では，観測値から，この対応関係を学習する．このメリットとしては，雑音による観測値のゆらぎなどが，確率値の重みとして反映された，“ソフト” な対応関係になっていることである．これにより，一対一のリジッドな対応関係を用いるよりも，ロバストな推定が行えることが期待される．

本報告では，学習用サンプルとして，会議室において，話者が単独で発話したデータ (テレビなどの妨害音源はないが，空調などの背景雑音は存在する) を用いた．話者の位置は，–30° 〜 +30° , 5° 毎とし，30 秒程度の時間，断続的に発話する．このデータに対して，人手により発話区間を検出し，検出された区間とそのときの音源位置を，教師として与えた．

図 5 は，この学習サンプルから得られた音響と画像の CPT である．この図から，発話状態に対応する話者の位置 (角度) と，音響情報における位置 (角度)，画像情報における位置 (pixel) との対応関係がわかる．対応関係は一対一ではなく，複数の領域にまたがる場合が見られ，確率値による重みづけがなされている．これにより，人物位置のゆらぎなどをある程度吸収できるものと考えられる．

4 音声インターフェース

ここでは，前節で述べた発話イベント検出を，音源分離と音声認識に適用した音声インターフェースについて述べる．

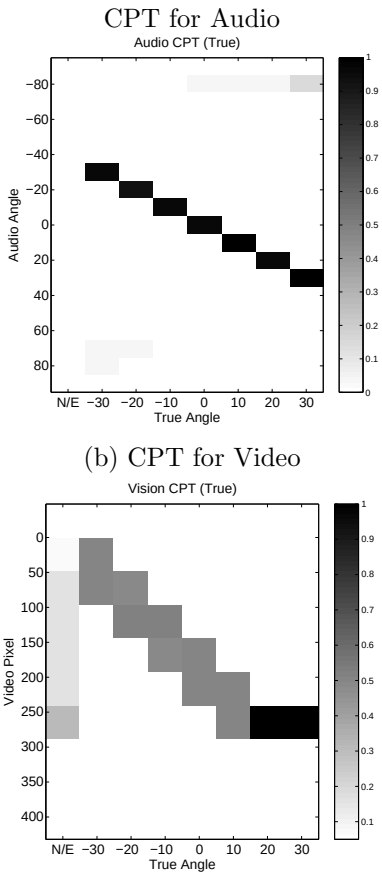


Figure 5: CPT obtained from the real data.

4.1 概要

図 6 に，音声インターフェースの概要を示す．

音響信号は，マイクロホンアレイで観測され，音源位置推定 (Sound Localization) モジュールにより，音源の位置が推定される．一方，ステレオカメラからの画像は，人物追跡 (Human Tracking) モジュールに送られ，画像上の人物位置が推定される．これらの音響及び画像情報は，情報統合 (Information Fusion) モジュールに送られ，ここで，情報が統合され，発話イベントの時間と位置が推定される．

この発話イベントの情報は，音源分離 (Sound Separation) モジュールと音声認識 (Speech Recongnition) モジュールへ送られる．音源分離では，発話イベントの情報に基づいて音源分離フィルタを更新し (詳しくは，次節で述べる)，ターゲットとなる話者の音声を，他の雑音と分離する．一方，音声認識モジュールでは，音源分離モジュールから，雑音の除去された音声信号を受け取り，さらに，発話イベント情報を用いて，音声区間を切り出し，認識する．また，モデル適応 (Model Adaptation) モジュールでは，音源分離において残留した雑音に適応するため，音源分離モジュールの出力にオンラインで適応し，音響モデルを更新する．

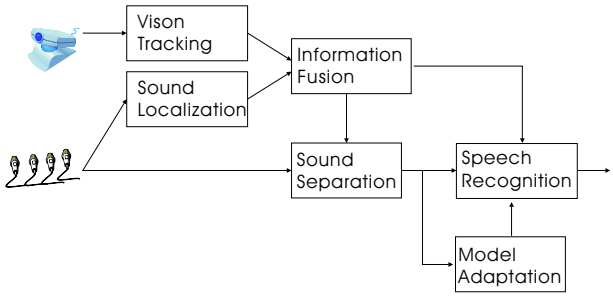


Figure 6: Block diagram of the proposed speech interface.

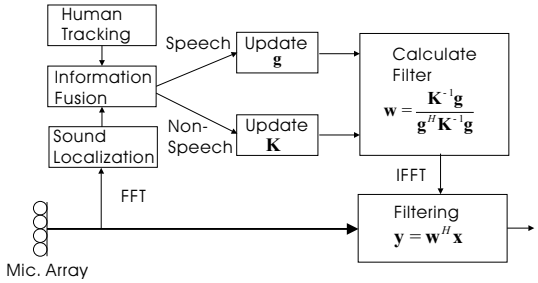


Figure 7: Block Diagram of the ML Beamformer.

4.2 音源分離

音源分離モジュールでは，最尤推定法を用いた，適応ビームフォーマを用いている [8, 9]．最尤推定法では，次式により，音源スペクトルが推定される．

$$y(\omega, t) = \mathbf{w}^H(\omega)\mathbf{x}(\omega, t), \tag{9}$$

ここで，ビームフォーマ係数は，次式により与えられる．

$$\mathbf{w}(\omega) = \frac{\mathbf{K}^{-1}(\omega)\hat{\mathbf{g}}(\omega)}{\hat{\mathbf{g}}^H(\omega)\mathbf{K}^{-1}(\omega)\hat{\mathbf{g}}(\omega)}. \tag{10}$$

行列 $\mathbf{K}(\omega)$ は雑音の空間相関行列であり，非発話区間において推定する．一方，ベクトル $\hat{\mathbf{g}}(\omega)$ は，目的音源の位置ベクトルであり，発話区間において推定する．

図 7 は，最尤推定ビームフォーマのフィルタ係数更新を行うブロック図である．情報統合モジュールにより，発話区間と推定された時間ブロックでは，Bayesian Network の出力ノードの状態 (発話者の位置) に従い，$\hat{\mathbf{g}}(\omega)$ が更新される．一方，非発話区間と推定された場合は，観測された相関行列は，雑音の相関行列 $\mathbf{K}(\omega)$ となり，更新される．

4.3 音声認識及びモデル適応

音声認識モジュールでは，音源分離モジュールから送られた，雑音と分離された音声信号に対し，情報統合モジュールからの発話情報に基づいて，音声区間を切り出し，認識を行う．音源分離により，SNR(signal-to-noise ratio) はかなりの程度改善されているものの，特に，部屋の反射音などは，分離するのに限界があり，雑音が多少残留している．この残留雑音に対して，音声認識のロバスト性を高めるため，音声認識のバックグラウンドで，音声認識の音響モデルの適応を行う．本研究では，MLLR[10] と MAP[11] を組み合わせた手法を用いて，教師なし適応を行った [12].

Figure 8: A scene of experiment.

Table 1: Sound source configuration in Exp.1

Source	Signal	Direction
S1(human)	Speech	-20°
S2(human)	Speech	0°
N1(TV)	Speech+Music	+30°
N2(Loudspeaker)	Music	–90°

5 実験

5.1 実験条件

実験は，中程度の会議室 (残響時間 0.5s) で行った．図 8 に，実験環境を示す．マイクロホンアレイは，円状のもの (直径 0.5m，8 素子，等間隔) を用いている．カメラは，Pointgray Research 社製のステレオカメラ Digiclops を用いている．

5.2 実験 1

実験 1 では，ある程度リアリスティックなシナリオを用意し，発話区間検出のテストを行った．この実験では， 表 1 に示すように，+30° 方向にテレビがあり，音声+音楽を放射している．また， –90° 方向にラウドスピーカがあり，音楽を放射している．話者は 2 人で，一人の話者は，常に観測空間内で椅子にすわり，$t = 15$s 前後で短い発話を行う．他の話者は，観測空間外にいて，$t = 20$s 付近で観測空間に歩いて入り，$t = 25$s 付近で発話して，観測空間外に歩いて出て行く．図 4 は，このシナリオに対する音響及び画像の特徴ベクトルを抽出した結果である．これに対し，Bayesian Network により発話区間を推論した結果を，図 9(a) に示す．図 9(b) は，真の発話区間である．両者を比較すると，$t = 10$s 付近に，一部誤検出はあるものの，発話区間とその位置が良好に推定されているのがわかる．

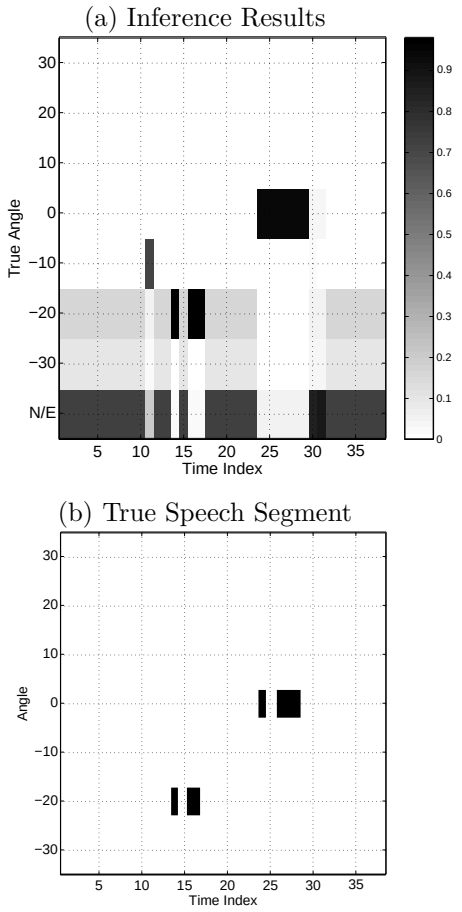


Figure 9: The detected and the true speech events.

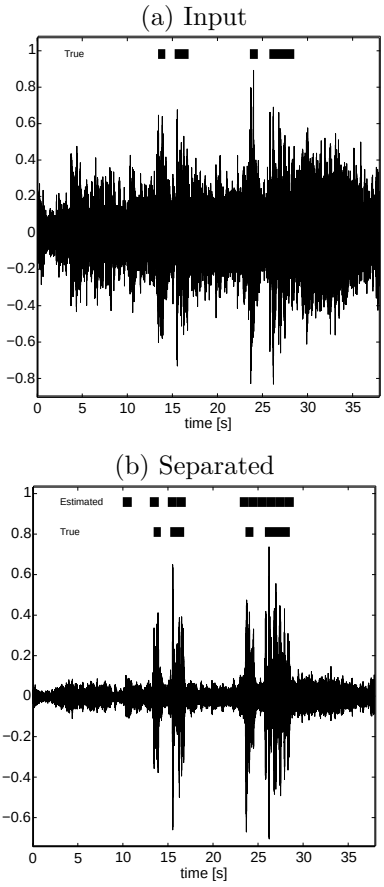


Figure 10: Input/Separated Waveform. The bars indicate the true and the detected speech events.

Table 2: Sound source configuration in Exp.2

Source	Signal	Direction
S1(human)	Speech	+15°
S2(human)	Speech	–25°
N1(Loudspeaker)	Music	–90°

図 10 は，発話区間の情報に基づき，1s ごとに適応ビームフォーマの係数を更新して，音源分離を行った結果である．図中には，検出された発話区間 (推論結果のうち，$= \{-30°, \cdots, +30°\}$ の確率値が 0.7 以上の区間) も示してある．この図から，ほぼ雑音に埋もれた音声波形が，かなりの程度回復しているのがわかる．

5.3 実験 2

実験 2 では，定量的な評価を行うため，区間検出率及び音声認識率による評価を行った．音源配置を表 2 にまとめる．話者 S1 及び S2 は，位置は変えずに，交互に，492 語の日本語単語を発生する．SNR は，概ね 0dB となるよう，雑音源のパワーを事前に調整した．表 3 に，検出率 (R_d)，適合率 (R_f)，再現率 (R_r) を下式に基づいて算出した結果を示す．

$$R_d = \frac{\text{正解した音声/非音声区間数}}{\text{区間総数}} \quad (11)$$

$$R_f = \frac{N_d\text{中の正解した音声/非音声区間数}}{\text{音声区間と検出された区間数 } (N_d)} \quad (12)$$

$$R_r = \frac{N_s\text{中の検出された発話区間数}}{\text{音声区間数 } (N_s)} \quad (13)$$

表中のマージン (margin) とは，検出された発話区間の前後に，0.5 秒 (B) あるいは 1.0 秒 (C) だけ，発話区間を延長したものである．これは，語頭，語尾のパワーの小さい子音などは，提案した発話区間検出により検出できない場合があるためである．この結果から，0.5 秒程度のマージンをとれば，ほとんど全ての音声区間をカバーできることがわかる．一方で，さらに 1 秒までマージンを伸ばすと，非発話区間の混入も多くなる．したがって，次に述べる音声認識実験では，マージンを 0.5s として，実験を行った．

音声認識実験では，マイクロホンアレイの中心からの距離を $r = 1.5$m 及び $r = 3.0$m として，認識実験を行った．認識結果 (word accuracy) を表 4 に示す．word accuracy は次式により定義される．

$$R_a = (H - I)/N_w \quad (14)$$

ここで，H:正解数，I:湧き出し誤り数，N_w:単語数である．Det. は，発話区間に基づいた単語の切り出し，Sep. は音源分離，Adp. は音響モデルの適応を表し，例えば，Det.

Table 3: Detection rate.

Margin	R_d[%]	R_f[%]	R_r[%]
A (0s)	85.7	81.0	82.2
B (0.5s)	74.5	60.3	98.8
C (1.0s)	53.4	44.4	99.8

Table 4: Speech recognition rate.

Condition	Det.	Sep.	Adp.	r=1.5m	r=3.0m
A	On			29.7%	9.6%
B	On	On		79.7%	54.3%
C		On		26.0%	-36.0%
D	On	On	On	91.1%	80.1%

のみが On の場合は，発話区間に基づいた単語の切り出しのみを施し，音源分離，音響モデルの適応は行わなかったことを示す．この表から，前処理系である，音源分離と単語切り出しを行うことにより，$r = 1.5$m で 80%近い認識率が達成されている．一方，音源分離を行っても，単語切り出しを行わず，連続音声認識を行った条件 C の場合は，$r = 3.0$m で word accuracy が負の値になっており，湧き出し誤りが頻発しているのがわかる．音源分離，単語切り出しに加え，音響モデルの適応も行った条件 D では，高い認識率を達成しており，実用に近いレベルであると考えている．

6 リアルタイムシステム

6.1 概要

上述の情報統合に基づく発話検出と音源分離をリアルタイムで行うシステムを構築した．音源定位及び音源分離モジュールは，産総研で開発したリアルタイムハードウエア RASP を用いて実装してある．一方，画像による人物追跡及び Bayesian Network を用いた情報統合は，PC(Dell Prcession530 2.4GHz) に実装されている．また，それぞれのモジュール間通信は，マルチメディア情報の通信のためのプロトコル (RMCP:Remote Media Control Protocol)[13] を産総研で拡張して用いている．

6.2 マイクロホン用ハードウエア RASP

ここでは，マイクロホンアレイ用ハードウエア RASP について簡単に述べる．このシステムは，以下のような開発コンセプトに基づいて作られている．

1. マイクロホンアンプ、Anti-aliasing filter、A/D、信号処理ユニットなど、必要なものを全てひとつの匡体に納める．

2. 信号処理は、FFT や FIR Filter など演算は単純だが、リアルタイム性の高い部分と、音源位置推定や分離フィルタの計算など、複雑な演算を必要とする部分に大別される．前者を**リアルタイム処理**、後者を**学習**と呼び、それぞれに異なるアーキテクチャを採用することにより、システムの効率化と使い易いプログラミング環境を両立させる．

3. 学習におけるプログラミング環境は、一般的な Linux+C 言語とし、DSP のパイプライニングのような職人的知識をほとんど必要としない．

4. リアルタイム処理は、システム側で専用回路を提供することにより、リアルタイム性を保証し、開発者の負担を軽減する．

図 11 に、システムのアーキテクチャを示す．システムは、(1) アナログボード、(2) 信号処理ボード、(3)CPU ボード、の 3 枚のボードから構成される．これらの 3 枚のボードは、図 12 に示す匡体に納められている．外形は、60x135x165 mm である．

アナログボードは、マイクロホンアンプ、Anti-aliasing filter、8 チャネル A/D, 2 チャネル D/A から構成されている．サンプリング周波数は，16kHz が初期値となっているが，変更可能である．単体の A/D,D/A ボードとしても使用できるよう、USB 端子を備えており、ノート PC などに接続して、多チャネルの信号収録だけを行うこともできる．

信号処理ボードは、主にリアルタイム処理を担当し、FPGA(Xilinx 社 VirtexII) を搭載していて、この部分に、フィルタ、FFT などのリアルタイム処理を「回路として」作りこむことができる．FPGA は、アーキテクチャの制約が少なく、FFT やフィルタなど、並列演算を多用することにより高速化できる信号処理モジュールの製作に向いている．現在は，音源分離用の 8ch-in/2ch-out, 1024 タップ/ch の FIR フィルタが実装されている．

CPU ボードは、信号処理のうち、信号処理ボードで行うことのできない、音源位置推定や音源分離フィルタの計算など、高度な演算 (学習) を行うことを目的としている．このボードは、PowerPC を CPU とした、汎用 PC メザニンカードであり、OS としは、Linux(MontaVista) が搭載されており、通常の Linux+C 言語の環境で、プログラミング可能である．また、このボードには、LAN のポートがあり、音源位置推定結果や、音源分離した信号を LAN 経由で音声認識システムに送る、といったことも可能である．汎用メザニンカードを使用しているため、より高速の CPU を搭載したカードに交換することで、アップグレードすることができる．

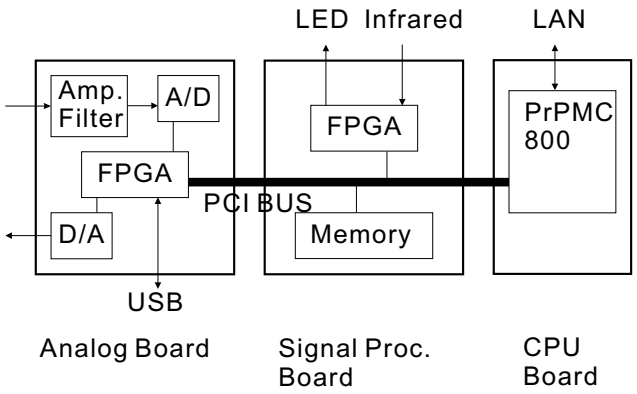


Figure 11: Architecture of the system.


Figure 12: Appearance of the system.

7 まとめ

本報告では，音響と画像の情報統合による発話区間検出システムと，これを用いた音声インターフェースについて述べた．現状では，音源が静止しているものと仮定し，音源定位や発話区間検出の最小単位を 0.5-1.0s 程度としている．今後は，時間分解能を向上させ，移動音源にも対応することを検討している．時間分解能をあげると，音源定位を行うための十分な統計データが得られず，推定分散が大きくなる．これに対処するため，Kalman filter や Particle filter などを用いて，現在は用いていないデータの時間構造の情報も考慮した手法についても，検討を行っている [14].

参考文献

[1] Virginie Gilg, Christophe Beaugeant, Martin Schoenle, and Bernt Andrassy, “Methodology for the design of a robust voice activity detectory for speech enhancement,” in *Proc. IWAENC 2003*, September 2003, pp. 131–134.

[2] Futoshi Asano, Yoichi Motomura, Hideki Asoh, Takashi Yoshimura, Naoyuki Ichimura, and Satoshi Nakamura, “Fusion of audio and video information for detecting speech events,” in *Proc. Fusion 2003*, 2003, pp. 386–393.

[3] Futoshi Asano, Yoichi Motomura, Hideki Asoh, Takashi Yoshimura, Naoyuki Ichimura, Kiyoshi Yamamoto, Nobuhiko Kitawaki, and Satoshi Nakamura, “Detection and separation of speech segment using audio and video information fusion,” in *Proc. Eurospeech2003*, September 2003, pp. 2257–2260.

[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] C. Eveland, K. Konolige, and R. C. Bolles, “Background modeling for segmentation of video-rate stereo sequences,” in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, 1998.

[6] Finn V. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2001.

[7] Y. Motomura and I. Hara, “Bayesian network learning system based on neural netwoks,” *Proc. Int. Symp. on Theory and Application of Softcomputing 2000*, 2000.

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

[9] Futoshi Asano, Masataka Goto, Katunobu Itou, and Hideki Asoh, “Real-time sound source localization and separation system and its application to automatic speech recognition,” in *Proc. Eurospeech*, Aalborg,Denmark, September 2001, pp. 1013–1016.

[10] C.L. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Computer speech and language*, vol. 9, pp. 171–185, 1995.

[11] J.L. Gauvain and C.H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observation of markov chains,” *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[12] Jun Ogata and Yasuo Ariki, “Unsupervised acoustic model adaptation based on phoneme error minimization,” *IEICE Trans. D-II (Japanese)*, vol. J85-D-II, no. 12, pp. 1771–1780, December 2002.

[13] 後藤真孝，根山亮，村岡洋一，“RMCP:遠隔音楽制御用プロトコルを中心とした音楽情報処理,” **情報処理学会論文誌**, vol. 40, no. 3, pp. 1335–1345, March 1999.

[14] 麻生英樹，本村陽一，吉村隆，山本潔，市村直幸，緒方淳，原功，浅野太，“パーティクルフィルタを用いた複数話者の位置と発話状態の追跡,” in *2003* **年ベイジアンネットワークセミナー予稿集**, 2003.

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-0318-5 (11/13)

階層的音源分離に基づく混合音声の認識

Recognition of the Mixed Speech based on multi-stage Audio Segregation

澤田知寛 関矢俊之 小川哲司 小林哲則

Tomohiro Sawada Toshiyuki Sekiya Tetsuji Ogawa Tetsunori Kobayashi

早稲田大学 理工学部

Department of Computer Science, Waseda University

{sawada,sekiya,ogawa,koba}@tk.elec.waseda.ac.jp

Abstract

A novel speech segregation method using a microphone array with the harmonic structure based band selection is proposed and applied to the preprocessing of speech recognition under the existence of disturbance speech. The band selection technique is very effective for audio segregation. However the residual noise spectrum caused by the band selection errors deteriorates the performance. In this paper we try to remove the band selection errors using the harmonic structure in order to improve the performance. The experimental results of double talk recognition with 20K vocabulary showed that the proposed method reduced errors by 18% compared to the naive band selection method.

1 はじめに

ハンズフリー音声認識においては周囲の雑音による認識性能の劣化が問題となる．たとえばカーナビゲーションシステムではエンジン音，助手席の人間の声，車内で流す音楽，車外の喧騒といった雑音が存在する環境下で目的音声を認識しなければならない．そのため，音声強調，雑音抑圧の技術が重要となり，これまで様々な音声強調手法が提案されている [Aoki, 2001] [Okuno, 1996] [Ichikawa, 2003] [水町, 1999] [Kiyohara, 1996]．この中で，伝達特性の異なる 2 系統の入力における帯域毎の振幅成分の比較により，それぞれの成分がどの音源に属するかを推定する，帯域選択に基づく手法 [Aoki, 2001] [Okuno, 1996] は，比較的単純な方法で高性能の分離性能を与えている．青木らの SAFIA は，指向性マイクを利用して 2 系統の差を作り，奥乃らの BiHBSS は，ロボットの頭部伝達特性に基づいて 2 系統の差を作っている．これらの方法は，多くの場合高精度に音源を分離するが，ともに 2 系統の伝達特性の非常に微妙な差によって帯域選択を行うため，音源の位置関係や周波数などの条件によっては良好に動作しない場合がある．

我々はマイクロホンアレー処理を帯域選択の前段に置くという階層的な音源分離の処理構造を提案してきた [Sekiya, 2003]．この手法では，帯域選択に利用する系統毎の指向特性を，アレー処理によって実現する．これによって，従来の手法に比べ特性の差が明確になり，分離性能の向上が期待できる．これまでに，この方法により単体でのアレー処理に比べ，50 %もの誤り削減率を達成してきた．しかし，帯域選択では残響，窓関数などの影響によって帯域選択誤りが生じ，不要な妨害音成分が目的音声に挿入されてしまう．特に，遅延和アレーでは低周波において十分な S/N が得られないため，帯域選択誤りは低周波において生じやすい．そこで本研究では帯域選択を行って得た音声から調波構造を抽出し，この情報を用いて妨害音声の成分を取り除く，3 階層の音源分離手法について検討する．

また，分離を行った音声は人間の聴覚上は違和感無く聞こえたとしても，周波数領域での分離処理によりスペクトル変形が生じてしまい，認識性能が劣化する．そこで，分離音声を用いて音響モデルの適応や学習を行うことでスペクトル変形を吸収する音響モデルを作成し，認識性能の向上を試みる．

以下，2 節で，今回用いた階層的音源分離手法について述べる．そして 3 節で実環境での同時発話を対象とした連続音声認識の結果について述べ，4 節でまとめとする．

2 階層的音源分離

本研究では Figure 1 に示すように 3 階層の音源分離処理を行った．第 1 階層は遅延和アレー処理による音声の強調であり，第 2 階層は帯域選択による音源の分離である．また，第 3 階層は調波構造の抽出に基づく妨害音声の帯

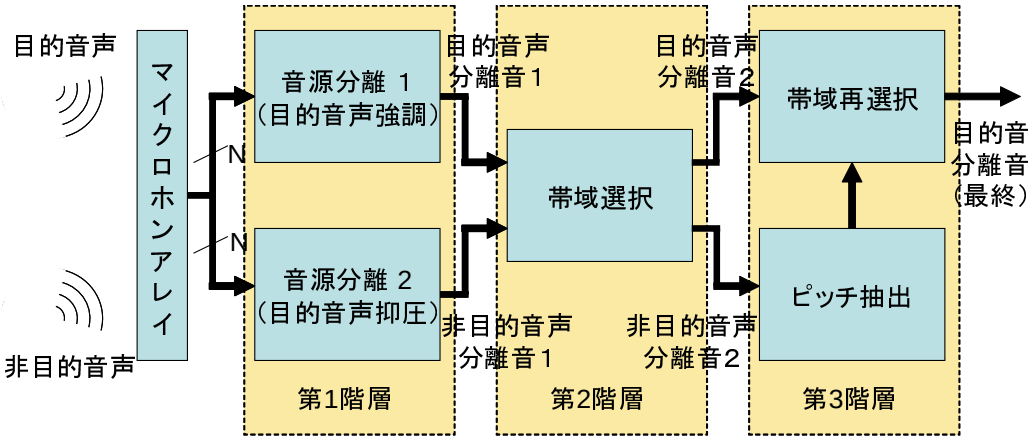


Figure 1: 階層的音源分離の過程

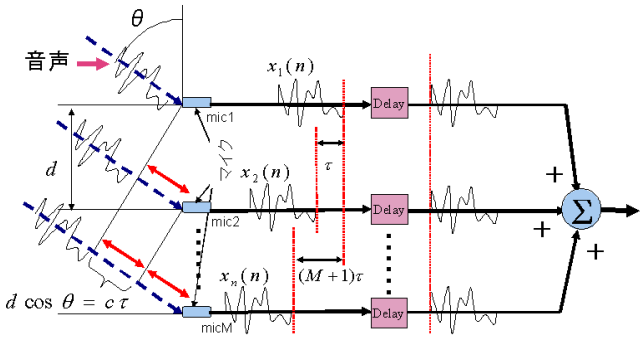


Figure 2: 遅延和アレーの原理

域成分の除去となっている．また，2.4 節では音響モデルの適応や学習について述べる．

2.1 遅延和アレー

第 1 階層では，通常のマイクロホンアレーの処理により複数の音源分離を行う．この段においては，ひとつは目的音声を強調（あるいは妨害音声を抑圧）し，もうひとつは妨害音声を強調（あるいは目的音声を抑圧）する形でアレー処理を行う．アレー処理に，音源分離に使われるものであれば何でも構わないが，本稿における実験では，遅延和アレー (DSA : Delay and Sum Array) を用いた．

遅延和アレーとは，各マイクロホンエレメントで受音した所望波の位相がそろうように制御することにより，所望音の強調を行う手法である．

Figure 2 に示すように θ 方向から到来する周波数 f の複素数平面波に対して，素子数 M，素子間隔 d の等間隔直線形状マイクロホンアレーで受音を行うとする．このとき x_i は x_{i-1} よりも

$$\tau = \frac{d\cos\theta}{c} \qquad (c \text{ は音速}) \qquad (1)$$

だけ時間遅れとなる．ここで，式 (2) のように各素子の受

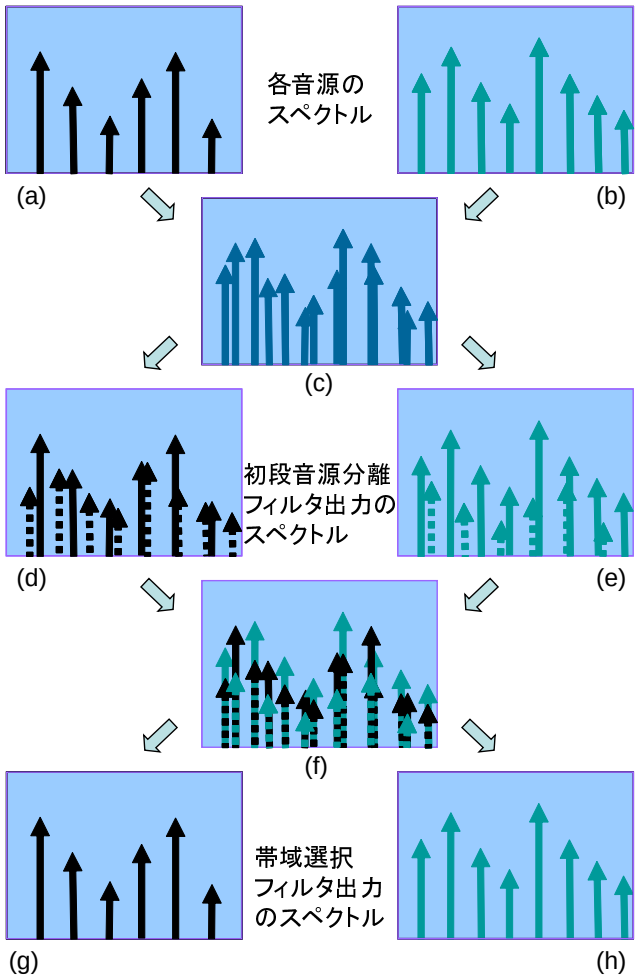


Figure 3: 帯域選択による音源分離の概念図．(a)(b) は各々右左の音源のスペクトル．(c) は受音スペクトル．(d)(e) はマイクロホンアレーにより各々左右の音源を強調したスペクトル．(f) は，(d)(e) を重ね合わせ，帯域毎に大小を比較した図．(g)(h) は比較の結果，各々(d)(e) の値が優位だったものを選択して構成して作ったスペクトル．これらの処理により，(a)(b) のスペクトルが復元できる．

音信号を同相化して加算する．

$$y(t) = \sum_{i=1}^{M} x_i\left(t-(-(i-1)\tau)\right) = \sum_{i=1}^{M} x_i \exp\left(j2\pi f(i-1)\frac{d\cos\theta}{c}\right) \quad (2)$$

このとき θ 方向から到来する信号は M 倍されて出力され，θ 方向以外の方向から到来する信号は同相化されず加算しても強調されない．以上の結果，θ 方向に対して感度の高い指向性が形成される．

2.2 帯域選択

第 2 階層においては，初段で行った 2 系統の音源分離の出力を用いて，帯域選択を行う．

帯域選択の過程を Figure 3 に示す．

Figure 3(a)(b) は各々の音源の振幅スペクトルとする．

これらを，あるマイクロホンエレメントで受音すれば，その混合音の振幅スペクトルは (c) となる．このとき，どの周波数がどの音源に属するものかはわからない．

これを第 1 階層の音源分離を施した結果が (d)(e) となる．即ち，(a) に指向性を向けてフィルタをかけた音の振幅スペクトルが (d) であり，(b) に指向性を向けてフィルタをかけた音の振幅スペクトルが (e) となる．(d) は，(a) の振幅スペクトル (実線) と (b) を抑制した形の振幅スペクトル (破線) との合成によって表される．同様に (e) は，(b) の振幅スペクトル (実線) と (a) を抑制した形の振幅スペクトル (破線) との合成によって表される．

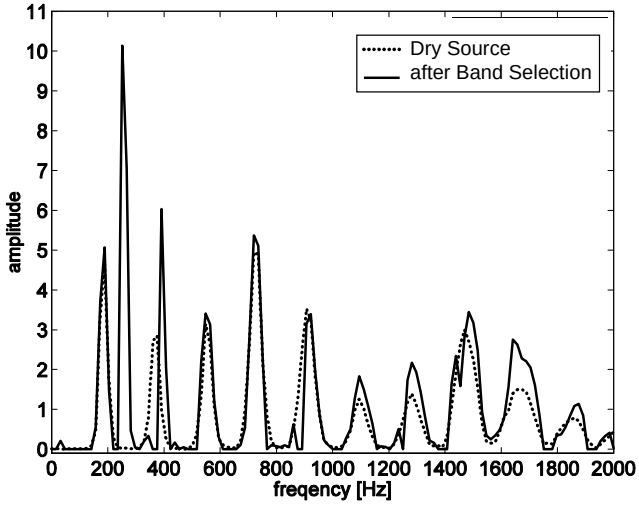
(f) は (d) と (e) のスペクトルを各帯域ごとに大小比較を行なっている図である．(d) における破線は (b) の抑制されたもの，(e) における破線は (a) の抑制されたものであるから，(d) と (e) とが持つ成分は完全に重なり，常に実線の高さが破線のそれを上回ることになる．

(d)(e) の各帯域について，(f) の比較の結果他方より大きな値を与えたもののみを残すことにすると，それぞれ (g)(h) の振幅スペクトルが得られ，音源分離を実現することができる．

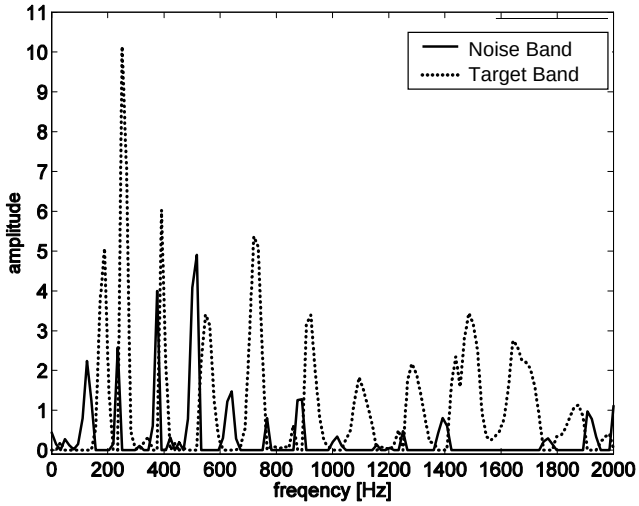
この方法は，目的音声の周波数成分と妨害音声のそれとに重なりがないという仮定が成り立ち，かつ高分解能で周波数分析することができれば，理論的には完全な音源分離ができる．音声の有声部は，比較的疎な線スペクトルであるため，この仮定が成立する可能性は高い．

2.3 基本周波数を用いた雑音成分除去

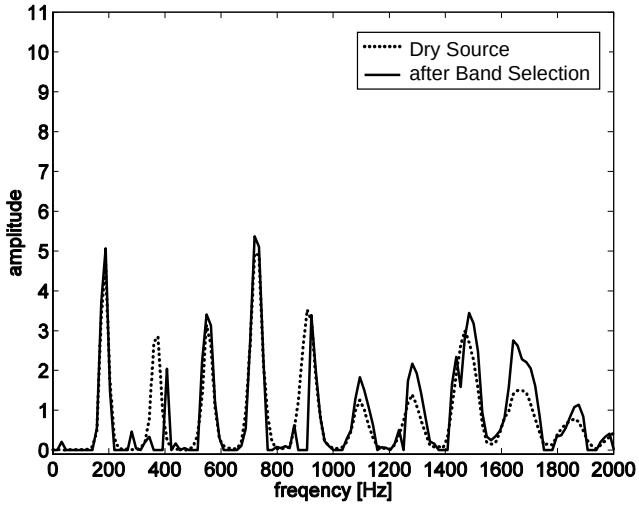
アレー処理と帯域選択を階層的に行なうことにより高精度な分離が可能となる．しかし，各帯域毎にパワーの大小比較のみによって帯域を選択するため，窓関数によるスペクトルの広がりや残響の影響のために選択誤りが生じることがある．選択誤りにより，他音源のスペクトルの特性が残ってしまうので認識性能は劣化してしまう．特に，



(a)



(b)



(c)

Figure 4: 各処理後のスペクトル特性

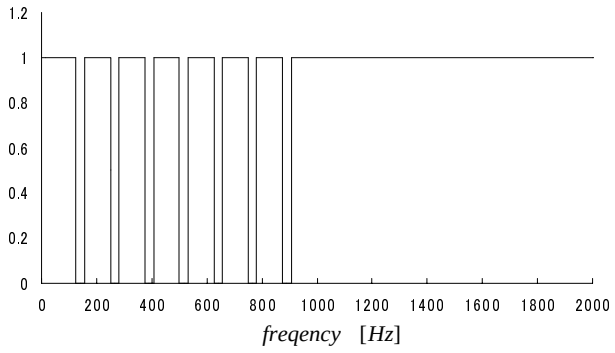


Figure 5: 櫛歯状フィルタ

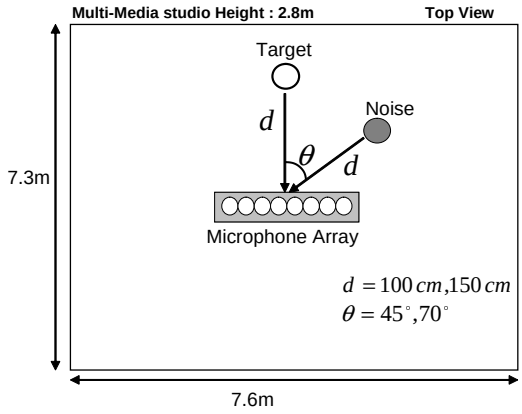


Figure 6: 実験環境

Table 1: アレーマイクの仕様／実験条件

アレー形状	等間隔直線状
素子配置	素子数 8 素子間隔 3cm
素子	無指向性コンデンサマイク
標本／量子化	32kHz，16bit
フレーム長	2048 サンプル (64ms)，ハミング窓
音声	ASJ-JNAS の学習対象話者以外の 男性話者 20 人計 100 文
音源配置	Figure 6 のとおり
位置ベクトル	65536 点 TSP にて測定 [Suzuki, 1995] インパルス長 1024 サンプル

第１階層で遅延和アレーを用いた場合，低域で鋭い指向特性を実現することは難しく，誤りが多く生じる傾向がある．そこでここでは，第３階層として，音源の調波構造を利用することで帯域選択誤りの影響を除去するフィルタを導入する．

二つの音源から発生された混合音声をマイクロホンアレーで受音して帯域選択により目的音声に分離する．Figure 4 (a) は分離した音声とそのドライソースのスペクトル特性であり，(b) は (a) と同一フレームにおける分離後の 2 音声のスペクトル特性である．

Figure 4(a) より，分離音声は概ねドライソースに近い調波構造を持つが，250Hz 付近に不要な周波数成分が含まれていることがわかる．ここで，Figure 4(b) に着目すると (a) における不要成分は，他方の音源のスペクトル特性のピークに対応していることがわかる．つまり，帯域選択誤りは，他方の音源のスペクトル特性においてピークが立つ周波数帯で起きやすいと考えられる．そこで目的音声から，もう一方の音声の調波構造を取り除くことにより帯域選択誤りがもたらす影響を除去することを試みる．

具体的には，妨害音声の基本周波数を求めることで雑音音声のピークが存在する個所を推定し，ピーク個所を削除する Figure 5 のような櫛歯状のフィルタを目的音声に対して施すことで雑音成分の除去を行った．

Figure 4(c) は雑音音声の調波構造を取り除いて帯域選択を行った時のスペクトル特性である．Figure 4(a) および (b) で存在した帯域選択誤りが削除されていることがわかる．

2.4 MLLR 適応，音響モデルの学習

音声認識システムは，接話型マイクで収録された音声に対しては非常に高い認識性能を示すが，実環境下で収録された音声に対しては性能が劣化する．これは，音響モデルの学習を行った環境と実際にデータの収録を行った環境のミスマッチにより生じる．このミスマッチを解消するため，データ収録を行った部屋のインパルス応答を測定し，インパルス応答を畳み込んだ音声を用いて音響モデルの学習を行い認識性能を向上させる試みがなされている [Giuliani, 1999].

同様に，音声認識システムはスペクトル変形を含んだ音声に対しても，音響モデルを学習した音声のスペクトル特性との間のミスマッチにより認識性能が劣化する．本手法で分離した音声は人間の聴覚上は違和感無く聞こえるが，周波数領域での帯域選択により若干であるがスペクトル変形が生じてしまう．そのため認識性能が劣化する．そこで，本研究ではスペクトル変形にロバストな音声認識を行うため，分離音声を用いて MLLR による音響モデルの適応と音響モデルの学習を行う．これにより，スペクトル変形を音響モデルで吸収し，認識性能の向上を試みる，

3 音声認識実験

3.1 収録条件

音声データの収録環境を Figure 6 に示す.

まず，スピーカをマイクロホンアレーから一定距離 d だけ離してアレーマイクの正面に配置し単独音声の収録を行った．d としては，100cm と 150cm の２通り行った.

次に二つのスピーカでの収録を行った．目的音声用のスピーカをマイクロホンアレーの正面，距離 d（ d は 100cm あるいは 150cm）のところに設置し，また妨害音用スピーカをアレーから同じく距離 d で，アレーを挟んだ目的音用スピーカとの角度が θ となるよう配置した． θ としては 45 度と 70 度の２通り行った．よって，合計 4 通りの

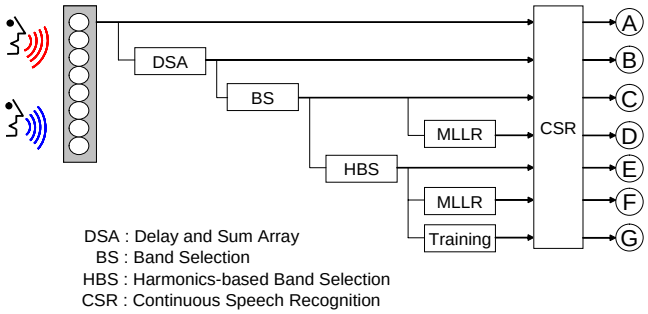


Figure 7: 音声処理方法

音源配置となる．

実験で用いたアレーの仕様，および収録条件を Table 1 に示す．

3.2 評価項目

収録した音声データに対する処理を Figure 7 に示す．ここで DSA は遅延和アレー (Delay and Sum Array)，BS は帯域選択 (Band Selection)，HBS は調波構造を考慮しての帯域選択誤り除去 (Harmonics-based Band Selection)，CSR は連続音声認識 (Continuous Speech Recognition) を表している．MLLR や Training はそれぞれ適応，学習を行った音響モデルで認識することを表している．図に示す 7 種類の処理を施した音声に対し，2 万語彙の連続音声認識を行い，単語認識精度により各々の評価を行なう．なお，帯域選択により生じるスペクトルの不連続性による認識性能の劣化を解消するため，帯域選択や雑音帯域削除を行なったデータには音声認識の前処理として 25dB の計算機ノイズを重畳した [山出, 2002]．

3.3 実験条件

認識の際に用いた音響特徴量を Table 2 に示す．音響モデルには ASJ-JNAS の男性話者約 100 人のクリーン音声 (約 20000 文) から学習を行ったものを用いた．言語モデルは CSRC 提供の語彙数 2 万語の trigram を使用し，認識器には当研究室開発のワンパストライグラムのデコーダ [柴田, 2002] を用いた．

MLLR の適応データは 20 名の認識対象以外の男性話者による音素バランス文を選択し，評価データと同じく 3.1 で述べた条件で収録されている．

音響モデルの学習には，大量の音声データが必要であるが，実環境下で同時発話音声を大量に集めることは困難である．そこであらかじめ収録しておいたインパルス応答をドライソースに畳み込むことにより，擬似的に同時発話音声を作成した．音声データには ASJ-JNAS の男性話者を用いた．話者や音源位置にロバストに学習するために，話者や音源の位置はランダムに選択して学習データを作成した．

Table 2: 特徴量算出パラメータ

プリエンファシス	0.97
フレーム長	25ms
フレーム周期	10ms
周波数分析	12 チャネル 等メル間隔フィルタバンク
特徴量 (25 次元)	MFCC+ΔMFCC+Δ power

Table 3: 単一音源 音声認識実験結果 (単語正解精度:[%])

手法	Baseline	$d = 100$	$d = 150$	平均
接話マイク	94.2	–	–	94.2
遠隔マイク	–	90.9	87.0	89.0

Table 4: 二音源 音声認識実験結果 (単語正解精度:[%])

手法			100[cm] 話者間隔 45[度]	100[cm] 話者間隔 70[度]	150[cm] 話者間隔 45[度]	150[cm] 話者間隔 70[度]	平均
混合音声		A	4.6	8.1	5.8	7.9	6.6
遅延和アレー		B	22.4	30.7	19.3	26.5	24.7
帯域選択		C	76.6	78.3	60.4	72.2	71.9
	適応	D	79.5	80.6	69.6	76.5	76.6
帯域選択 + 雑音帯域削除		E	80.9	82.5	68.6	75.7	76.9
	適応	F	80.9	83.4	74.3	77.1	78.9
	学習	G	81.8	82.9	71.3	77.8	78.5

3.4 実験結果

単一話者での接話型マイク，および遠隔マイクにおける認識率を Table 3 に示す．単一話者の場合，接話型マイクにおける認識率は 94.2% ，遠隔マイクにおける認識率は 89.0%であり，大きな劣化は見られなかった．

二話者の同時発話音声の認識結果を Table 4 及び Figure 8 に示す．同時発話音声では，何も処理を施さない混合音声の場合は認識率が 6.6%である．遅延和アレー処理を施すことで 24.7%となり，若干性能が向上するが十分とは言えない．これに帯域選択を行うことで認識率は大幅に改善され 71.9% になった．これは，遅延和アレーに対して 63%のエラーを削減したことになる．そして，調波構造を考慮して帯域選択誤りを消去する手法を用いた場合の認識率は 76.9%となり，帯域選択のみの場合と比べ 18%のエラーを削減した．このことから雑音源の調波構造を考慮して帯域選択を行うことが有効であると言える．

さらに，分離時に生じるスペクトル歪みに対して音響モデルの MLLR 適応を行うことで認識率は 78.9%となった．クリーン音声で学習した音響モデルに対して 9%のエラー削減にあたり，無処理の場合に対して 77%のエラー削減にあたる．歪みをシミュレートしたデータで音響モデルを学習することで，さらに性能が上がることを期待したが，認識率は 78.5%にとどまった．

4 まとめ

本研究では遅延和アレー処理，帯域選択処理，雑音成分除去処理を階層的を行なうことで，音源分離処理の高精度

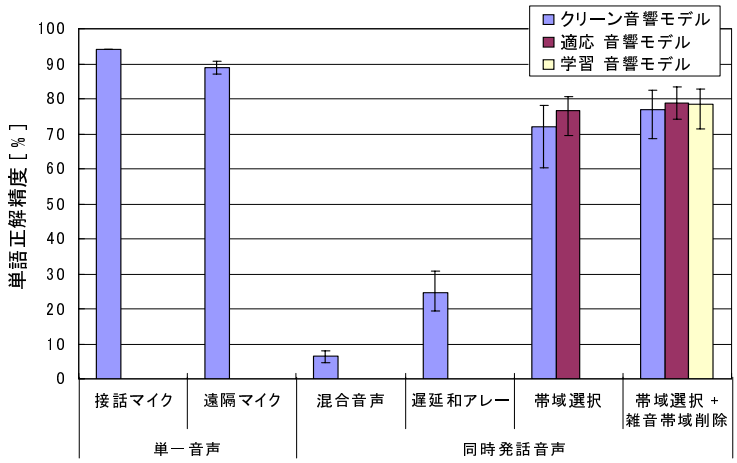


Figure 8: 単語正解精度．棒グラフの値は平均値を示し，棒グラフ上の線は条件毎の最大値と最小値を表す．

化を試みた．また，分離処理で生じる音声のスペクトル変形による認識性能の劣化を防ぐために，分離音声を用いて MLLR による音響モデルの適応と音響モデルの学習を行った．

実環境下での二話者の同時発話音声を対象に連続音声認識実験を行ったところ，調波構造を考慮し帯域選択誤りを除去することで，従来の帯域選択のみの手法と比べ 18 %のエラーを削減することができた．また MLLR による音響モデルの適応や分離音声による学習を行うことで認識率をさらに向上させることができた．

参考文献

[Aoki, 2001] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda: Sound source Segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones, J.Acousic. Soc.vol.22, No.2, pp149–157, 2001.

[Okuno, 1996] H. G. Okuno, T. Nakatani, and T.Kawabata: A new speech enhancement:speech stream segregation, In Proceedings of 1996 International conference on Spoken Language Processing , pp2356–2359, ASA.

[Ichikawa, 2003] O. Ichikawa, T. Takiguchi, and M. Nishimura: Speech Enhancement by Profile Fitting Method, IEICE Trans.Inf.&Syst., vol.E86-D, No.3, March 2003

[水町, 1999] 水町光徳，赤木正人: マイクロホン対を用いたスペクトルサブトラクションによる雑音除去法, 電子情報通信学会論文誌, '99/4, vol.J82-A, No.4, 1999.

[Kiyohara, 1996] K. Kiyohara, Y. Kaneda, S. Takahashi, H. Nomura, and J. Kojima: A Microphone Array System for Speech Recognition, Proc.ICASSP96, 1996.

[Sekiya, 2003] T. Sekiya, S. Serizawa, T. Ogawa, and T. Kobayashi: Speech Recognition of Double Talk using SAFIA-based Audio Segregation, Eurospeech2003.

[Giuliani, 1999] D. Giuliani, M. Matassoni, M. Omologo, P. Svaizer: Training of HMM with filtered speech material for hands-free recognition Proc.ICASSP, vol.1, pp449–452, March 1999.

[Suzuki, 1995] Y. Suzuki, F. Asano, H. Y.Kim, and T. Sone: An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses, J.Acoustic. Soc. Am. vol.97 (2), pp.1119-1123, 1995.

[山出, 2002] 山出慎吾，李晃伸，猿渡洋，鹿野清宏: 雑音に頑健な音韻モデルと教師なし話者適応，信学技法, SP2002-124，pp19–24，2002.

[柴田, 2002] 柴田大輔，小林哲則: ワンパストライグラムデコーダにおける単語履歴の束ね処理に関する検討, 日本音響学会秋季講演論文集，pp151–152，2002.

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-0318-6 (11/13)

ロボットを対象とした散乱理論による三話者同時発話の定位・分離・認識の向上

Improvement of Robot Audition System by Scattering Theory

中臺 一博 †, 奥乃 博 ‡, 辻野広司 †

Kazuhiro Nakadai†, Hiroshi G. Okuno‡, and Hiroshi Tsujino†

† (株) ホンダ・リサーチ・インスティチュート・ジャパン,
‡ 京都大学大学院情報学研究科

† **HONDA Research Insutitute Japan, Co. Ltd.,**

‡ **Graduate School of Infomatics, Kyoto University**

nakadai@nakadai.com, okuno@nue.org, tsujino@jp.honda-ri.com

Abstract

This paper addresses sound source localization, separation and recognition for three simultaneous speeches. We have reported such system for humanoids with a pair of microphones. The system uses interaural phase difference (IPD) and interaural intensity difference (IID) for localization and separation. To estimate IPD and IID mathematically, we proposed *auditory epipolar geometry*. The auditory epipolar geometry, however, has two problems. One is that localization and separation in higher frequency is weak because it estimates IPD rather than IID that is a dominant parameter in higher frequency. The other is that IPD estimation by the auditory epipolar geometry is inaccurate against sounds from side directions. To solve these problems, *scattering theory* in physics is introduced, and accurate estimation of IPD and IID based on the scattering theory is implemented in the system. As a result, improvement of sound source localization, separation and recognition of three simultaneous speeches is attained.

1 はじめに

ヒューマノイドを始め、将来的に日常環境での動作が期待されるロボットでは、実時間・実環境で同時に様々な音を聞き分ける必要がある。このような問題に対し、これまで、動作を知覚向上に利用するアクティブオーディションシステム、ストリームベースの視聴覚統合による実環境・実時間複数人物追跡システム、アクティブ方向通過型フィルタを用いた実時間音源分離システム、これらのシステムを統合した三話者同時発話の認識を報告した[9]。しかし、報告したシステムが用いているロボット頭部の音響モデルは精度が悪く、高周波数域や音源方向が正中面から遠ざかる場合に十分な定位、分離、認識精度が得られなかった。そこで、本稿では、物理学で用いられる散乱理論を適用して、高精度のロボット頭部音響モデルの構築を行い、定位・分離・認識精度の向上を図る。

2 ロボット頭部の音響モデル

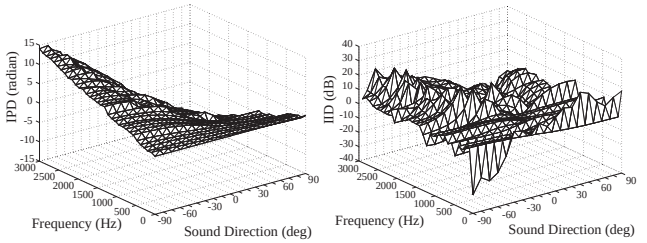
2 本のマイクを耳部に備えたロボットの頭部音響モデルを表す際の主要なパラメータとして、両耳間位相差 (IPD), 両耳間強度差 (IID) が挙げられる。これらは、一般には、頭部伝達関数 (HRTF) の測定によって得られる。しかし、HRTF は、一般に特定の環境 (主に無響室) で計測した離散関数であるため、残響や動的な音響環境の変化に追従させることが難しい。また、各方向からの測定が必要であるため、測定にも時間がかかるといった欠点を抱えている。このため、これまで、音源方向の変化に対して、連続的に IPD を推定できる聴覚エピポーラ幾何を提案した[9]。しかし、IID に関しては、単に正面、左右の 3 方向を推定するだけであったため、IID が支配的な高周波域では、推定精度が低かった。また、後頭部を回ってくる散乱波に対する考慮がされていなかったため、散乱波の影響が出やすい横方向からの音源に対しては、頭部音響モデルの精度が低いという問題があった。このことは、一般に、横方向の音源に対しては、カメラの視野が狭いため、視覚情報も利用できないことを考えると、大きな問題であった。

そこで、本節では、これらの問題を解決するため、散乱理論[5]を導入したロボット頭部の音響モデルを紹介し、既存の手法との比較を通じて、その妥当性を示す。

2.1 散乱理論によるロボット頭部の音響モデリング

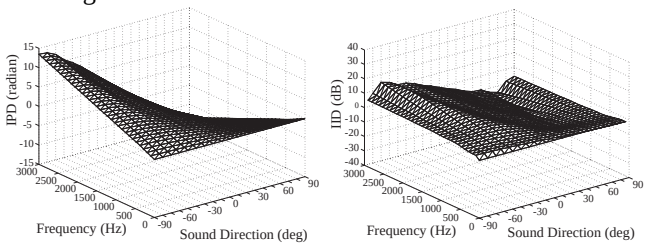
散乱理論を用いると、ロボット頭部による散乱波を考慮しつつ、IPD, IID 双方を計算的に推定することができるため、高精度のロボット頭部音響モデル構築が可能である。なお、使用しているロボット *SIG* の頭部形状は、真球ではないが、本稿では便宜上、これを真球とみなす。また、マイクの位置は、頭部を球体とみなしたときに 180° をなす位置 (つまり球の直径の両端) に設置されているものとする。

まず、ロボットの頭部を半径 a とする。極座標系 (r,θ,φ)



a) IPD b) IID

Figure 1: HRTF Measurement in Anechoic Room



a) IPD b) IID

Figure 2: Estimation by Scattering Theory

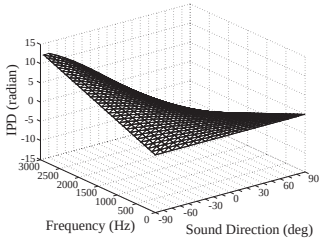


Figure 3: IPD estimation by auditory epipolar geometry

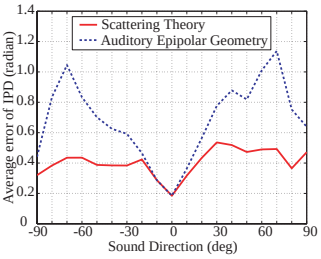


Figure 4: Error of IPD in auditory epipolar geometry and scattering theory

を用いて、点音源 $\boldsymbol{r_0} = (r_0, 0, 0)$ を仮定すると、観測点 $\boldsymbol{r}$ における直接音によるポテンシャルは式 (1) で定義される。

$$V^i = \frac{v}{2\pi Rf} e^{i\frac{2\pi Rf}{v}}, \tag{1}$$

ここで f と v は周波数と音速を表す。また R は音源位置 $\boldsymbol{r_0}$ と観測位置 $\boldsymbol{r}$ の距離 $|\boldsymbol{r_0} - \boldsymbol{r}|$ を示す。

この時、頭部表面 $r = a$ における直接音と散乱音によるポテンシャルは、式 (2) で定義される。

$$\begin{aligned} S(\theta, f) &= V^i + V^s \\ &= -\left(\frac{v}{2\pi af}\right)^2 \sum_{n=0}^{\infty} (2n+1) P_n(\cos\theta) \frac{h_n^{(1)}(\frac{2\pi r_0}{v} f)}{h_n^{(1)'}(\frac{2\pi a}{v} f)}, \end{aligned} \tag{2}$$

ここで、P_n と $h_n^{(1)}$ は第一種 Legendre 関数と第一種球ハンケル関数を示す。

左右のマイクの位置をそれぞれ、$M_l = (a, \frac{\pi}{2}, 0)$ と $M_r = (a, -\frac{\pi}{2}, 0)$ とすると、左右のマイク位置でのポテンシャル S_l と S_r は、式 (3),(4) で表される。

$$S_l(\theta, f) = S(f, \frac{\pi}{2} - \theta), \tag{3}$$

$$S_r(\theta, f) = S(f, -\frac{\pi}{2} - \theta). \tag{4}$$

従って、IPD $\Delta\varphi_s$ と IID $\Delta\rho_s$ は、それぞれ式 (5), (6) によって算出することができる。

$$\Delta\varphi_s(\theta, f) = \arg\left(S_l(\theta, f)\right) - \arg\left(S_r(\theta, f)\right), \tag{5}$$

$$\Delta\rho_s(\theta, f) = 20\log_{10}\frac{|S_l(\theta, f)|}{|S_r(\theta, f)|}. \tag{6}$$

2.2 散乱理論による IPD, IID の推定精度

まず、実際の頭部の音響効果を調べるため、HRTF を計測した。HRTF は、無響室で、音源・マイク間距離を 1 m とし、10° 間隔で ±90° の範囲で、インパルス応答を計測することによって取得した。計測したインパルス応答に対する左右のスペクトルを $Sp_l(f)$ と $Sp_r(f)$ とする (f は周波数)。この時、IPD $\Delta\varphi$ と IID $\Delta\rho$ は以下のように表すことができる。

$$\Delta\varphi = \arg\left(Sp_l(f)\right) - \arg\left(Sp_r(f)\right) \tag{7}$$

$$\Delta\rho = 20\log_{10}\frac{|Sp_l(f)|}{|Sp_r(f)|}. \tag{8}$$

図 1a),b) に示された HRTF から得られた IPD と IID より、次のような音響効果がロボット頭部にあることがわかる。

1. IID は 0° から最大となる 60° 近辺まで増加し、以後は減少する。
2. IPD は 0° から 90° まで単調的に増加する。
3. これらの傾向は周波数帯域によらない。

音源方向が 90° の時、両耳間の音路差が最大となるため、IID は IPD と同様に、90° で最大値を持つことが期待される。しかし、散乱波が後頭部に回り込むため、音源と反対方向では、逆にパワーが大きくなる。従って、実際には、IID は 60° 近辺で最大となる。

次に比較のため、聴覚エピポーラ幾何を用いて IPD と IID の推定を行った。聴覚エピポーラ幾何では、IPD は下式により取得する。

$$\Delta\varphi_e(\theta, f) = \frac{2\pi fr}{v}(\theta + \sin\theta) \tag{9}$$

ここで、f, v, r, θ は、周波数、音速、頭部の半径、音源方向を示す。IID は、簡単に音源方向 θ が正面の場合は 0, 左方向の場合は正、右方向では負と定義する。実際に聴覚エピポーラ幾何によって、推定された IPD を図 3 に示す。また、散乱理論で推定された IPD と IID を、それぞれ、図 2a), b) に示す。これらより、散乱理論による IPD と IID の推定精度が高いこと、特に、IID は、聴覚エピポーラでは、3 方向しか推定できないのに対し、散乱理論では、実測値 (図 1b)) とよく一致している。

図 4 は、散乱理論と聴覚エピポーラ幾何の HRTF に対する IPD 誤差を示している。推定誤差は、散乱理論の方が小さいことがわかる。特に、聴覚エピポーラ幾何では、

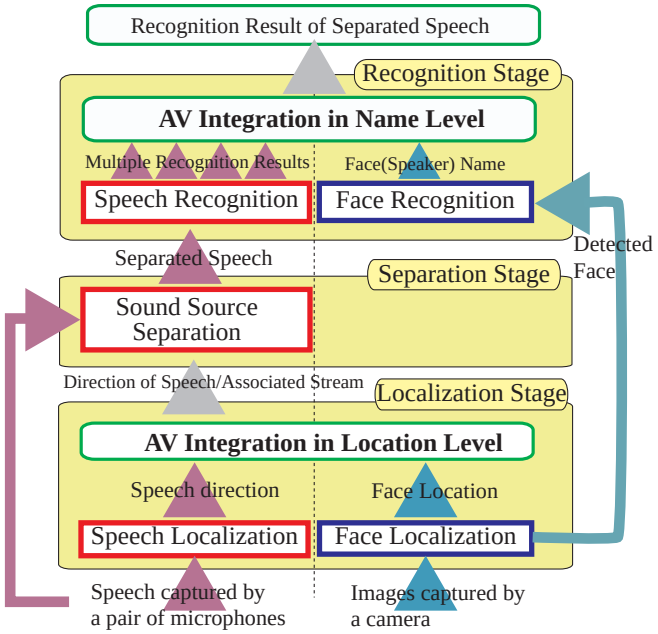


Figure 5: Speech Recognition by two-layered AV Integration

音源方向が 30° 以上になると誤差が顕著になる。これは、聴覚エピポーラ幾何は、後頭部を回り込んでくる散乱波を考慮に入れていないため、IPD の推定誤差は、音源が正面方向から離れるほど大きいためである。これに対し、散乱理論では、横方向の音源でも精度が高い。また、IID を計算的に推定できるため、IID が支配的である高周波領域でも推定精度がよい。

3 散乱理論によるロボット聴覚システム向上

散乱理論による頭部音響モデルを導入し、同時発話認識を行うロボット聴覚システム [9] の精度向上を試みた。テムは、図 5 に示すように、“定位”、” 分離”、“認識” の 3 つのステージからなっている。定位ステージでは、位置レベル、認識ステージでは名前レベルと階層的な視聴覚統合を行い、ロバストな実環境動作を可能にしている。

本稿では、音源定位、音源分離モジュールに、散乱理論ベースの頭部音響モデルを適用し、各々の処理精度向上を通じて、三話者同時発話の認識の向上を試みた。

3.1 定位ステージ – 音声・顔の定位

定位ステージでは、音源定位と顔定位を統合し、ロバストな音源方向を推定する。しかし、顔定位が利用できるのは、正面に近いカメラの画角内のみである。横方向の音源では、顔定位情報が利用できない上に、IPD 推定精度が低いため、定位精度が低くなるという問題があった。そこで、以下のように散乱理論を導入し、精度向上を試みた。

音源定位は、±90° の範囲で水平方向の定位を行う。まず、スペクトル上のローカルピークを抽出し、調波構造を利用してグルーピングを行う。各グループが、一つの音に対応するものとし、各グループに含まれるピークの IPD と IID を計算する。各ピークに対して、5° ごとに式 (5), (6) を用いて、IPD, IID 仮説を生成する。周波数が f_{th} 以下のピークについては、IPD が IID より支配的であるため、IPD に関して仮説と入力の距離を計算する。f_{th} は、両耳間距離によって決まる定数で、*SIG* では、1500 Hz である。

$$d_{\mathrm{IPD}}(\theta) = \frac{1}{n_{f \le f_{th}}} \sum_{f=\mathrm{F0}}^{f_{th}} (\varphi_h(\theta, f) - \varphi_s(f))^2 / f \qquad (10)$$

ここで φ_h と φ_s は、仮説と対応する入力の IPD であり、$n_{f \le f_{th}}$ は、1 グループに含まれるピークのうち、周波数が f_{th} 以下のピーク数である。

f_{th} 以上の周波数では、IID が支配的であるため、IID に関して仮説と入力の距離を計算する。

$$d_{\mathrm{IID}}(\theta) = \frac{1}{n_{f > f_{th}}} \sum_{f=f_{th}}^{f_{max}} (\rho_h(\theta, f) - \rho_s(f))^2 \cdot f) \qquad (11)$$

ここで ρ_h と ρ_s は、仮説と対応する入力の IID である。$n_{f > f_{th}}$ は、1 グループに含まれるピークのうち、周波数が f_{th} 以上のピーク数である。

次に、式 (10),(11) で得られる距離 $d_{\mathrm{IPD}}, d_{\mathrm{IID}}$ を確率密度関数を用いて、確信度 $P_{\mathrm{IPD}}, P_{\mathrm{IID}}$ に変換する。

最後に IID と IPD の確信度を式 (12) で示される Dempster-Shafer 理論[2]を用いて統合し、最大の $P_{\mathrm{IPD+IID}}$ を持つ θ を音源方向とする。

$$\begin{aligned} P_{\mathrm{IPD+IID}}(\theta) \;=\; & P_{\mathrm{IPD}}(\theta) P_{\mathrm{IID}}(\theta) + (1 - P_{\mathrm{IPD}}(\theta)) P_{\mathrm{IID}}(\theta) \\ & + P_{\mathrm{IPD}}(\theta)(1 - P_{\mathrm{IID}}(\theta)) \end{aligned} \qquad (12)$$

顔定位は、肌色抽出と相関演算に基づくパターンマッチングの組合せで顔領域を検出し、検出領域の 3 次元空間座標変換により、実現している [4]。

次に、位置レベルでの視聴覚統合を行うために、各時刻の定位情報を時系列に接続し、音源、顔ごとにストリームを形成する。このストリーム生成により、一時的な定位エラーが訂正される。また、複数のストリームをアソシエーション機構を用いた視聴覚統合[9]を行い、視聴覚情報の曖昧性を相互補完する。最終的に、ストリームから得られる音源の方向情報が分離ステージに送られる。

3.2 分離ステージ – 音声分離

分離ステージでは、アクティブ通過型フィルタ (ADPF) [9] による実時間音源分離を行う。ADPF は、指定方向からの音響信号を IPD, IID を利用して、サブバンド選択を行うことにより、音源分離を行うフィルタである。このため、入力情報である音源方向の精度が分離精度に大きく依存する。従って、位置レベルの視聴覚統合によって、ロバストな音源方向を取得することは音源分離でも有効である。また、サブバンド選択では、IPD, IID を利用するため、IPD, IID を推定するためのロボット頭部音響モデルの精度も大きく分離精度に影響する。このため、散乱理論に基づく高

精度な頭部音響モデルを導入することも分離精度の向上に大きく寄与する。具体的には、以下のようなアルゴリズムで分離を行っている。

1. 音源方向 θ から、式 (5),(6) を用いて、IPD $\Delta\varphi_s(\theta)$, IID $\Delta\rho_s(\theta)$ を推定する。
2. 音源方向 θ に従い、ADPF の方向通過幅 $\delta(\theta)$ を選択する。δ は、音源定位の精度に応じて、正面方向で最小値、90° の方向で最大値を持つ実験的に定めた。ここで、便宜上、$\theta_l = \theta - \delta(\theta)$、$\theta_h = \theta + \delta(\theta)$ とする。
3. 入力信号の IPD, IID と照合し、以下の条件を満たすサブバンドを選択する。

$$\begin{aligned} f \le f_{th} &: \Delta\varphi_s(\theta_l, f) \le \Delta\varphi(f) \le \Delta\varphi_s(\theta_h, f), \\ f > f_{th} &: \Delta\rho_s(\theta_l, f) \le \Delta\rho(f) \le \Delta\rho_s(\theta_h, f). \end{aligned} \quad (13)$$

4. 選択されたサブバンドから波形の再合成を行い、音源方向 θ からの信号を分離する。

3.3 認識ステージ – 分離音声と顔の認識

認識ステージは３つのモジュールからなる。1 つ目は、複数の方向話者依存音響モデルを利用した音声認識である。音響モデル数と同数の音声認識が並列実行され、一つの入力に対して、複数の結果が出力される。2 つめは顔認識であり、顔認識結果の第 3 位までのリストとそれぞれの確信度が出力される。3 つめは名前レベルでの視聴覚統合、つまり、音声と顔認識の統合である。音声認識の複数の結認識果と顔認識の話者名が統合される。

3.3.1 複数の音響モデルを用いた音声認識

ADPF による分離音の特徴が方向ごとに異なること、また顔認識による話者の情報を効果的に利用するため、方向話者依存音響モデルを用いる。方向話者依存音響モデルは、方向ごと、話者ごとの音声情報を元に構築される。本稿では、水平方向に $\pm 90°$ の範囲で、10° おきに 17 方向、話者 3 名 (男性 2、女性 1) を扱い、計 51 個の音響モデル (トライフォン) を HTK(Hidden Markov Model Toolkit) [1] を用いて作成した。また、語彙数は、数字、色、フルーツなどからなる 150 語を用いた。

音声認識エンジンには “Julian” [10]を用いる。Julian は、認識結果に対して尤度に応じたスコアが出力可能なので、これを確率密度関数を用いて、確信度 P_s に変換する。この 51 個の認識結果と確信度が名前レベルの視聴覚統合モジュールに送られる。

3.3.2 顔認識

音声認識における視聴覚統合にはリップリーディング[6, 8]を用いるのが一般的である。しかし、ロボットでは、距離が遠くなると、良好な解像度を得られず、唇を抽出できなくなるためリップリーディングは、必ずしも有効ではない顔は、一般的に唇より抽出しやすいため、顔認識はリップリーディングより利用しやすいといえる。顔認識は、既存のシステム[9]を利用している。第 3 位までの結果と確信度 P_v を名前レベルの視聴覚統合モジュールに送る。

[1] http://htk.eng.cam.ac.uk/

3.3.3 名前レベルの視聴覚統合

名前レベルの視聴覚統合モジュールは、51 個の認識結果を顔認識結果と統合し、最尤の結果を出力する。投票ベースで、複数の音声認識システムの結果を統合する手法として ROVER[3] が挙げられる。本稿では、多くの方向に対応するため、51 個の方向依存音響モデルを利用しており、51 個の音声認識結果を統合する必要がある。このような場合、誤認識も増加するため、単純投票や多数決などを用いた統合では、誤認識結果がシステムに悪影響を及ぼしてしまう。実際、投票ベースの統合を行った結果、10 個程度の認識結果の統合には有効だったが、51 個の認識結果を統合する今回のケースでは、うまく統合ができなかった。そこで、式 (14) で示す各方向話者依存音響モデルの音声認識率に基づく統合を行った。

$$\begin{aligned} V(p_e) \;=\; & P_v(p_e) \left(\sum_d r(p_e, d)\; v(p_e, d)\; P_s(p_e, d) \right. \\ & \left. +\; \sum_p r(p, d_e)\; v(p, d_e)\; P_s(p, d_e) - r(p_e, d_e)\; P_s(p_e, d_e) \right). \\ v(p, d) = & \begin{cases} 1 & if \quad Res(p, d) = Res(p_e, d_e), \\ 0 & if \quad Res(p, d) \neq Res(p_e, d_e). \end{cases} \end{aligned} \quad (14)$$

ここで、d_e は位置レベルの視聴覚統合で得られた音源方向である。p_e は、評価べき話者名である。$r(p, d)$ は、話者 p_e, 方向 d_e の入力に対して話者 p, 方向 d の音響モデルを用いた場合の認識率であり、$Res(p, d)$ は、話者 p, 方向 d の音響モデルを用いた場合の認識結果である。$P_v(p_e)$ は、顔認識で得られた確信度である。最終的に最大の $V(p_e)$ を持つ、話者 p_e と認識結果 $Res(p_e, d_e)$ が出力される。

4 実験と評価

以下の実験により、散乱理論導入の効果を示す。使用した部屋は、$3\,m \times 3\,m$、残響時間は 0.2 – 0.3 秒、スピーカとロボットの距離は 1 m とした。

1. 単音 (調波構造音) の定位と背景雑音からの分離
2. 三話者同時発話に見る孤立単語認識

実験 1 に関しては、100 Hz の調波構造音 (倍音数 30) をスピーカから出力した。定位実験では、音源方向を 10° 単位で 0° から 90° に変化させ、聴覚エピポーラ幾何を用いた場合と散乱理論を用いた場合を比較した。なお、従来システムでは、IPD 推定に式 (9) を用いており、IID からは正面、左右の 3 方向を推定を行っている。分離実験では、音源方向を 0°, 30°, 60°, 90° とし, ADPF の方向通過幅を ±5° おきに ±5° から $\pm 90°$ に変化させ、HRTF を用

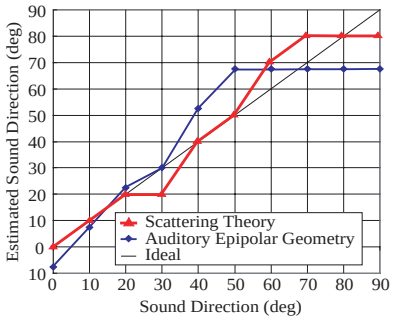


Figure 6: Localization of Harmonic Sound of 100 Hz

いた場合と聴覚エピポーラ幾何を用いた場合を比較した。HRTF を用いる場合は、式 (13) の代わりに式 (15) をサブバンド選択に使用した。

$$\begin{aligned} f \le f_{th} \quad &: \quad \Delta\varphi_H(\theta_l) \le \Delta\varphi \le \Delta\varphi_H(\theta_h), \\ f > f_{th} \quad &: \quad \Delta\rho_H(\theta_l) \le \Delta\rho \le \Delta\varphi_H(\theta_h) \end{aligned} \tag{15}$$

聴覚エピポーラ幾何を用いる場合は式 (16) をサブバンド選択に使用した。

$$f \le f_{th} : \Delta\varphi_e(\theta_l) \le \Delta\varphi \le \Delta\varphi_e(\theta_h). \tag{16}$$

それぞれにおいて、式 (17) で定義される音源抽出率 R を計測した。

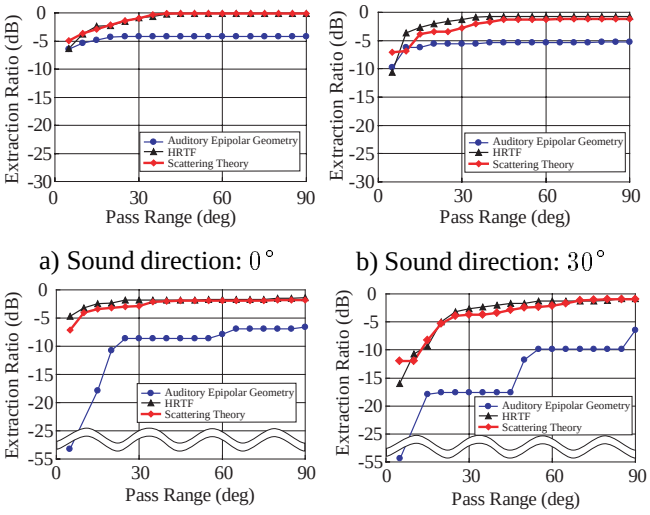
$$R = 10\log_{10}\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}(|sp(i,j)| - \beta|sp_o(i,j)|)^2}{\sum_{j=1}^{n}\sum_{i=1}^{m}(|sp(i,j)| - \beta|sp_s(i,j)|)^2}. \tag{17}$$

ここで、$sp(i,j), sp_o(i,j), sp_s(i,j)$ は、元信号、ロボットのマイクで収音した信号、ADPF で抽出した信号のスペクトルである。また、m と n はサブバンド数とサンプル数であり、β は、元信号に対する観測信号の減衰率である。

定位の結果を図 6 に示す。0°, 30°, 60°, 90° の抽出結果をそれぞれ図 7a) – d) に示す。

図 6 において、散乱理論ベースの定位は聴覚エピポーラ幾何ベースの定位より精度がよい。これは、散乱理論ベースの定位では、高周波域で IID を効率的に利用できることを示している。両者とも定位は、音源方向が正面から離れるにつれ、悪化している。これは、両耳聴の場合、音源が正面方向で精度が高く、横方向では精度が低いという聴覚中心窩と呼ぶ現象に起因すると考えられる。しかし、定位誤差は散乱理論ベースのシステムの方が小さい。特に音源方向が 50° 以上の場合は、この傾向が顕著である。これは、図 4 に示したように IPD の推定誤差が横方向で大きくなるという結果と一致する。

図 7a) – d) では、散乱理論ベースシステムは HRTF ベースシステムとほぼ同様のパフォーマンスを示している。これは、IPD と IID を精度よく推定できる散乱理論が音源抽出に対して有効であることを示している。散乱理論ベースのシステムは、HRTF ベースのシステムと比較し、事前測定が不要であるという利点がある。また、ADPF の方向通



a) Sound direction: 0° b) Sound direction: 30°

c) Sound direction: 60° d) Sound direction: 90°

Figure 7: Extraction of Harmonic Sound of 100 Hz

過幅が同じであっても、散乱理論ベースシステムの方が音源抽出率が良いことから、ADPF の方向通過幅を狭く取ることができる。通過幅が狭いと背景雑音の除去に有効に働くため、この点も大きな利点である。聴覚エピポーラ幾何ベースの音源抽出のパフォーマンスは式 (16) で、$f > f_{th}$ の場合を考慮に入れていないため、散乱理論ベースのものより悪い。散乱理論はベースの音源抽出は f_{th} 以上の周波数で精度の高い IID を使ってサブバンド選択が行えるため精度がよい。図 7c), d) では、横方向で IPD の推定精度が良くないため、聴覚エピポーラ幾何ベースの音源抽出はさらに悪化している。一方、散乱理論ベースの音源抽出は、HRTF ベースとほぼ同等に精度が高い。

次に、実験 2 として、HRTF、聴覚エピポーラ幾何、散乱理論を用いた場合の三話者同時発話の認識を行った。実験では、実際の人物ではなく、スピーカの前に各話者の写真を貼って代用した。図 8 に結果を示す。横軸はスピーカ間の角度、縦軸が 3 つのスピーカから出力された単語の平均認識率である。また、“AEG”, “HRTF”, “ST” はそれぞれ、聴覚エピポーラ幾何、HRTF, 散乱理論を用いた場合の結果である。

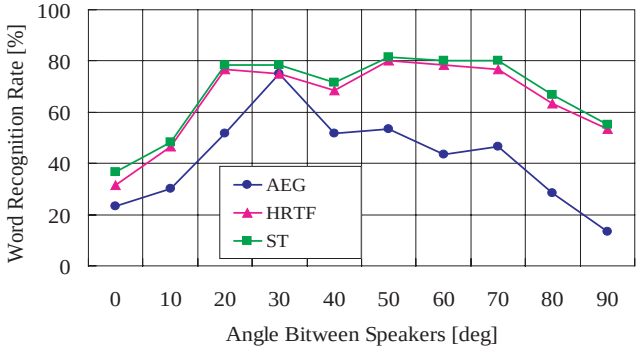


Figure 8: Word recognition rate

スピーカ間の角度差が 20° から 70° では、HRTF と散乱理論ベースの手法を用いた場合、音声認識率は約 80%で

一定しており、90° に近い (つまり真横からの音源がある) 場合でも認識率の変化が少ないことから、聴覚エピポーラ幾何と比較し、横方向の音源に対して良好な精度が得られていることを示している。HRTF と散乱理論ベースの手法では、ほとんど変化がないが、HRTF は測定が必要であることを考慮すると、散乱理論ベースの手法は、ロボットへ適用に適しているといえる。

5 考察と課題

実時間、実環境で処理を行わなければならないロボット聴覚システムは、観測、認識、動作などすべての処理においてノイズが混入することを考慮しなければならない。特に、聴覚の場合、一般に視覚ベースのシステムと比較して、広範囲な情報が取得できる分、曖昧性が大きく、システムの精度やロバスト性を向上させることが大きな課題である。こうした精度やロバスト性を向上という問題に対して、本稿では位置レベル、名前レベルの二階層のモデルを実際に構築し、三話者の同時発話認識に適用し、その有効性を示した。

よりロバストな認識を行うため、これを発展させた図 9 に示すような視聴覚統合モデルを目指している。このモデルでは、信号レベル、音素・口経索レベル、位置レベル、名前レベルといった様々な情報レベルにおいて 対応する視聴覚間の統合が行われ、また、異なる情報レベル間でも統合が行われる。これにより、一部の情報が失われていたり、曖昧性が大きい場合であっても、システム全体として、最尤の内部状態を構築することにより、これを解決するモデルである。

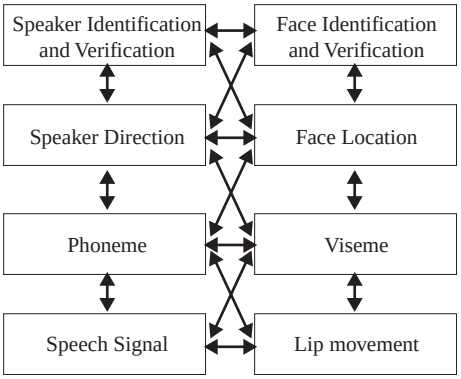


Figure 9: Hierarchical AV Integration Model

この他にも、より汎用的な処理に向け、上下方向の音源定位、調波構造を積極的に利用したり、重複周波数成分を扱うことができる音源分離の検討が必要であろう。また、話者非依存の音響モデルを用いた未知話者の音声認識、ワードスポッティングなどの技術を導入した複雑な文章の音声認識、雑音が混入することを前提とし、missing data や missing feature など分離データの性質を考慮した音声認識エンジンの改良[1, 7]も大きな課題である。

6 結論

本稿では、散乱理論を導入して高精度なロボット頭部の音響モデルを構築することにより、2 本のマイクを備えたロボットを対象に混合音の定位・分離・認識を行うロボット聴覚システムを向上させた。その結果、音源定位・分離精度が向上し、分離音の音声認識率も向上できた。今後は、現状の課題を踏まえ、様々な聴覚要素技術をインテグレーションしたシステムとして、より高精度で高次処理が可能なロボット聴覚の実現に向けた研究を行う予定である。

謝辞

本研究に対して、有意義な議論をいただいた東工大の松浦大輔氏に感謝する。本研究は主に著者が科技団 ERATO 北野共生システムプロジェクト在籍中に行われたものである。統括責任者である北野宏明氏に感謝する。また、*SIG* に関する一連の研究は “http://winnie.kuis.kyoto-u.ac.jp/SIG/” を参照されたい。

参考文献

[1] J. Barker, M.Cooke, and P.Green. Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *EUROSPEECH-2001*, volume 1, pages 213–216. ESCA.

[2] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[3] J.G. Fiscus. A post-processing systems to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *ASRU-97*, pages 347–354. IEEE, 1997.

[4] K. Hidai, H. Mizoguchi, K. Hiraoka, M. Tanaka, T. Shigehara, and T. Mishima. Robust face detection against brightness fluctuation and size variation. In *IROS-2000*, pages 1397–1384. IEEE.

[5] P. Lax and R. Phillips. *Scattering Theory*. Academic Press, NY., 1989.

[6] J. Luettin and S. Dupont. Continuous audio-visual speech recognition. In *Proceeding of 5th European Conference on Computer Vision (ECCV-98)*, volume II of *Lecture Notes in Computer Science*, pages 657–673. Springer Verlag, 1998. IDIAP-RR 98-02.

[7] P. Renevey, R. Vetter, and J. Kraus. Robust speech recognition using missing feature theory and vector quantization. In *EUROSPEECH-01*, volume 2, pages 1107–1110. ESCA, 2001.

[8] P.L. Silsbee and A.C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 1996.

[9] 中臺 一博, 奥乃 博, and 北野 宏明. アクティブオーディションによる複数音源の定位・分離・認識. In *AI Challenge* **研究会**, pages 1043–1049. 人工知能学会, 2002.

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

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-0318-7 (11/13)

人間との円滑なコミュニケーションを目的とした ヒューマノイドロボットの心理モデルの構築

Construction of Mental Model of Humanoid Robot for Natural Communication with Human

三輪　洋靖 [1]，伊藤　加寿子 [2]，髙信　英明 [3, 4]，高西　淳夫 [1, 4]

Hiroyasu Miwa [1], Kazuko Itoh [2], Hideaki Takanobu [3, 4], Atsuo Takanishi [1, 4]

*1 早稲田大学，*2 早稲田大学大学院，*3 工学院大学，*4 早稲田大学ヒューマノイド研究所
*1 Department of Mechanical Engineering, Waseda University
*2 Graduate School of Science and Engineering, Waseda University
*3 Dept. of Mechanical Systems Engineering, Kogakuin Univ.
*3 Humanoid Robotics Institute, Waseda University

takanisi@waseda.jp

Abstract

The authors have been developing humanoid robots in order to develop new mechanisms and functions for a humanoid robot that has the ability to communicate naturally with a human by expressing human-like emotion. We believe that in the future, it will be necessary for personal robots to interact bilaterally between human and robot. Therefore, the “Need Model” consisting of the “Appetite,” the “Need for Security” and the “Need for Exploration” was introduced to the mental model for humanoid robots. We also defined the “Need Matrix” and introduced the “Equations of Need” to describe the robot needs. Robots with the need model can generate and express active behavior according to their need. Finally, we implemented the new mental model to the Emotion Expression Humanoid Robot WE-4R (Waseda Eye No.4 Refined) developed in 2003.

1. はじめに

現在，産業用ロボットは工場において組み立てや搬送作業など，さまざまな用途で活躍しているが，あらかじめプログラムされた動作しかできず，新しい動作の定義には高い専門知識を必要とする．しかし，将来の普及が期待されているパーソナルロボットには，人間との共同作業や共同生活が求められており，産業用ロボットのような画一的な動作ではなく，接する人間によって動作を変化させたり，自発的に行動を起こしたりする必要があると考えられる．また，それらの行動を人間とのコミュニケーションの中で構築する必要があると考えている．そこで，筆者らは情動表出ヒューマノイドロボットを開発することにより，ロボットのパーソナリティを表現し，人間らしい情動表出を行うことで，人間と円滑なコミュニケーションを取るのに必要な機能の実現を目標としている．

ロボット分野におけるコミュニケーションロボットに関する研究としては，小林らが Ekman の 14 の Action Unit[1]を用いた顔ロボットを製作している．表情 19 自由度，眼球 2 自由度，首部 3 自由度の合計 24 自由度を有し，メインアクチュエータとして空気圧を用いており，人間と同速度で 6 基本表情の表出を実現できる[2, 3]．また，星野らは小型二足歩行エンターテインメントロボット SDR-4XII を用いたユーザとの対話行動を実現している[4, 5]．

これに対し，われわれは人間形頭部ロボット WE-3 (Waseda Eye No.3)シリーズにおいて，眼球と頭部の協調運動，眉・唇・顎・顔色を用いた表情表出，感覚器として，視覚・聴覚・触覚・嗅覚を実現している．さらに，2003 年，片腕 9 自由度の心理志向型ロボットアームが統合され，表情・体幹・腕を用いた情動表出が可能な情動表出ヒューマノイドロボット WE-4R (Waseda Eye No.4 Refined)を開発した．一方で，心理モデルとして，3 つの独立したパラメータを持つ心理空間，2 次の気分ベクトル，ロボットパーソナリティを導入し，2 次情動方程式による心理制御を行った．さらに，2003 年，欲求モデルに基づく行動生成を WE-4R に組み込んだ．本報告では，WE-4R に導入した欲求モデルについ

て以下に詳しく述べる．

2. 心理モデル

2.1 従来のヒューマノイドの心理モデル

われわれは，人間とロボットとの円滑なコミュニケーションの実現を目指した，ヒューマノイドの心理モデルを構築するにあたって，人間の心理モデルの定式化を行ってきた．そして，2002 年に開発した人間形頭部ロボット WE-4 では，心理学において相互に関係のある心理的要素を Fig. 1 のようなモデルで記述した．以下に，これまでの心理モデルを簡単に説明する．

まず，ロボットの心理空間として Fig. 2 に示すような快度・覚醒度・確信度の 3 軸からなる心理空間を定義しており，ロボットの心理状態は，心理空間内に定義された情動ベクトル **E** によって表される．

$$E = (E_P, E_A, E_C) \quad (1)$$

われわれは情動ベクトル **E** の動きを式(2)に示す運動方程式をモデルにした 2 次情動方程式によって定義した．

$$M\ddot{E} + \Gamma\dot{E} + KE = F_{EA} \quad (2)$$

M : *Emotional Inertia Matirx*

Γ : *Emotional Vis cos ity Matirx*

K : *Emotional Elasticity Matirx*

F_{EA} : *Emotional Appraisal*

ここで，情動的評価(Emotional Appraisal) $\mathbf{F_{EA}}$ とはロボットの内部・外部からの刺激によって引き起こされる心理状態への作用量で，これは感受個性(Sensing Personality) **Ps** と気分ベクトル(Mood Vector) $\mathbf{M_d}$ の関数として表される．

$$F_{EA} = f_{EA}(M_d, P_S) = k_m \cdot M_d + P_S \quad (3)$$

k_m : *Mood Influence Matrix*

感受個性とはロボットに入力された刺激が心理モデルにどの程度作用するかを決定する要素であり，感受個性は式(4)によって表される．

$$P_{SP} = f_{SPP}(S_t, I_t)$$
$$P_{SA} = f_{SPA}(S_t, I_t) \quad (4)$$
$$P_{SC} = f_{SPC}(S_t, I_t)$$

S_t : *External Stimuli*

I_t : *Internal Stimuli*

また，われわれは，刺激を条件刺激と非条件刺激に分類し，条件刺激に対して，式(5)に示すようなレスポンデント条件づけ学習の数理モデルとして知

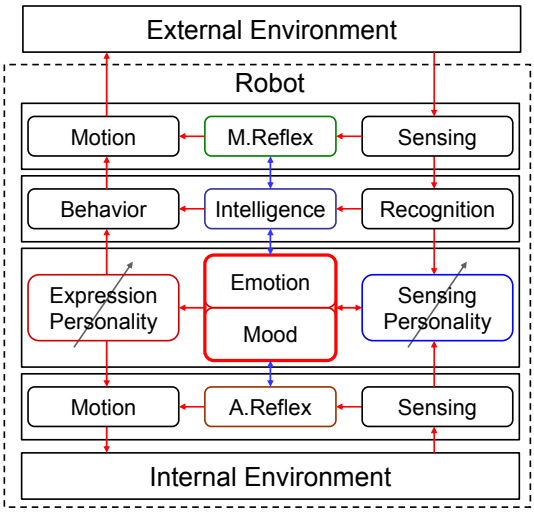


Fig. 1 Previous Mental Model

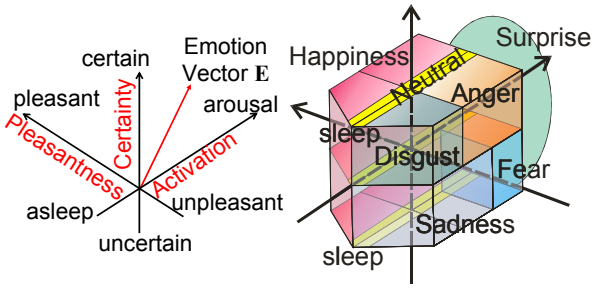


Fig. 2 3D Mental Space

られている Rescorla-Wagner Learning Rule[6]をモデルとした学習システムを適用することで非条件刺激に対する感受個性の学習を実現している．

$$\Delta P_{S\_CS} = \alpha(P_{S\_US} - P_{S\_CS}) \quad (5)$$

P_{S_US} : *SPT for Unconditioned Stimulus*

P_{S_CS} : *SPT for Conditionned Stimulus*

α : *Learning Rate*

さらに，ロボットの心理状態は刺激だけでなく，気分による影響を受ける．気分とは比較的長時間での弱い心理状態の変化である．われわれは，式(6)のように快度と覚醒度の 2 軸から構成される気分ベクトル $\mathrm{M_d}$ を導入した．

$$M_d = (M_{dP}, M_{dA}, 0) \quad (6)$$

気分ベクトルの快度は，外部からの刺激によって少しずつ変化するものとし，その変化は式(7)に示すような情動ベクトルの快-不快成分の積分としている．一方，気分ベクトルの覚醒度は，人間における睡眠・覚醒のような生体リズムに相当すると考えた．そこで，式(8)のように，Van del Pol 方程式を用いることで，ロボットの覚醒レベルに周期性を持たせ，体内時計を表現している．

$$M_{dP} = \int E_P dt \quad (7)$$

$$\ddot{M}_{dA} + \left(1 - M_{dA}^2\right)\dot{M}_{dA} + M_{dA} = 0 \qquad (8)$$

ロボットの感情は式(9)のような感情行列 E_m によって表される．心理空間には 7 つの感情領域が図 3 のようにマッピングされており，ロボットの感情 $\mathbf{E_m}$ は情動ベクトル $\mathbf{E}$ が各領域を通過することで決定される．

$$E_m = \begin{bmatrix} E_{m\_Neutral} \\ E_{m\_Surprise} \\ E_{m\_Happiness} \\ E_{m\_Sadness} \\ E_{m\_Anger} \\ E_{m\_Fear} \\ E_{m\_Disgust} \end{bmatrix} \qquad (9)$$

最後に，式(10)で表される表出個性によって，ロボットの感情を表情表出やロボットの行動に対し，どの程度出力するかが決定される．表出個性行列は 7×7 の正方行列で，単位行列が最も基本的な表出個性を表すことになる．最後に，ロボットは情動表出により，自身の心理状態を表出する．

$$\begin{aligned} &E_{mo} = P_E \cdot E_{mi} \\ &E_{mo} : Expressed\ Emotion \\ &E_{mi} : Current\ Robot\ Emotion \\ &P_E : Expression\ Personality\ Matrix \end{aligned} \qquad (10)$$

また，2 つのパーソナリティを変更することで，多様なパーソナリティの生成可能となっている．

しかし，従来のヒューマノイドの心理モデルではロボットの外部からの刺激，もしくは内部状態に起因する刺激が心理空間に作用することによって，感情に変化が生じ，その感情を表情によって表現していたため，一方向的なインタラクションであった．そこで，本研究では，双方向インタラクションを実現するために，欲求モデルを心理モデルに導入し，ロボットによる自発的な行動を生成可能とした．

2.2 人間の欲求

心理学において人間の欲求に関する研究は古くからなされており，いくつかの理論が提唱されている．そのような中，われわれは A. H. Maslow の欲求階層論[7]に着目した．欲求階層論とは，人間の欲求は生理的欲求・安全欲求・所属と愛情の欲求・尊重の欲求・自己実現の欲求の 5 層構造になっており，低階層の欲求が満たされると，すぐに 1 つ上の欲求が現れるという理論である．しかし，探索欲求のような一部の欲求に関しては 5 階層に分類できないと考えられている．そこで，われわれは欲求階層論において低次の階層である生理的欲求および安全欲求に加え，ロボットの好奇心である探索欲求の 3 つの欲求で構成される欲求モデルを構築した．

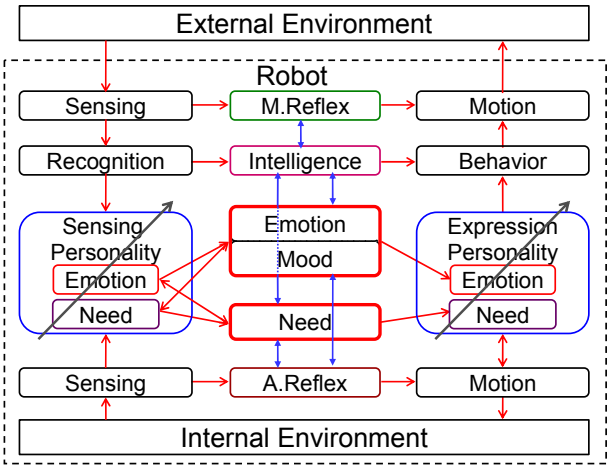


Fig. 3 Mental Model with Need

2.3 欲求モデルを統合した心理モデル

ロボットの心理モデルにおいて，欲求の位置づけとして，欲求は外部刺激や内部状態が作用するため，感情と非常に近いものと考えられる．そこで，われわれは欲求を表現するため，Fig. 3 のような心理モデルを構築した．Fig. 3 において，欲求と感情は 2 層構造となっており，欲求は感情よりも低い層に位置づけた．また，感受個性を介すことで感情と欲求は相互作用も可能とした．

2.4 欲求方程式

われわれはロボットの欲求状態を式(1)に示す欲求行列によって表した．本研究では，食欲・安全欲求・探索欲求の 3 要素によって構成したが，将来，欲求の要素数にしたがって，欲求行列を拡張することも可能である．そして，時刻 t における欲求状態を N_t，. t 後における欲求を $N_{t+.\,t}$ とし，. t 後のロボットの欲求状態を式(12)で表した．

$$N = \begin{bmatrix} N_a & N_s & N_e \end{bmatrix}^T \qquad (11)$$

$$N_{t+.\,t} = N_t + P_N \cdot .\ N \qquad (12)$$

$$\begin{aligned} &P_N : Need\ Personality\ Matrix \\ &.\ N : Small\ differences\ between\ two\ need\ states \end{aligned}$$

ここで，. **N** は式(13)のように入力された外的および内的刺激情報と感受個性よって決定される．また，P_N は式(14)のような 3 次の正方行列であり，欲求に対する個性を表し，対角項を変化させることで，多様な個性を表現できる．心理学では各欲求は独立であるため非対角項は 0 となるが，非対角項に 0 以外とすることで，心理学にない新しい欲求状態のシミュレートも可能であると考えている．

$$
\begin{aligned}
&. N_a = f_{NA}(I_t, S_t, E_t) \\
&. N_s = f_{NS}(I_t, S_t, E_t) \\
&. N_e = f_{NE}(I_t, S_t, E_t) \\
&I_t : Internal\ Stimuli, S_t : External\ Stimuli \\
&E_t : Emotion\ Vector
\end{aligned} \tag{13}
$$

$$
P_N = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & 1 \end{bmatrix} \tag{14}
$$

以上より，式(12)はさまざまな入力刺激およびロボット自身の内部状態によりロボットの欲求を決定する式であり，連続系における微分方程式，離散系における差分方程式に相当すると考え，われわれは式(2)を「欲求方程式」と名付けた．欲求方程式を導入することで，われわれは欲求モデルを数式化した．

2.5 食欲

食欲は人間の消費エネルギーに依存し，安静にしている状態でも消費する基礎代謝エネルギーと運動により消費するエネルギーの和として表される．つまり，ロボットの基礎代謝エネルギーを A_{BM}，運動による消費エネルギーを A_{EA}，単位時間あたりのロボットの消費エネルギー. A とすると，欲求方程式における. N_A は式(15)のように表すことができる．

$$
\begin{aligned}
&. N_A = f_{NA}(. A) \\
&. A = . A_{BM} + . A_{EA}
\end{aligned} \tag{15}
$$

また，基礎代謝エネルギーはロボットの心理状態によって変化し，ロボットの消費エネルギーはロボットに流れる総電流量など，内的もしくは外的刺激に依存すると考え，式(16)のように記述した．

$$
\begin{aligned}
&. A_{BM} = f_{ABM}(E_t) \\
&. A_{EA} = f_{AEA}(I_t, S_t) \\
&I_t : Internal\ Stimuli, S_t : External\ Stimuli \\
&E_t : Emotion\ Vector
\end{aligned} \tag{16}
$$

2.6 安全欲求

安全欲求は人間が持つ外界に対する防衛態度の一種である．近似の反応として生体防御反射があるが，生体防御反射が強い刺激に対する反射的な回避行動であるのに対して，安全欲求は反射行動よりも時間的に長い刺激に対する防御反応となり，弱い危険刺激であっても，連続的に入力されることで，その危険性を認識し，危険回避や防衛態度などの行動を引き起こす．われわれはロボットが外界から危険刺激を感じたときに，刺激が入力された部位と強度を記憶させることで安全欲求を実現した．

Table 1 Sensing Personality Table for Need

Stimulus	Sensation	P_{S_A}	P_{S_P}	P_{S_C}
Appetite	$N_A < T_A$	0	0	0
	$N_A > T_A, . N_A > 0$	-	+	+
	$N_A > T_A, . N_A > 0$	+	-	-
Need for Security	$N_S < T_S$	0	0	0
	$N_S > T_S, . N_S > 0$	-	+	+
	$N_S > T_S, . N_S > 0$	+	-	-
Need for Exploration		0	0	0

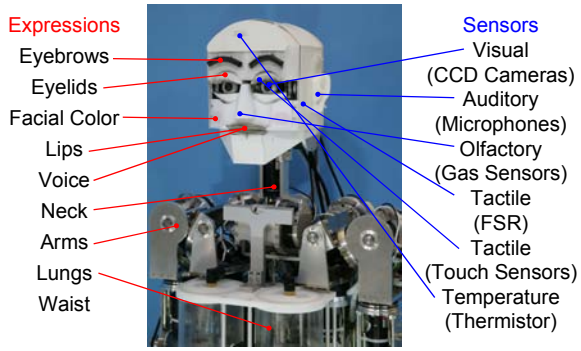


Fig. 4 WE-4R

2.7 探索欲求

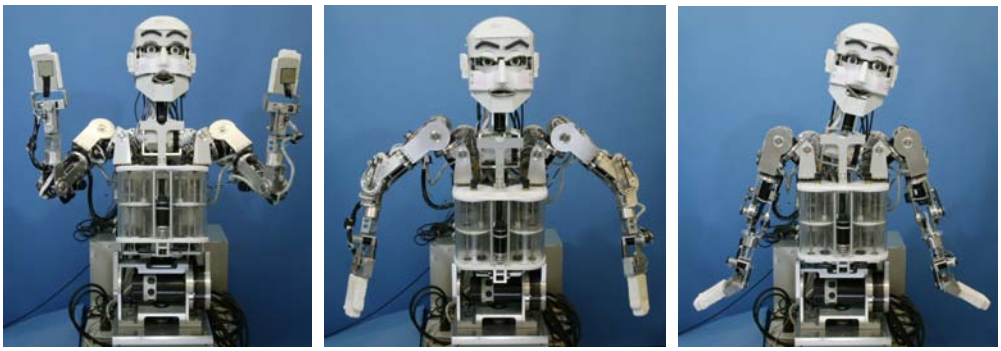
探索欲求とは人や動物が新しい場面や対象に出会うと，好奇心を示して探索行動を起こすという基本的欲求の 1 つであるが，探索欲求を欲求階層論で分類することはできないと言われている．

われわれは，ロボットに入力された視覚刺激とその対象物が持つ属性情報を関連づけて記憶させることで探索欲求を実現した．未知の視覚刺激は，新奇性が高いため，高い探索欲求が生じる．そのような刺激に対して，ロボットは対象物がどのような属性を有しているかを学習する．一方，既知の刺激は新奇性が低く，探索欲求を生じることはない．そのような刺激に対しては，対象物と関連づけて記憶している情報に従った情動の変化や行動を表出する．

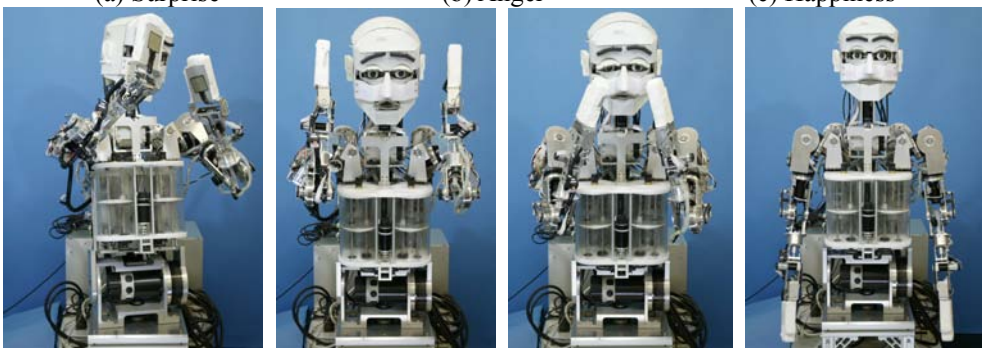
2.8 欲求による行動生成

ロボットの欲求が高まると，欲求を満たすためにロボットが自発的に行動を選択し，その行動を表出する．その結果，欲求が満たされない場合，欲求が解消されるまで，欲求を満たそうとする．例えば，食欲の場合，食欲が閾値以上に達すると，ロボットは食べ物を探索するようになり，食べ物を発見すると他の視標がロボットの近傍にあっても選択的に食べ物を注視したり，食べ物を獲得したりして食欲を満たそうとする．

また，欲求は行動の生成だけでなく心理状態への作用も引き起こす．われわれは刺激によるロボットの心理状態の変化を情動方程式によって記述しており，その作用量は SPT によって定義している．そ



(a) Surprise (b) Anger (c) Happiness



(d) Disgust (e) Fear (f) Sadness (g) Neutral

Fig. 5 Emotional Expression

こで，欲求もロボットの内的刺激の一種とみなし，欲求に対する SPT を Table 1 のように定義した．Table 1 は欲求が心理状態に対して正に働くか負に働くかを表しおり，正負やその程度を変化させることで，容易に多様な個性を定義可能である．

3. 情動表出ヒューマノイドロボット WE-4R

2003 年に開発された情動表出ヒューマノイドロボット WE-4R (Waseda Eye No.4 Refined)の写真を Fig. 4 に示す．WE-4R は情動表出だけでなく，自発的行動を表出するために，片腕 9 自由度の心理指向型のアームを有している．自由度構成は全 47 自由度となっており，Fig. 4 に示したセンサによって，視覚・聴覚・触覚・嗅覚の 4 感覚および身体情報の検出が可能となっている．そして，眉・眼瞼・口唇・顔色・声・首・腰・腕の動きによる情動表出が可能である．WE-4R による情動表出を Fig. 5 に示す．WE-4R に第 2 章で述べた心理モデルを組み込むことにより，外部刺激である視覚・聴覚・触覚・嗅覚刺激およびロボット自身が持つ内部刺激によって WE-4R の心理状態が変化し，WE-4R は自身の心理および欲求状態を行動や情動という形で表出する．

4. 評価実験

欲求が統合された心理モデルを実装した情動表出ヒューマノイドロボット WE-4R を用いて，3 種類の欲求に対する評価実験を行った．

4.1 食欲

ロボットの行動による食欲の変化と食欲による行動の変化を確認する実験を行った．実験は WE-4R に視標追従をさせながら食べ物を呈示し，WE-4R が食べ物を求める行動をした場合のみ，食べ物を与えることとし，このときの WE-4R の食欲と心理状態の変化を計測した．実験結果を Fig. 6 に示す．実験の結果，視標追従中，基礎代謝と運動による食欲の上昇し，ロボットは空腹による不快感を示した．また，食欲が上昇すると WE-4R は食べ物を獲得するため，手を伸ばして食べ物を求める行動を示した．WE-4R が食べ物を獲得すると，快状態を示し，食欲は急速に解消された．以上より，ロボットが食欲の変化に従い，空腹時に自発的に食べ物を獲得しようとする行動が確認できた．また，食欲による心理状態への作用が確認できた．

4.2 安全欲求

安全欲求によって WE-4R が危険と感じたときの行動を確かめる実験を行った．実験は WE-4R の左頬に「たたき」・「なで」の 2 種類の刺激を与えたときの WE-4R の行動の変化と心理状態の変化を計測した．実験結果を Fig. 7 に示す．実験の結果，危険刺激である「たたき」を 1 回感じただけでは WE-4R

の安全欲求は高まらないが，連続して感じることで安全欲求が高まることを確認した．逆に，「なで」のような安全な刺激は安全欲求を下げることを確認した．さらに，安全欲求が 10000 に達した地点で，WE-4R は左腕を頬の横に上げ，身を守る行動を示した．以上より，ロボットが安全欲求に従って，自発的に防衛行動を取れることを確認した．

4.3 探索欲求

最後にロボットに未知視標を呈示したときの，探索欲求による行動の変化を確認する実験を行った．実験では，WE-4R に未知視標を呈示したとき，および，同じ視標を再度呈示したときのロボットの行動と心理状態を計測した．実験結果を Fig. 8 に示す．その結果，WE-4R は未知視標を発見すると手を伸ばして獲得する行動が確認できた．また，未知視標とその属性の関係を学習したため，再度，同じ視標を呈示した場合，視標を獲得することはなく，過去の学習に基づいた心理反応を示した．以上より，探索欲求を導入することで，ロボットが未知視標に興味を持ち，その属性を自発的に調査する行動を実現でき，その結果を学習できることを確認した．

5. 結論と今後の展望

(1) ヒューマノイドロボットの心理モデルとして食欲・安全欲求・探索欲求から構成される欲求モデルを導入した
(2) 欲求方程式を導入することで欲求を数式化した
(3) 欲求の導入により，ヒューマノイドロボットによる自発的行動が実験により確認された

現在は欲求による行動の種類が限られているが，将来的には，行動の多様化と最適化を行うことで，より人間に近い行動を表出可能と考えている．

謝辞

本研究は早稲田大学ヒューマノイド研究所で行われた．本研究所のヒューマノイドコンソーシアムへの参加企業に対して感謝の意を表します．また，本研究の一部は岐阜県からの委託である WABOT-HOUSE プロジェクトにより行われた．ここに謝意を表します．さらに，研究に協力頂きました(株) 長田中央研究所，(株) NTT ドコモ，ソリッドワークス・ジャパン(株)，早稲田大学理工学総合研究センター，早稲田大学文学部木村裕教授に感謝の意を表します．

参考文献

[1] 工藤，P. Ekman，W. V. Friesen: “表情分析入門 -表情に隠された意味をさぐる-”，誠心書房，1987

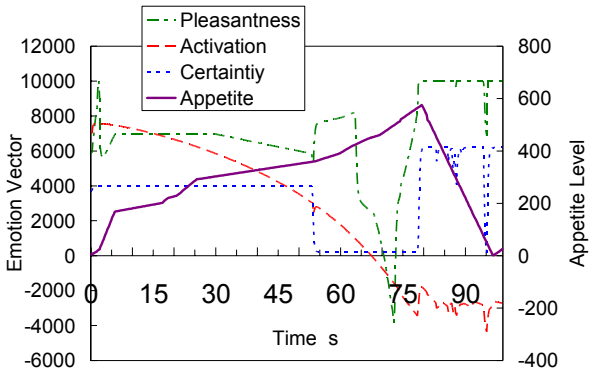


Fig. 6 Result of Appetite

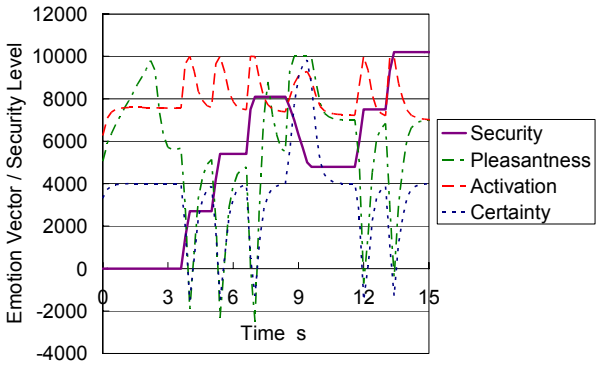


Fig. 7 Result of Need for Security

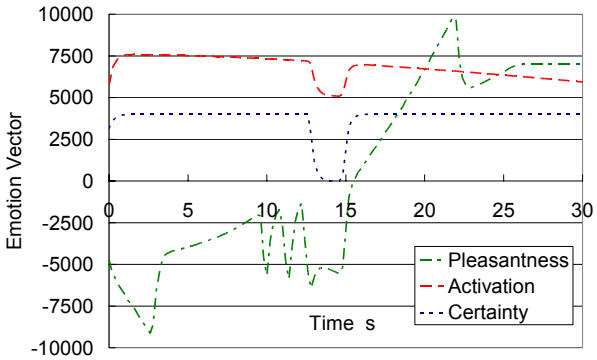


Fig. 8 Result of Need for Exploration

[2] 小林，原他: “アクティブ・ヒューマン・インタフェース(AHI)のための顔ロボットの研究(顔ロボットの機構と 6 基本表情の表出)”，日本ロボット学会誌学術論文，Vol.12，No.1，pp.155-163，1994

[3] 小林，原: “顔ロボットにおける 6 基本表情の動的実時間表出”，日本ロボット学会論文集，Vol. 14，No. 5，pp. 677-685，1996

[4] 星野，青山他: “パーソナルロボットにおけるユーザ固有情報を用いたインタラクション”，1E28，2003

[5] 星野，高木他: “パーソナルロボットにおける行動モジュールを用いた”，第 21 回日本ロボット学会学術講演会，2A18，2003

[6] 今田:“学習の心理学”，pp.107-1224，培風館，1996

[7] D. H. Maslow: “Motivation and personality”，Harper & Row，1970

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

人工知能学会研究資料
JSAI Technical Report
SIG-Challenge-0318-8 (11/13)

ユビキタスセンサ環境における音と画像の直接統合

A direct fusion method of video and audio in ubiquitous sensor environment

池田 徹志　　石黒 浩　　浅田 稔

Tetsushi IKEDA　Hiroshi ISHIGURO　Minoru ASADA

大阪大学大学院

Osaka University

ikeda@ed.ams.eng.osaka-u.ac.jp

Abstract

One of the required features of the ubiquitous sensor system is paying its attention to our signals, such as uttering keywords and footsteps. To detect and localize these signs, it is useful to fuse visual and audio information. The sensor fusion in previous works is performed in the task-level layer through individual representations of the sensors. This paper proposes another method that fuses sensory signals based on mutual information maximization in the signal-level layer. As an example, this paper shows an experimental result of a sound source localization by audio-visual fusion.

1 はじめに

環境中に多数のセンサを配置してネットワークで結合することにより，人間の日常活動を知的に支援する環境を構築する研究が盛んに行われている [3][7][8][9].

これらの知覚情報基盤 [7] における基本的な問題は，人間の行動をいかに認識して人間の行動や状態に応じた機能を提供するかにある．環境中には様々なセンサを設置することが可能であり，人間の行動をロバストに認識するためには，性質の異なる異種のセンサを組み合わせて用いることが有効であると考えられる．

従来の異種センサの統合行う研究では，各センサで認識処理を独立に行い，認識結果を統合するアプローチが主流であった．このアプローチの問題点は，各センサで信号からの特徴抽出を行った後に統合処理を行っているため，特徴抽出によって失われた情報は統合の時点で利用できないことである．

これに対し近年，信号の統計的性質に注目し，統合処理を早期の段階で行う手法が提案されてきている [1][4][5][6]. これらの手法の特徴は，信号間の相互情報量などの統計量を用いることで，性質が全く異なるセンサの統合を信号レベルで行うことができる点である．

しかしこれらの手法では，センサの信号間の統計的性質が定常であることが仮定されており，統合処理が固定的なことが本質的な問題である．そのため，知覚情報基盤の中で対象が移動し，対象を観測するセンサが切りかわって行くような場合に適用することは困難である．

本稿では，信号間の統計的性質を利用したセンサ統合のアプローチを拡張し，センサの信号間の統計的性質が変化する場合に適用する．環境中の対象の動きの軌跡を求め，統計量を求められた軌跡に沿って計算する手法を提案する．信号レベルの統合のアプローチの例として，カメラとマイクロホンの信号間の相互情報量を求め，画像上で音源の位置を同定を行った結果を示す.

2 相互情報量を用いた信号統合

2.1 信号レベルの統合とタスクレベルの統合

従来行われてきたセンサ統合のアプローチは，主に以下の二つに分けられる.

- 同種のセンサを複数用い，センサによる信号の違いに注目して直接的に統合する.
- 各センサで信号から特徴を抽出し，次に抽出した特徴を統合する.

前者のアプローチの例としては，アンテナアレイなどが挙げられる．信号の位相差を利用することにより，特定方向から来る信号に対し効果的に検出や抑制をすることができる．これは複数の信号を直接的に統合する手法と考えられ，ここでは**信号レベルの統合**と呼ぶ．このアプローチは同種のセンサの統合にしか適用されていない.

後者のアプローチの例としては，カメラや距離センサを用いてそれぞれ対象の位置を抽出し，抽出した結果を統合し位置を精度の良く求める手法が挙げられる．これはタスクの実行に必要な特徴を求めた後に統合する手法と考えられ，ここでは**タスクレベルの統合**と呼ぶ．このアプローチでは，抽象化した各センサの信号の性質が特徴抽出の段階で失われるという問題がある.

異種のセンサの統合を行う後者のアプローチを分析すると，まずセンサの信号の抽象化を行い，次に統合を行うという順の処理である (図 1 の実線)．これに対し異種センサの統合に置いても，前者の統合アプローチのようにより直接的な統合を先に行い，その後に抽象化をすると

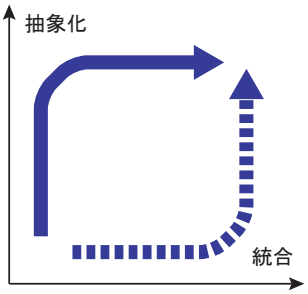


図 1: センサ統合のアプローチ

いうアプローチがあって良い (図 1 の点線)．これは**異種センサの信号レベルの統合**と言える．

近年，信号の統計的性質を利用することにより，異種センサの信号レベルの統合を行う手法が提案されている．Becker[1][2] は，画像を入力とする 2 つのニューラルネットワークを，出力間の相互情報量を最大化するという規範で学習することにより，入力画像間の視差に相当する特徴を抽出できることを示した．Hershey et al.[6] は，音声と画像中の各画素値の時系列との相互情報量を求めることにより，話者を画像上で特定できることを示した．Fisher et al.[5] は相互情報量最大化の規範で音声と画像に対する変換を学習し，話者の画像上での特定を行い，また指定された画像領域に存在する話者の音声を強調するフィルタリングを行った．

しかし提案されてきた異種センサの信号レベルの統合手法では，信号間の静的なモデルを仮定しており，動的な環境に適用することは難しい．

2.2 動的な環境下での信号統合

[5][6] で提案された手法では，音信号と画像 (画素) の関係を相互情報量を用いて評価し，高い相互情報量を示した画素を抽出することで音源の位置を同定している．この際に相互情報量を計算する間はセンサの信号間の統計的関係が変化しないことを前提としており，対象が移動する場合等の動的な環境に適用できない問題がある．

2 種類のアレイセンサ A,B を用いて同一の対象を観測した場合の例を図 2 に示す．ここでセンサ A は A1,A2 等の要素センサから構成され，センサ B も同様とする．対象が環境中で一定位置に存在する場合，センサ A1 とセンサ B1 の間の関係は定常である．

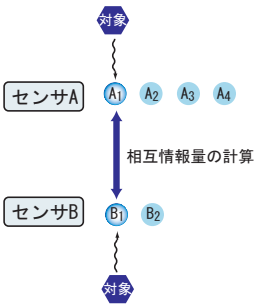


図 2: センサ間の関係が静的な場合

しかし環境中で対象が移動すると，一般に対象を最も良く観測できるセンサが時間と共に切りかわってゆく (図 3)．観測を開始した時点で，センサ A1 とセンサ B1 の信号の間に現れた関係は，対象の移動と共に，センサ A1 とセンサ B2 の間で観測されるようになる．したがって，長時間の観測によってセンサ A1 とセンサ B1 の間の統計的関係を求めても，対象が発する信号の間の関係をとらえることはできない．

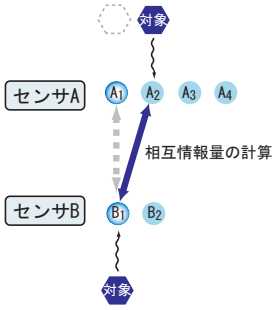


図 3: センサ間の関係が動的な場合

この問題に対処するため，各センサ上で対象を追跡し対象を最も良く観測するセンサの軌跡を求め，その軌跡に沿ってセンサを切換えながら相互情報量を計算することを提案する (図 4)．

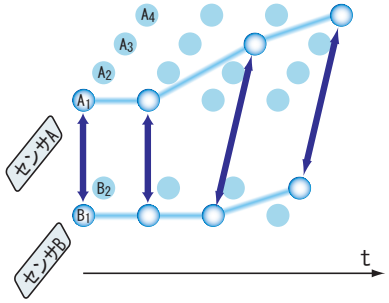


図 4: 求めたセンサの軌跡に沿った相互情報量の計算

2.3 信号間の時間同期

これまでの信号レベルの統合を行う研究では，異種センサの信号の間の相互情報量を抽出する際に，時間軸上での同時性の問題が扱われていなかった．

対象からセンサに信号が到達する時間や，信号が到達してからセンサが反応するのに要する時間は，センサの種類によって様々である．また情報基盤に設置されたセンサはネットワークで接続されており，厳密な時間同期をとるのが難しいという側面もある．さらに対象が複数のセンサを通じて観測された際に，最も良く対象の性質を抽出するためには，対象の示す動作に応じてあえて時間差をつけた方が良い場合も考えられる．

このような問題に対処するため，得られた信号を時間軸上で適切にシフトした後に統合することが必要と考えられる．

3　信号の直接統合による音源の定位

ここては提案する統合手法を，環境中に設置されたカメラとマイクロホンを用いて音源の位置同定を行うタスクを例として，具体的に述べる．

3.1　信号間の相互情報量の計算

音および画像中の画素の時系列信号をそれぞれ $A(t), V(t)$ とすると，$A(t)$ と $V(t)$ の間の相互情報量は以下のようにして求められる．

$$\mathrm{I}(A;V) = \mathrm{H}(A) + \mathrm{H}(V) - \mathrm{H}(A,V) \tag{1}$$

ここで $\mathrm{H}(A)$，$\mathrm{H}(V)$ はそれぞれ $A(t), V(t)$ のエントロピ，$\mathrm{H}(A,V)$ は両者の結合エントロピである．

$$\mathrm{H}(A) = -\sum_t \mathrm{p}(A(t)) \ \log\ \mathrm{p}(A(t)) \tag{2}$$

$$\mathrm{H}(V) = -\sum_t \mathrm{p}(V(t)) \ \log\ \mathrm{p}(V(t)) \tag{3}$$

$$\mathrm{H}(A,V) = -\sum_t \mathrm{p}(A(t), V(t)) \ \log\ \mathrm{p}(A(t), V(t)) \tag{4}$$

$A(t)$ と $V(t)$ が同時ガウス分布に従うと仮定することで，相互情報量は以下のように求めることができる．

$$\frac{1}{2}\log\frac{1}{1-\rho(A,V)^2} \tag{5}$$

ここで $\rho(A,V)$ は $A(t)$ と $V(t)$ の相関係数である．

3.2　対象の軌跡に沿った相互情報量の計算

異種センサを信号レベルで統合する従来の研究では，典型的には式 (2)-(4) に従って統合を行っていた．この手法を音源が移動する状況に適用した場合に，位置同定に失敗する様子を図 5 に示す．

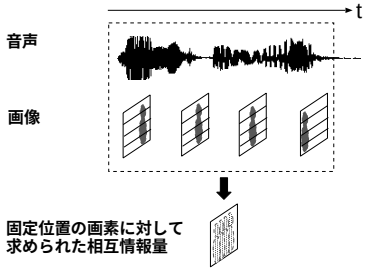


図 5: 従来手法による移動音源の位置同定

この問題に対処するため，音と画像を統合する前に，環境中の対象を背景差分によって検出し，対象の移動する軌跡を求める．$V(t)$ を求められた軌跡にそって移動させたときの相互情報量を計算することにより，センサ間の関係を安定して検出できる．提案手法の流れを 図 6 に示す．

音声，画像上の対象の位置をそれぞれ $x(t), y(t)$ とすると，一般に式 (2)-(4) は以下のように位置を含んだ式となる．ただし，本稿の実験では画像上での位置変化のみを扱っている．

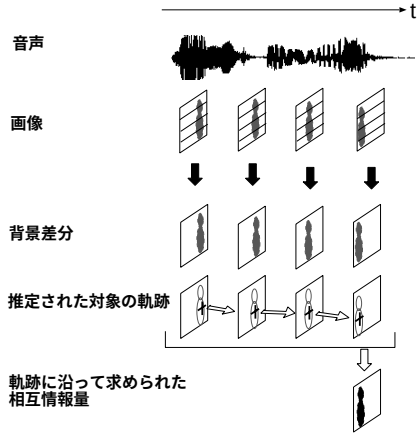


図 6: 提案手法による移動音源の位置同定

$$\mathrm{H}(A) = -\sum_t \mathrm{p}(A(t,x(t))) \ \log\ \mathrm{p}(A(t,x(t))) \tag{6}$$

$$\mathrm{H}(V) = -\sum_t \mathrm{p}(V(t,y(t))) \ \log\ \mathrm{p}(V(t,y(t))) \tag{7}$$

$$\mathrm{H}(A,V) = -\sum_t \mathrm{p}(A(t,x(t)), V(t,y(t))) \ \log\ \mathrm{p}(A(t,x(t)), V(t,y(t))) \tag{8}$$

3.3　音信号と画像信号の時間軸上での調整

対象を観測するセンサや対象の動きに応じて，音声信号を dt だけ遅延させたのちに画像信号との相互情報量を計算する．式 (6)-(8) は以下のように拡張される．ここで $A'(t,x(t)) = A(t-dt, x(t-dt))$ とした．

$$\mathrm{H}(A) = -\sum_t \mathrm{p}(A'(t,x(t))) \ \log\ \mathrm{p}(A'(t,x(t))) \tag{9}$$

$$\mathrm{H}(V) = -\sum_t \mathrm{p}(V(t,y(t))) \ \log\ \mathrm{p}(V(t,y(t))) \tag{10}$$

$$\mathrm{H}(A,V) = -\sum_t \mathrm{p}(A'(t,x(t)), V(t,y(t))) \ \log\ \mathrm{p}(A'(t,x(t)), V(t,y(t))) \tag{11}$$

4　実験

提案手法の有効性を検証するため，音信号と画像信号を統合し，画像上で移動する音源の位置を同定する実験を行った．部屋に２人がいる環境で，１人は足音を立てて移動し，もう１人は音を立てず手を振っている様子を，カメラ１台とマイク１本を用いて観測した．

画像は 30 フレーム毎秒で取得し，それぞれの人の領域を背景差分によって検出し，領域の重心の動きを推定した．音は 16kHz でサンプリングし，画像の１フレームに対応する 1/30 秒ごとの平均パワーを求めたものを用いた．音と画像の信号の例を図 7，図 8 に示す．

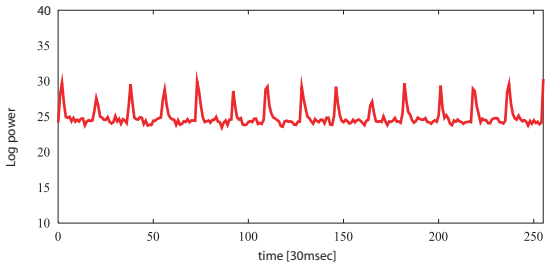


図 7: 音信号の例

(a) 画像例 (1)

(b) 画像例 (2)

(c) 背景画像

図 8: 画像信号の例

4.1 従来手法による音源位置同定

従来の信号レベルの統合手法 (式 (2)-(4)) に基づき，画像の各画素の値と音の平均パワーとの間の相互情報量を求めた結果を図 9 に示す．音源の移動に伴い，計算された相互情報量は画像上で広がり，この結果から歩行者の位置を同定をするのは困難である．

(a) frame 64

(b) frame 256

図 9: 従来手法による結果

4.2 提案手法による音源位置同定

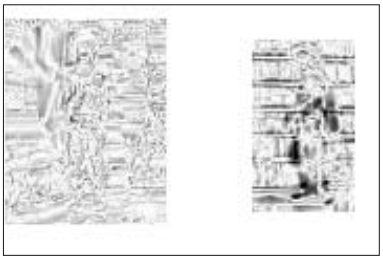
対象の位置を抽出し，求めた軌跡上を式 (6)-(8) に従って計算した結果を図 10 に示す．相互情報量の計算は抽出された領域の外接長方形に対してのみ行っている．手を振っている人 (画像内の左) の領域に対して，歩行者 (右) の領域が高い相互情報量を示している．この結果から，しきい値等を用いて音源の位置を検出することができると考えられる．

4.3 統合前に時間差をつけた場合の結果

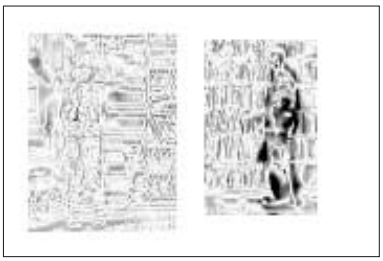
式 (9)-(11) に従って音声信号を dt だけ遅延させた時の結果を図 11 に示す．ここで dt を様々な値を設定したときの結果を別々に求めている．今回の実験では dt の値を $+2$ 付近にした場合に，音信号と画像信号の間の関係を最も良く抽出できていることが分かる．

4.4 信号波形の比較

図 12，図 13 に求められた軌跡上の画素値と，音信号の信号の様子を抽出したものを示す．図 12 は手を振る人の領域上の 1 点 (手付近) での画素値を，図 13 は歩行者の領域の 1 点 (足付近) での画素値の変化を示している．音信号は比較のため，図 7 と同じものを並べて示した．



(a) frame 64



(b) frame 256

図 10: 提案手法による結果

図 12 では信号間に関係は見られないのに対し，図 13 ではピークが同期して現れていることが分かる．

5 むすび

相互情報量に基づき，異種のセンサを信号レベルで統合する手法を提案した．提案手法では，画像上で検出された対象の重心を追跡することにより，センサで観測される対象が移動した場合にも対応できる．

信号統合の例として，画像上で音源の位置を同定する実験を示した．画像から複数の人が検出される場合でも，足音を手がかりにして歩行者の位置を同定することができることを示した．今後の課題として，音源が複数ある場合への拡張や，ノイズが存在する場合での評価を行ってゆきたい．接触センサ等の信号の性質が大きく異なるようなセンサとの統合にも興味を持っている．また本稿で提案した手法では，各センサ上で求めた軌跡上で相互情報量を計算したが，センサ間の相互情報量を最大化するようなセンサの軌跡を求めることを検討中している．

参考文献

[1] S. Becker and G. E. Hinton. A self-organizing neural network that discovers surfaces in r andom-dot stereograms. *Nature*, 355(9):161–163, 1992.

[2] S. Becker. Mutual information maximization: Models of cortical self-organization. *Network: Computation in Neural Systems*, 7(1), 1996.

[3] R. A. Brooks. Intelligent room project. In *Proc. of the Second International Cognitive Technology Conference*, 1997.

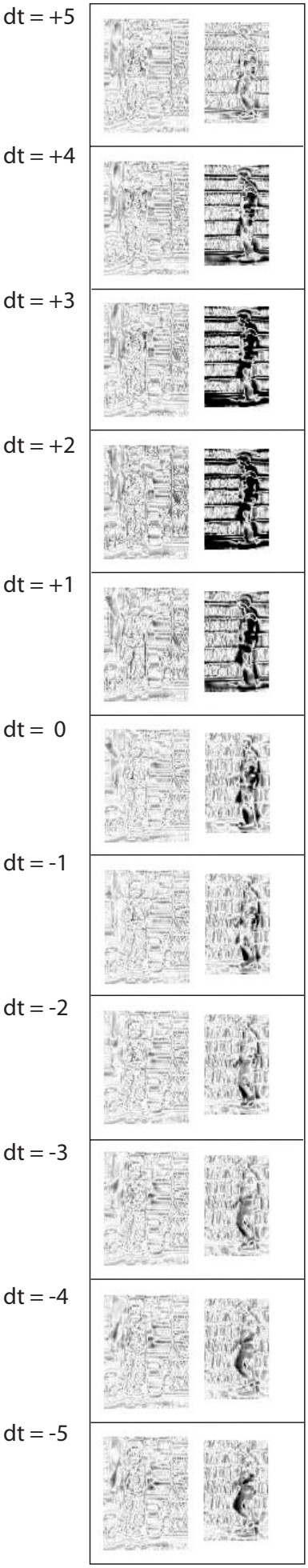


図 11: 統合前に時間差をつけた場合の結果

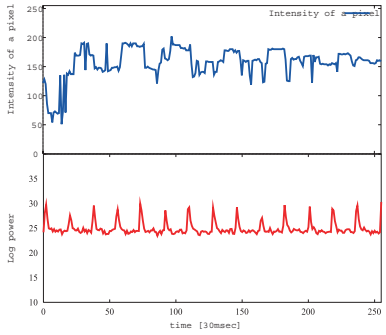


図 12: 画素値 (上) と音 (下) の信号例 (手を振る人)

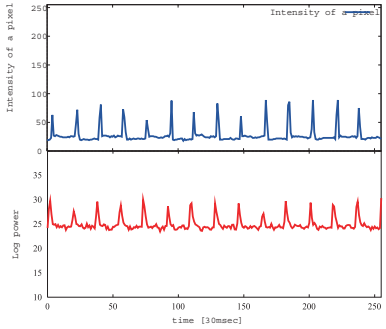


図 13: 画素値 (上) と音 (下) の信号例 (歩く人)

[4] R. Cutler and L. Davis. Look who's talking: Speaker detection using video and audio correlation. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2000.

[5] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems*, 2000.

[6] J. Hershey, H. Ishiguro, and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Proc. of Neural Information Processing Systems (NIPS'99)*, 1999.

[7] H. Ishiguro. Distributed vision system: A perceptual information infrastr ucture for robot navigation. In *Proc. Int. Joint Conf. Artificial Intelligence*, pp. 36–41, 1997.

[8] A. Pentland. Smart rooms. *Scientific American*, 274(4):68–76, 1996.

[9] 佐藤, 森, 原田. ロボティックルームの知能－ユービキタス知能－. 日本ロボット学会誌, 20(5):482–486, 2002.

頭部運動に追従するダミーヘッドシステム －テレヘッドⅡ －

Advanced version of a dummy head that tracks head movement: *TeleHead II*

平原達也、戸嶋巌樹、川野洋、青木茂明

Tatsuya HIRAHARA, Iwaki TOSHIMA, Hiroshi KAWANO, Shigeaki AOKI

日本電信電話株式会社　NTT コミュニケーション科学基礎研究所

NTT Communication Science Laboratories,　NTT Corporation

hirahara@idea.brl.ntt.co.jp, toshima@avg.brl.ntt.co.jp, kawano@avg.brl.ntt.co.jp, aoki@avg.brl.ntt.co.jp

Abstract

It is natural to take head and body movements into account in discussing the auditory sound localization function, because, when we hear sound, there is often some accompanying movement of the head and body. If our brain receives consistent auditory and motion-related information, we can localize virtual three-dimensional sounds fairly well. This paper describes the design concept, architecture and performances of the *TeleHead II,* an advance version of a steerable dummy head system that tracks three-dimensional human head movement in real time.

１．はじめに

私たちは五感を通じて外界の情報を脳に取り込み、脳は取り込んだ情報の断片から、その場の状況とつじつまが合うように、頑健かつ迅速に外界の様子を再構築する。そして脳は、その再構築された外界の様子に基づいて、次に取るべき行動を決める。この外界の様子を再構築するプロセスにおいて、外界の物理的な環境および自らの身体形状とその運動は重要な役割を果たす。なぜならば、私たちは自らの体を動かして外界とインタラクションを持つ「動物」だからである。

音響情報から外界の様子を再構築する聴覚においても、身体形状とその運動は重要である。聴覚系は、音の時間差と音圧差とスペクトル差を元にして、音の到来方向を計算している[1]。左右の耳が受ける音信号の差は両耳の間にある頭部と耳介の形状によって形成される。また、頭部の動きは音像定位精度を上げることが知られている[2-7]。さらに、頭部を自発的に動かせる場合には頭部伝達関数を精密に再現しなくとも音像定位精度が悪化しないことも分かっている[8,9]。人間と同等もしくは人間を超える聴覚機能を機械に与えたり、聴覚が本来持つ諸機能を有効に利用できる情報通信技術を創り出すためには、このような頭部形状と頭部運動を含めた聴覚の総合的な理解とそれらを考慮したシステムの設計が必要である。

我々は、ある場所の音環境を離れた場所で高臨場感に再現する方法、すなわち、あたかも自分がその場所に居るかのように音環境を耳元でリアルに再現する方法として、図１に示すような遠隔ロボットを用いるテレロボティクス方式を検討している。具体的には、受聴者の動きに追従するダミーヘッドを遠隔地に置き、ダミーヘッド耳に装着したマイクロフォンに到達した音をバイノーラル技術[10-12]を利用して使用者の耳元で再生するテレヘッド(TeleHead)である[13-16]。

このテレヘッドは、ダミーヘッドが置かれた音場における音響信号の「計算」を物理的な形状を持つダミーヘッドが実行するために計算コストが極めて低いという利点と、耳元で再生される音から脳が外界の様子を再構築する際に頭部運動情報と聴覚情報との整合性が取れるという利点とを合わせ持つ。信号処理を駆使して３次元音場を再現しようとするモダンなＶＲ方式では、３次元空間に配置される複数の音源と受聴者との相対関係の計算や、音源の移動や受聴者自身の運動に伴う音源と受聴者の相対的な位置関係の変化に対処するための複雑な計算が必要になる。

本稿では、先に報告したテレヘッドⅠ[14]を通じて明らかになった問題点を改善して構築したテレヘッドⅡの構造と性能、そして残された課題についてについて述べる。

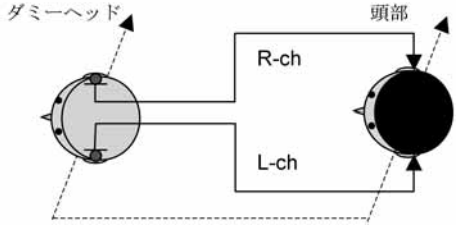


図１　テレロボティクス方式の概略

2. テレヘッド1号機 – *TeleHead Ⅰ* –

テレヘッドは、頭部姿勢検出部、ダミーヘッド部、ダミーヘッドを頭部運動に追従動作させる駆動部、ダミーヘッドで集音した音響信号を伝達する音響信号伝達部から構成される。図２に示すテレヘッドⅠの各部の構成は以下のとおりであった。

2.1　頭部姿勢検出部

受聴者の頭部の姿勢情報は頭頂に装着した 3 次元位置姿勢計測器（Polhmus, Fastrak）を用いて 120Hz で取得し、制御用 PC がサーボ系へ位置指令を出力した。

2.1　ダミーヘッド部

ダミーヘッドは、FRP 樹脂の骨格に発砲ポリウレタンを約 2cm 緩衝材として付着させ、その上に厚さ 1 mm の軟質ウレタンの表面形状を張り付けた 3 層構造であった。頭部の表面形状は、石膏を用いて取った型に軟質ウレタン樹脂を流し込んで再現した。また、耳介部分については別途石膏による型取りを行い、塩化ビニルゾル素材で成型し、頭部に装着した。ダミーヘッドの重量は約 1kg であった。

2.3 駆動部

ダミーヘッドは球面関節に取り付け、屈曲・伸展(Pitch)方向と側屈(Roll)方向はそれぞれ AC サーボモータ（400W）を用いてワイヤとプーリで駆動した。駆動に伴う機械的な雑音の発生を抑えるために、ギヤは使用しなかった。回旋(Yaw)方向については、ダミーヘッドと屈曲・伸展方向と側屈方向の駆動機構を含めた上部構造物全体を、静粛性に優れた DD サーボモータ（トルク 2.1Nm, 120rpm 時）で直接駆動した。

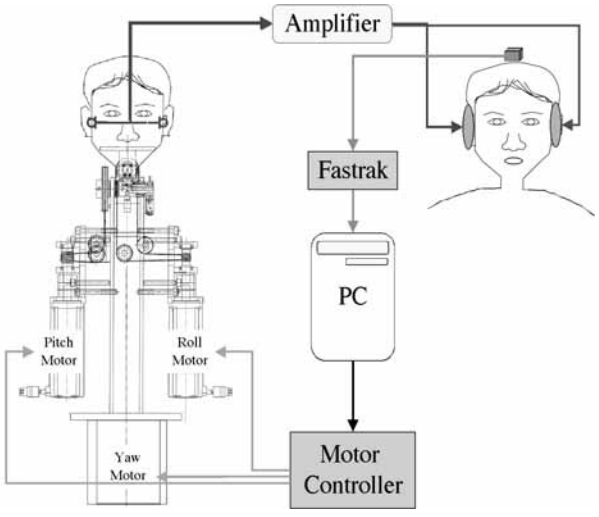


図２　テレヘッドⅠの構成

ダミーヘッドの可動域は，屈曲方向54°、伸展方向26°、側屈方向片側30°、回旋方向片側90°とした。また、運動速度の最大値は360°/sec　とした。なお、駆動機構部分は防音のため着脱式のウレタンフォームの胴体で覆った。

2.4 音響信号伝達部

ダミーヘッドの外耳道入り口から2mm奥にマイクロフォン(SONY, ECM77B) を設置した。その出力をオーディオアンプを経てヘッドフォンHDA200 (Sennheiser)[17]を介して受聴者の耳に導いた。ヘッドフォンと外耳道の伝達特性の補正は行わなかった。

2.5 テレヘッド1号機の問題点

このような構成のテレヘッド１号機の騒音特性、頭部運動の追従特性、ダミーヘッドの音響特性はいずれをとっても、満足できるものではなかった。

外部放射騒音は最大でも 40 dB SPL 程度で比較的低騒音であった。しかし、マイクロフォンからヘッドフォンにいたる音響信号ラインに混入する騒音レベルは 200Hz で 67dBSPL、1kHZ で 47dB と外部放射騒音よりも高いレベルであった。この原因は、モータやプーリの機械系の振動が、機構部分とダミーヘッドの骨格部分を伝わってマイクロフォンのダイヤフラムを振動させたためであった。騒音対策には機構部分の根本的な改善が必要であった。

頭部運動の追従動作には 200ms の遅延があり、高速で運動させた場合には 70%もの位置のオーバーシュートがあった。この原因は、人間の頭部運動において最も速くかつ頻繁に動かされる Yaw 方向の駆動に対して最も負荷の大きいモータ配置を取ったことと、Roll および Pitch 方向の駆動に低剛性のワイヤ駆動方式を採用したことにあった。これらを改善するためには、Yaw 方向駆動時の負荷が少ない機構と、追従性能と軌道の安定性を高める機構を検討する必要があった。

ダミーヘッドの形状の再現精度は 10mm 程度であったが、別途作成して取り付けた耳介部分の位置と角度がずれていた。そのため、ダミーヘッドと実頭の HRTF は必ずしも一致しなかった。この原因は、型取りした頭部と耳介を合わせる際の不手際と耳介の位置や角度の情報が不足していたためであった。また、HRTF 測定時の技術的な問題もいくつかあった。[18]

３．テレヘッド２号機 -*TeleHead Ⅱ*- の構成

テレヘッドⅠの問題点を基にして、以下のような構成で図３に示すテレヘッドⅡを構築した。[19,20]

3.1　頭部姿勢検出部

頭部検出部の構成はテレヘッドⅠと同じとした。

3.2　ダミーヘッド部

ダミーヘッドは新規に作成し直した。まず、着席したモデルに石膏を流して肩から上部の型を取り、その型に粘土を流し込んで造形物を作成した。複雑な形状をもつ耳介部分については別途型を取り造形物を作成した。次に、MRI および 光３次元計測装置（NEC, Danae-R）で計測した３次元の頭部形状データと各方向から写した写真を用いて、粘土の造形物を手作業で補正し最終的な形状を確定させた。耳介部分はこの補正作業の間に、頭部と合体させた。そして、この形状が確定した造形物を元にして、再び型を作成した。

今回は一つの型から 2 種類のダミーヘッドを作成した。一つは、FRP 樹脂とシリコンを別々に型に流し込んで骨格部分と表皮を作成し、FRP の骨格にシリコン表皮をかぶせた 2 層構造のダミーヘッド(DH1a)である。このダミーヘッドには頭髪を付けてあり、テレヘッドⅡに装着した。このダミーヘッドの重量は 2.2kg である。もう一つは、硬質発泡ポリウレタンを型に流し込んで作成したダミーヘッド（DH1b）である。このダミーヘッドはスキンヘッド状で、音響特性の評価用として用いた。図４に実頭と２種類のダミーヘッドの写真を示す。

テレヘッドⅠで問題となった、駆動機構部分とダミーヘッドの骨格部分を伝わる振動を低減させるために、マイクロフォンの取り付け部分はFRPの骨格を除去し、音響インピーダンスの異なる材料でマイクロフォンを支える構造にした。

RH1　　DH1a　　DH1b

図４　実頭（RH）と２種類のダミーヘッド(DH1a,b)

表１．実頭とダミーヘッドの主要寸法（mm）

	幅	高さ	奥行き	耳甲介腔	耳介幅
RH1	155	150	117	19	30
DH1a	151	152	111	21	29
DH1b	145	145	107	20	31
RH2	153	149	114	18	26

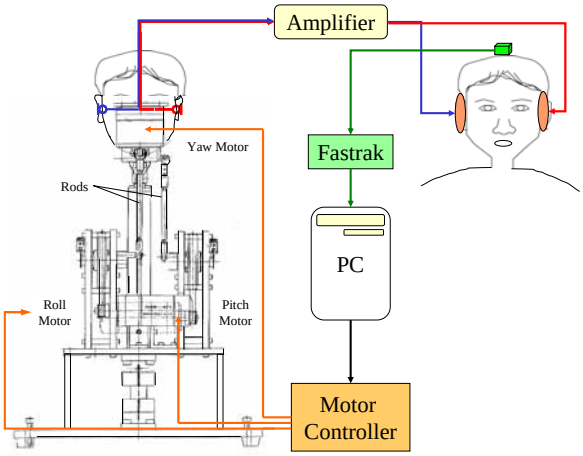


図３　テレヘッドⅡの構成

3.3 駆動部

駆動部はテレヘッドⅠと同じく３個のモータでYaw, Roll 及び Pitch 方向の 3 自由度を実現した。Yaw 方向はダミーヘッドの頭部に小型 DD モータ（連続定格トルク 0.47 Nm 60 rpm 時）を配置して直接ダミーヘッド部を駆動する方式を採用し、Yaw 方向駆動時の負荷の低減を図った。Roll 及び Pitch 方向は，ダミーヘッドの下部に AC サーボモータ(200 W)を配置し，剛性の高いベルトとロッドでダミーヘッド部へ動力を伝達する方式を採用した。なお、駆動部は、着脱式の FRP 樹脂の胴体で覆った。さらに，制御系の自由度を高めるために、位置制御に加えてトルク制御のインタフェースを追加した。頭部の可動域と運動速度はテレヘッドⅠと同じとした。

3.4 音響信号伝達部

音響信号伝達部はテレヘッドⅠと同じである。

4. テレヘッド２号機 -*TeleHeadⅡ*- の性能

4.1 ダミーヘッドの形状

実頭とダミーヘッドの形状の比較は、光３次元計測装置で測定したデータに基づいて行った。この計測装置の測定誤差は 1mm、これに加えて、複数の方向から測定したデータを貼りあわせて3次元再構成する際に生じる誤差が 1mm 強であり、合計 2mm 程度の誤差がある。また、光を反射しにくい頭髪部分はストッキングを被って頭髪を押さえた状態で計測を行った。DH1a については頭髪を付ける前に計測を行った。なお、複雑に入り組んだ形状をもつ耳介部分には計測しにくい部分があり、2mm 以上の誤差が生じる。

HRTFへの寄与率が高いといわれるいくつかの頭

部形状の寸法[21]を、モデルの実頭（RH1）と二つのダミーヘッド（DH1a、DH2b)、および別人の実頭（RH2）について比較したものを表１に示す。頭部の幅、高さ、奥行きがRH1に最も近いのはRH2、次いでDH1b、DH1aはRH1との差異が最も大きかった。頭部の直径で見ると、DH1aはRH1より約10mm小さかった。一方、ダミーヘッドの耳甲介腔（外耳道入り口周辺の窪み）と耳介の幅は、実頭との差異が少なかった。また、耳介部分も精度よく再現されていた。RH1とRH2では耳介部分の差異は4mmあったが、これは個人差である。

このように今回作成したダミーヘッドは全体的に実頭よりも小さめであったが、テレヘッドＩで用いたダミーヘッドよりは精度よく実頭の形状が再現できた。実頭とダミーヘッドの形状に差が生じた原因は、型取りの際に生じた誤差、造形物を修正するために使用した光３次元計測に含まれる誤差、ダミーヘッドを構成する材料が硬化する際の縮みなどであると考える。なお、別人の実頭RH2の形状がモデルの実頭RH1と類似していたのは単なる偶然である。

4.1 ダミーヘッドの頭部伝達関数（HRTF）

4.1.1 HRTFの計測方法

実頭とダミーヘッドの頭部伝達関数（HRTF）は、頭部の中心から音源までの距離を1.2mとして、無響室内で測定した。実頭の測定では、測定方位は全方位角と仰角-40～90° で合計143点を測定点とした。各測定点は、正中面と水平面は10° おきに、その他の点は隣り合う測定点との間の仰角と水平角が最大でも20° 以内に収まるように設定した。これは、HRTF の測定時間を90分以内に納めて、被験者の負担を減らすためである。ダミーヘッドは長時間の測定に耐えられるので、全方位角と仰角-40～90° で5° おきに1873点、ないし10° おきに469点を測定点とした。実頭の測定点はダミーヘッドの測定点のサブセットとなっている。

HRTF の測定には、RH1の左右の外耳道入り口付近をシリコン印象材で型取りした耳栓に装着した小型コンデンサマイクロフォン（Panasonic，WM62-AT102）を用いた。即ち、実頭もダミーヘッドも外耳道をマイクロフォン付きの耳栓で塞ぐ状態で HRTF を測定した[22]。もちろん、RH2のHRTF測定はRH2の耳で型取りした耳栓を用いた。

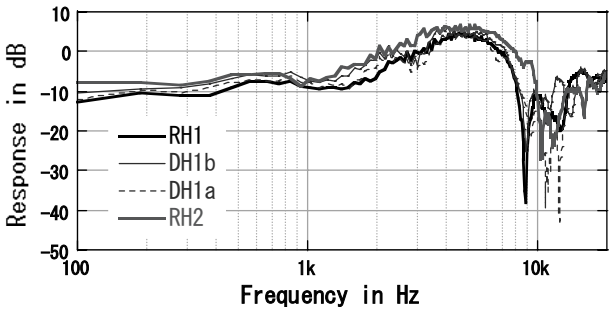


図５．実頭とダミーヘッドの HRTF の一例

HRTF はRH1について2回、DH1aについて4回、DH1bについて3回、RH2について1回測定した。なお、いずれの場合も、測定を開始する前に、レーザポインターを用いて両耳珠と鼻頭の位置が常に同じ位置にくるように位置あわせを行った。音源はサンプリング周波数 48 kHzの最適化引き延ばしパルス（TSP）信号[23]を使用し、10回の平均をとった。HRTF は512点のFFTで算出した。

4.1.2 結果

図５にRH1（太線)，DH1a（細線)，DH1b（点線)，RH2（灰色太線）の HRTF の一例を示す。この方位は RH1とDH1aおよびDH2bの HRTF の差が比較的大きい場合であるが、9 kHz付近のスペクトルの谷は一致しており，それ以下の周波数帯域におけるスペクトル概形も一致している。前述の通り，RH2の頭部形状はRH1と類似しているが、RH2のHRTFにはスペクトルの谷が10 kHz付近に現れており、RH1やDH1a,b とは明らかに異なっている。

ふたつの頭部で測定した HRTF の差 D は、式(1)によりスペクトル歪を算出して評価した。

$$D = \sum_d \left(\sqrt{\sum_\omega (H_i - H_j) / N_\omega} \right) \Big/ N_d \qquad (1)$$

ここで d は測定方位、ω は角周波数、N_ω は FFT のポイント数、N_d は測定方位の総数である。なお測定対象とする頭部 i の測定方位 d に対する HRTF を $H_i(\omega, d)$ を H_i と略記した。

実頭とダミーヘッドの間の HRTF の差を比較した結果を図６に示す。なお、HRTF の差の算出は 300Hz～10kHz の帯域で行った、帯域を制限した理由は、トラバースで移動させたスピーカの位置ずれによる 10 kHz 以上の HRTF の変動と、測定系の雑音による 300 Hz 以下の HRTF の変動の影響を排除するためである。計算対象の周波数帯域を 300Hz～20 kHz とすると全体に 2 dB 程度大きくなるが、各測定値間の関係は不変であった。図６には RH1, DH1a, DH1b それぞれについて複数回測定した HRTF の差、つまり HRTF の計測誤差も示してある。なお，全ての平均値間に有意差があるかどうか分散分析を行った結果、有意な差（$p<0.0001$）が認められた。

異なる２つの頭の間の HRTF のスペクトル差は、RH1とDH1b、RH1とDH1a、RH1とRH2の順で小さかった。これは、今回作成したダミーヘッドの HRTF が、他人の頭部よりも本人の頭部で測定した HRTF に近いことを示す。前述したように、形状はDH1aの方がDH1bよりも実頭に近いが、HRTF はその順序が逆転していた。これは、DH1の外耳道の直径がRHよ

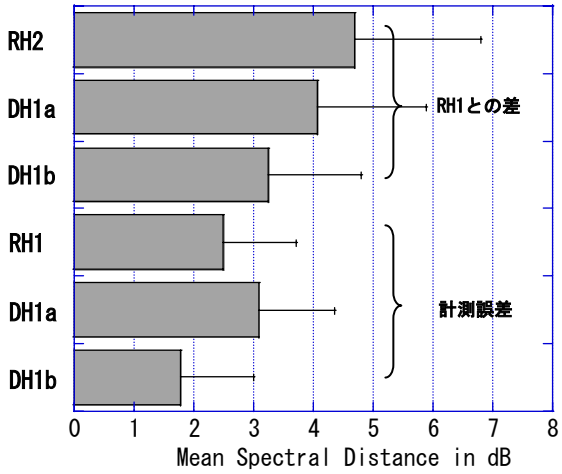


図６ 実頭（RH1）とダミーヘッド（DH1a,DH1b）と別人の実頭(RH2)で測定したHRTFの差の比較

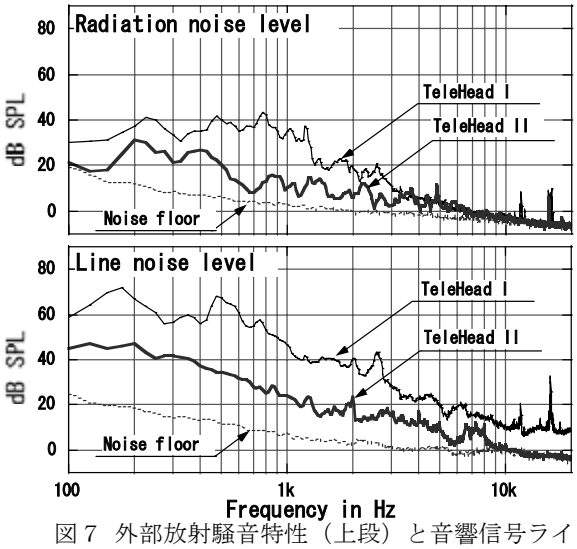


図７ 外部放射騒音特性（上段）と音響信号ラインへの混入騒音特性（下段）

りも狭かったために、計測の前後で耳栓マイクロフォンが5mm程外にせり出していた場合があったためである。つまり計測時のミスである。

１時間半に渡る計測中に姿勢を静止ことができない実頭RH1では、HRTFの測定誤差は2.5dBであった。一方、姿勢が変化しないダミーヘッドではこの測定誤差は小さくなり、DH1bでは1.8dB,DH1aでは3.1dBであった。DH1aの計測誤差がDH1b よりも大きかったのは、前述した耳栓マイクロフォンの不安定さの影響である。その後、DH1aに合った耳栓を作り直すことなどによって、DH1aのHRTFの測定誤差は1.0dB程度まで減少することを確かめている。

4.2 騒音特性

騒音特性の測定はテレヘッドⅡを無響室に入れ，制御用PCやその他の騒音源となる機器を無響室の外に出して行った．外部放射騒音はテレヘッドⅡの正面0.5mに設置したマイクロフォン（B&K, 4133）で測定した。ライン混入騒音は耳部に装着したマイクロフォンからヘッドフォンまでの音響系のゲインを1kHzの純音で校正した後に、ヘッドフォンの出力をIECカップラに装着した状態で測定した。いずれの場合も、測定中は常にヒトの頭部運動を追従させ運動の動作方向が偏らぬように注意した。

動作時の騒音特性を図７に示す。上段は外部放射騒音（実線）と無響室の暗騒音（点線），下段はライン混入騒音（実線）と音響信号ラインの暗騒音（点線）を示している。

テレヘッドⅠの騒音特性と比較すると、テレヘッドⅡの外部放射騒音とライン混入騒音は共に大きく減少した。特にライン混入騒音は、テレヘッドⅠにおいては聴感度の高い1～4 kHzの範囲で最大47 dB SPLであったが、テレヘッドⅡにおいては最大24 dB SPLであった。つまり、オフィス等の騒音環境下にテレヘッドⅡを設置した場合、駆動に伴う騒音はほとんど気にならない。

騒音が軽減した理由の一つは、FRP樹脂でできているダミーヘッド部の骨格とマイクロフォンとを音響的に縁を切る構造にしたことで，駆動によって生じる機械振動が音響信号ラインに混入しにくくなったためである。また、駆動部の位置制御のフィードバックゲインを下げたことで、微少な頭部姿勢の変動や３次元位置姿勢計測器に混入する電磁ノイズに起因する振動が少なくなったことも挙げられる。

4.3 頭部運動追従特性

制御用PCから正弦波状に変化する頭部運動の参照信号を各自由度毎に出力し、テレヘッドⅡの運動性能を測定した。正弦波状の運動が定常状態に至った周期(2～3周期目）における最大回転を与える時刻のずれを時間遅れ、参照信号と実測したテレヘッドⅡの運動振幅（回転角）の最大値の比率を到達率とした。正弦波の振幅はYaw方向は120°、Roll方向は40°、Pitch方向は60°とした。また正弦波の原点はYaw及びRoll方向では正面方向、Pitch方向では屈曲と伸展の可動範囲に差があることを考慮して水平面から10°屈曲（下向き）した方向とした。

追従性能の測定結果を図８に示す。上段が時間遅れ、下段が到達率である。横軸は動作速度として正弦波入力の周波数を示している。1.5Hzの動作では参照入力の70%程度の振幅で定常状態となり、その時の時間遅れは100～105msであった。0.5Hz以下の遅い動作では100%近い追従性能となり、時間

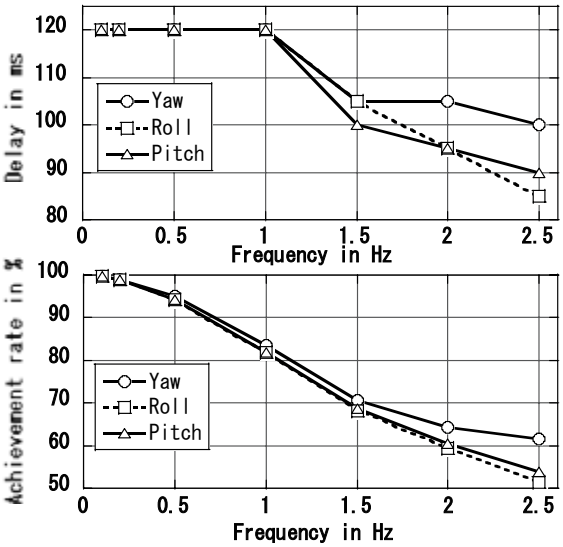


図８　テレヘッドⅡの正弦波入力に対する追従性能。
上段：時間遅れ，下段：到達率

遅れは120msに収束した。速い動作で時間遅れが短くなるのは、追従しきらない状態で定常状態に至るためである。なお、筆者らの頭部運動の速度を測定したところ、連続動作では概ね1.5 Hz以下であった。2Hz以上の速度は瞬間的には可能であるが、連続的には困難であった。

テレヘッドⅠでは0.5Hzの速度に相当する動作に対して200ms程度の遅延があったが，テレヘッドⅡの遅延は120ms程度まで改善した。また1.5HzのYaw方向動作時にテレヘッドⅠで認められた70%程度のオーバーシュートが、テレヘッドⅡにおいては30%程度のアンダーシュートに改善された。これらの結果は、Yaw方向駆動用モータへの負荷を軽減したことと、位置制御のゲインを下げて振動が残らないように調整したからである。

5. 今後の課題

5.1 アクチュエータ

テレヘッドⅠ・Ⅱでは、アクチュエータとして電磁式サーボモータを利用した。しかし、力の伝達機構やモータ本体が発生する機械振動や騒音の低減は頭の痛い問題である。また、1kgを超えるダミーヘッドを360°/sec（1.5Hz，Yow方向）にもおよぶ速度で回転させ、急に停止させる制御方法も頭の痛い問題である。ところが、ヒトは頚椎を構成する骨と筋肉によって、動作音を発生せずに、重い頭部を高速に運動させることができる。このメカニズムは巧妙であり、驚嘆に値する。

未来のアクチュエータの一つとして人工筋肉の研究が進んでいる。例えば、電歪ポリマーを用いた人工筋肉は11%の伸縮率、1.9MPaの駆動圧力、効率は80%以上、応答速度1ms以下という性能を出し、自然の筋肉を凌駕する面もある[24]。しかし、制御方法や耐久性、そして入手可能性などの観点から、すぐさまロボットのアクチュエータとして導入できる状況にはない。

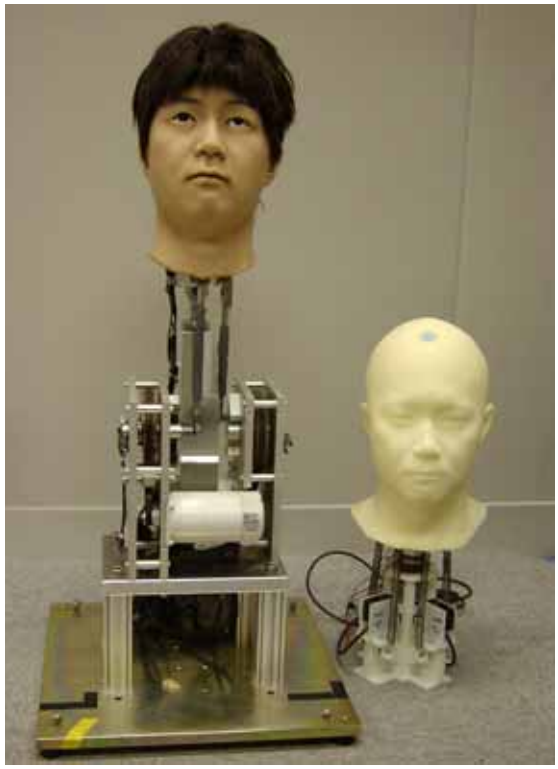


図９　電磁モータを用いたテレヘッドⅡ（左）と多自由度超音波モータを利用したテレヘッド（右）

我々は、高い静粛性と高トルク、そして多自由度の制御可能性を備えたアクチュエータとして、超音波モータに着目している。超音波モータは、形状の自由度が高くモータをリング状にも出来ることからカメラなどに用いられている。しかし、テレヘッドに用いるような多自由度超音波モータについては、その予圧機構、三自由度回転位置計測機構、負荷重量物が接続された状態での制御手法などが未確立である。詳細は別稿に譲るが、我々は上述した問題点を解決しつつあり、1kg の負荷重量物を載せた状態で、頭部運動と同様な三自由度運動の制御に成功している[25-28]。

超音波モータは、モータ自体が小型であり、電力供給なしに強いトルクを発生するのでブレーキが不要で、力の伝達機構も不要である。そのため、電磁式サーボモータを用いる場合と比べると、図９に示すようにテレヘッドの小型化が図れる。なお、使用した多自由度超音波モータは、東京工業大学上羽研究室で製作されたものである。

5.2 遅延時間と制御方法

テレヘッドⅡにおける120msの遅延は制御系サーボループの待ち時間に起因する。この遅延時間は総合的な運動の安定性とのトレードオフで現在の値になっており、さらに短くすることも可能である。また、現在は位置制御を行っているが、これをトルク制御に変更して制御方法を一新させることも検討中である。

機械系の運動の遅延がどこまで許容されるかについては、音源定位性能やリアリティの再現度といった知覚的な側面からの評価が必要であり、今後、この点について明らかにしていく必要がある。

5.3 頭部形状とHRTFと音源定位精度の関係

今回作成したダミーヘッドも、実頭の形状を100%の精度では再現できなかった。顔という柔らな皮膚で覆われた部分の３次元形状を正確に記述する難しさと、ダミーヘッドを構成する材料の選択の難しさがあると考えている。一方、自発的な頭部運動を許せばHRTFを鈍らせても音源定位精度は低下しないという知見[8]は、必ずしもダミーヘッドを精度よく再現しなくても構わない、ということを示唆している。すなわち、ツボさえ押さえればダミーヘッドの形状はある程度自由に設定できることになる。ダミーヘッド形状の変形がどこまで許容されるかについては、HRTFの再現精度とともに、音源定位性能やリアリティの再現度といった知覚的な側面からの評価も必要である。

5.4 ヘッドフォンの選択と外耳道特性の補正

ヘッドフォンを通じてバイノーラル信号を忠実に再生するためには、受音点すなわち外耳道入り口からヘッドフォンを見込んだ音響インピーダンスがヘッドフォンを装着していない場合とどれだけ近いかが重要といわれている[29,30]。今回利用したヘッドフォンHDA200 (Sennheiser)についても、その音響インピーダンスが測定されているが理想的な値ではない[31]。予備的に、他の開放型ヘッドフォン数種類と挿入型イヤフォンを試してみたが、定量的な評価は今後の課題である。

また、外耳道の容積は個人差があり、その音響特性を補正することによって音像定位精度が向上する[32]。このような音響的な伝達関数の補正がテレヘッドを利用する場合の音源定位性能やリアリティの再現度をどの程度向上させるかも検討する必要がある。

5.5 視覚、表情、発話機能

テレヘッドをテレイグジスタンス装置と見た場合、聴覚機能だけでなく視覚機能や発話機能、さらには表情表出機能や手腕運動機能を付加させることが必要と考える。五感全てとはいかずとも、これらの機能を具備させることによって、自分の分身を動かして遠隔地の環境をリアルに把握することができる、つまり、脳が遠隔地の様子を再構築できるようになると考えている。また、テレヘッドを置かれた側にとっても、表情変化がないダミーヘッドがリアルに首を振る姿はあまり気味が良いものではない。

5.6 ネットワーク接続

現在、受聴者が装着するヘッドフォンおよびヘッドトラッカーとテレヘッドとは電気的に直結されている。将来的には、その間に通信ネットワークを介在させ、何処にでもテレヘッドを持っていくことができるようにしたい。しかしながら、現在主流のIPネットワークでは、通信遅延とAV用エンコーダ・デコーダでの遅延が発生するために、本当のリアルタイム動作は期待できない。テレビ放送などの中継に利用されている専用回線と機器を利用すれば、これらの遅延はかなり短くできるが、通常のインターネットでは数百ミリ秒以上の遅延は不可避である。PHSや携帯電話などの通信回線でも事情は同様である。5.1で述べた機械系の運動の遅延と合わせたシステム全体としての遅延時間の許容範囲を明らかにするとともに、予測制御のような手法も導入する必要がある。

また、テレヘッドでは双方向のリアルタイム信号のやり取りが必要であり、このような双方向の信号の同期を確保するプロトコルも考案する必要がある。

6. おわりに

本稿では、性能を改善した頭部３次元運動に追従するダミーヘッドシステムーテレヘッドⅡーについて述べてきた。ダミーヘッドと駆動機構を改良することにより、騒音特性に関してはほぼ満足のいく性能が得られた。追従性能に関しては、運動軌道を安定化するとともに遅延時間を 40％短縮することができたが、まだ改良の余地が残されている。また、５章で論じたように、理想的なテレヘッドの完成に至るまでには様々な課題が残されている。今後は、これらの課題解決に向けた諸検討を進め、テレヘッドの設計指針を明確にしていく予定である。

参考文献

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

[2] Wallach H. (1940): "The role of head movements and vestibular and visual cues in sound localization,"

J.Exp.Psychol., **27**, 339-368
[3] Thrlow W.R. and Runge P.S. (1967) : “Effect of induced head movement in localization of direction of sound”, *J. Acoust. Soc. Am*., **42**, 480-488
[4] Thrlow W.R. and Runge P.S. (1967): “Head movements during sound localization”, *J. Acoust. Soc. Am*., **42**, 489-
[5] Perrett S. and Noble W. (1997): “The effect of head rotations on vertical plan sound localization” *J. Acoust. Soc. Am.* **102**, 2325-2332
[6] Perrett S. and Noble W. (1997): “The contribution of head motion cues to localization of low-pass noise”. *Perception & Psychophysics*, **59** (7), 1018-
[7] Wightman F. and Kistler D.J. (1999): “Resolution of front-back ambiguity in spatial hearing by listener and source movement,” *J. Acoust. Soc. Am.*, **105**, 2841-2853
[8] 植松尚, 柏野牧夫, 平原達也, (2001. 10)：“頭外音像定位における自発的な頭部回転の影響, ” 日本音響学会講演論文集, 501-502.
[9] Kato, M., Uematsu, H., Kashino, M. and Hirahara, T.(2003): "The effect of head motion on the accuracy of sound localization," *Acoustical Science and Technology* **24**, 315-317
[10] 三好正人 (1996)：″音場を創る″, 日本音響学会誌 **52**, 466-469
[11] Møller H. (1992): "Fundamentals of binaural technology," Applied Acoustics, **36**, 171-218.
[12] Wenzel E. M. (1992): "Localization in virtual acoustic displays," Presence, **1**, 80-107.
[13] 戸嶋巌樹, 植松尚, 平原達也 (2002. 10)：”頭部運動に追従するダミーヘッド”, 日本音響学会講演論文集, 467-468
[14] 平原達也, 戸嶋巌樹, 植松尚 (2002)：″頭部の３次元運動に追従するダミーヘッドシステム-テレヘッド(TeleHead)-″, 第16 回 AIチャレンジ研究会, 45-52
[15] Toshima, I., Uematsu, H. and Hirahara, T. (2003):"A steerable dummy head that tracks three-dimensional head movement: TeleHead," *Acoustical Science and Technology* **24**, 327-329
[16]戸嶋巌樹, 青木茂明, 平原達也 (2003)：″頭部形状と運動を考慮した高臨場感伝達ロボット：テレヘッド″, 日本ロボット学会学術講演会, 1I2a
[17]平原達也(1997)：”聴覚実験に用いられるヘッドホンの物理特性”, 日本音響学会誌, **53** (10), 798-806
[18] 植松尚, 平原達也 (2002. 04)：”頭部形状を精密に模擬したダミーヘッドの頭部伝達関数”, 日本音響学会講演論文集, 467-468
[19] 戸嶋巌樹, 青木茂明, 平原達也 (2002. 09)：″頭部運動に追従するダミーヘッドの性能改善″, 日本音響学会秋季研究発表会, 463-464
[20] 戸嶋巌樹, 青木茂明, 平原達也 (2002. 09)：″実頭と複製ダミーヘッドの頭部伝達関数″, 日本音響学会秋季研究発表会, 465-466
[21] 西野隆典、梶田将司、武田一哉、板倉文忠(2001)：″重回帰分析に基づく頭部伝達関数の推定, ″ 電子情報通信学会論文誌 **J84-A**, 260-268
[22]飯田一博, 岩根雅美, 矢入幹記, 森本政之 (1997. 03), ″正中面定位における耳介各部位の役割″, 日本音響学会春季講演会論文集, 439-440
[23] 鈴木陽一、浅野太、曽根敏夫 (1989)：音響系の伝達関数の模擬を巡って（その２）, 日本音響学会誌, **45**, 44-50 [31] 岩谷幸雄、渋谷亮輔、鈴木陽一 (2003. 9)：″ヘッドホンの自由空間等価特性(FEC)の個人差″, 日本音響学会講演論文集, 519-520
[24] Pelrine, R., Kornbluh, R., Joseph, J. (1998): "Electrostriction of Polymer Dielectrics with Compliant Electrodes as a Mean of Actuation", *Sensor and Actuators A: Physical* **64**, 77-85
[25] 川野洋, 平原達也 (2003. 09)：″予圧条件下における多自由度超音波モータの回転位置制御手法″, 日本音響学会秋季講演会講演論文集, 1034-1044
[26] 川野洋, 平原達也 (2003)：″多自由度超音波モータの高臨場感伝達ロボット-テレヘッド-への適用手法 -予圧機構と三自由度回転位置計測手法-″, 第 21 回日本ロボット学会学術講演会, 3D22,
[27] 川野洋, 平原達也(2003)：″多自由度超音波モータの回転位置制御 -予圧条件下における適応制御手法-″, 第 21 回日本ロボット学会学術講演会, 3D23
[28] Kawano, H. & Hirahara, T. (2003): "Three-DOF Angular Positioning Control using a Multi-DOF Ultrasonic Motor in the Pre-loaded Condition - Application to the Auditory Tele-Existence Robot ▪ lTeleHead▪ l", Proc. of 2003 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems.
[29] Møller H., Hammershøi D., Jensen C.B. and Sørensen M. F. (1995): "Transfer characteristics of headphones measured on human ears," J. Audio Eng. Soc., **43** (4), 203-217.
[30] Møller H., Hammershøi D., Jensen C.B. and Sørensen M.F. (1995), "Design criteria for headphones," J. Audio Eng. Soc., **43** (4), 218-232.
[31] 岩谷幸雄、渋谷亮輔、鈴木陽一 (2003. 09)：″ヘッドホンの自由空間等価特性(FEC)の個人差″, 日本音響学会秋季研究発表会, 519-520
[32] 小澤賢司、金澤永治、鈴木陽一(2000)：″ヘッドホンを用いたバイノーラル再生における個人性補正の効果″, 日本バーチャルリアリティ学会論文誌 **5**, 949-956

〒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://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/