特集 5

ロボット聴覚の課題と現状

奥乃　博（H.G.Okuno）
京都大学大学院情報学研究科知能情報学専攻
okuno@i.kyoto-u.ac.jp

中臺　一博
（株）ホンダ・リサーチ・インスティチュート・ジャパン
nakadai@jp.honda-ri.com

ロボットが家庭に入ってくるようになり，ロボットと人とのコミュニケーション，特に，ロボットに装備されたマイクロフォンを用いたコミュニケーションや音による環境知覚がますます重要になってきている．最近，ロボット自身の耳による聴覚機能がようやく活発になってきた．では，ロボットのための聴覚機能にはどのようなものが必要であろうか．

聴覚は人間にとって最も重要な感覚である．言語によるコミュニケーションが聴覚によって成立することは容易に理解されるが，「ヒトは聴覚によってのみ言語を獲得し，そこに文化が生まれ，継承される．書かれた言語は目によって伝承されるが，話す言葉は耳からしか得られない．話し言葉があって書く言葉が生まれる」ことを，多くの人が理解していないのは残念なことである．

鈴木淳一，小林武夫共著「耳科学－難聴に挑む」（中公新書 1598, 2001）

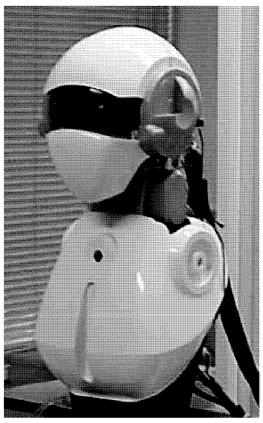
ロボットが身近に，でも音は苦手

21 年紀に入り，ヒューマノイドロボットが数多く開発されるだけでなく，犬，猫，熊などの形をしたペットロボットが家庭に入ってきている．これらのロボットは，2 本以上のマイクロフォンを搭載し，「耳」に相当する機能を備えている．しかし，視覚機能が，ボールを認識し，追跡できるようなレベルにあるのに対して，聴覚機能，特に，音声によるインタラクションはパソコンで提供されている音声認識システムと比較しても，相当機能が低いと言わざるを得ない．

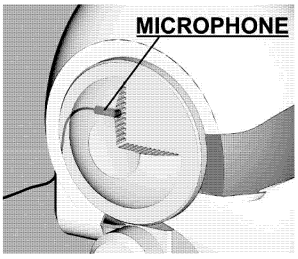
本稿では，ロボット音響についての研究課題について検討し，我々が現在行っている研究，特に，本年 5 月 NHK が放映した「鉄腕アトムをつくる」というドキュメンタリーで紹介された，三話者が同時に発話した単語を聞き分けるロボット SIG の機能についても簡単に報告する．なお，音声対話のような高度なレベルでの処理は，ヒューマンマシンインタフェースの高度化に不可欠ではあるが，本稿では触れない．また，我々は人工知能学会 AI チャレンジ研究会でロボット聴覚を特集した研究会をこれまでに 3 回開催してきた．本稿の執筆に合わせて，その報告集をオンライン化したので，個別の報告は下記の URL を参照していただきたい：
http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/

ロボット聴覚に求められる要求条件

ロボットが動き回ったり，人間社会の中で共生するためには，周囲の状況を知覚し，認知し，さらに，理解し

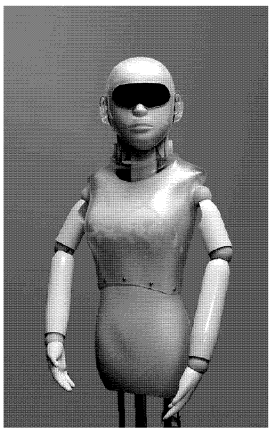


(a) SIG の外装

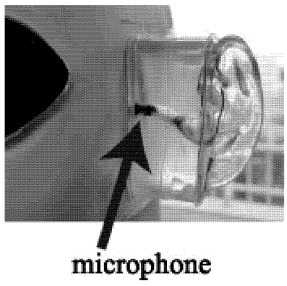


(b) マイクロフォンの実装部

図-1　ロボットの耳の例 1



(a) SIG2 の外装



(b) マイクロフォンの実装部

図-2　ロボットの耳の例 2

て行動することが求められる．このような動的な環境でのロバストな認知は，不良設定問題であり，重要な課題として，長年研究されている．

入力から出力まで，センサとしてのロボット聴覚に求められる要求条件を以下にまとめる．

- ロボットは自分に装着されたマイクロフォン（耳）で音を聞く必要がある．
- 入力音は複数の音源から生ずる混合音であり，音源は音声だけでなく，楽音，環境音などが含まれ，さらに，音源が動いても，逆に，ロボットが動いても聞き分けることができる必要がある．
- 音源分離で聞き分けられた音響信号は，後続の処理のために，音声認識，楽音認識，擬音語認識等により記号表現にグラウンディングする必要がある（シンボルグラウンディング）．

以下，各項目について説明をし，そのあと具体的なシステムとして，我々が科学技術振興事業団北野共生システムプロジェクトで開発してきたロボットSIG上の視聴覚を統合した実時間複数話者追跡システムと複数話者同時音声認識システムを解説する．

ロボットの耳とは

マイクロフォンをロボットの体に設置する場合，ロボット自身の出す音をできるだけ取り込まないようにするための工夫が必要である．たとえば，マイクロフォンの特性，特に，指向性か無指向性か，さらに，指向性の場合どの程度の指向性なのか，マイクロフォンに体を通じた伝導で音が伝わらないための工夫は，マイクロフォンを収納する耳朶は必要か，必要ならどのような形か．さらに，ロボットの外装のデザインからの影響も考慮する必要がある．

我々が使用しているロボットSIGとSIG2の外装と耳の形を**図-1**, **図-2** に示す．SIGでは，外装の耳の部分にあるマイクロフォン用のハウジングは薄い緩衝材に包まれているものの，外装からの伝導が少なからずある．SIG2では，医学教育用のシリコン製外耳道モデルを使用し，その鼓膜部分にマイクロフォンを装着している．いずれも，無指向性のマイクロフォンを左右1つずつ使用している．他のロボットのマイクロフォンの数は，ホンダASIMOが2本，ソニーSDR-4Xが7本，産総研のJiJo2が8本，NECのPaperoが3本である．

ロボットとの音声対話の研究では，音声認識精度向上のために，マイクロフォンを話者の口元にできるだけ近づける，というアプローチがとられてきた．MITのKismet[1]は，耳朶を有した1対のマイクロフォンを装備しているが，実際には話者の口元近くにつけたピンマイクロフォンを使用して対話を実現している．早稲田大学のHaradyや奈良先端大学のAskaは，従来はヘッドセットマイクロフォンを使用していたが，最近はロボット聴覚という観点から，ロボット自身の耳で取得した音声による人間との対話研究を進めている．

なお，ヘッドセットマイクロフォンであっても音声認識は難しい．たとえば，ロボット展示会場では認識率が雑音により大幅に低下するし，複数人が参加する対話では別話者の声が自分の体に反射してマイクロフォンに入り，あたかもその話者の発話として間違ってとらえてしまう．また，音声会議録で使用されるマイクロフォンアレイは，さまざまな形状でも，動く物体でも装着できるとされているが，現時点で著者が知り得る限り，音声強調ではなく，複数混合音を聞き分けるようなシステムは

図-3　無響室におけるSIGのHRTFの測定

報告されていない．

混合音を聞き分けるには

音環境理解（Computational Auditory Scene Analysis, CASA）は，工学的なアプローチにより，特定の音に特化しない一般の音を理解し，その認知機構を解明する枠組みを提供する研究分野である[2]．具体的には，さまざまな音響環境で，単一音源ではなく混合音を対象として，音声だけではなく，音楽など非音声を含めた音響信号を理解するための枠組みを扱っている[3]．特に，混合音から，同一音源に由来するなど一貫した属性を備えた音響ストリームを抽出・分離する音源分離については，音環境理解における主要な研究テーマの１つとして，これまでさまざまな手法が試みられてきた．

混合音の分離には，マイクロフォンアレイが一般的に使用される．マイクロフォンアレイで得られたマルチチャネルのデータから特定方向の音だけを抽出するには，ビームフォーミング，独立成分解析（ICA, Independent Component Analysis）といった技法がよく使われる．ビームフォーミングによる音源分離でよく使用されるのは，遅延加算型（delayed sum）である．

適応型マイクロフォンアレイの一種である，ナルフォーミングでは，N+1本のマイクロフォンを使えば，N個の音響的な死角を作成できる．独立成分解析では，情報論的に互いに独立なN個の音源を，N本のマイクロフォンで理論的に分離することができる．最近の研究では，音源数の制限や動的な音源数変化にも対応できる手法が開発されている．しかし，一般環境では，マイクロフォンの位置が変化したり，音源が移動したり，人が入ってきて音響環境に変化が生じるので，必ずしも理論だけでは解決できないことが少なくない．

2本のマイクロフォンによる音源定位

一般に，調波構造を持つ音は，基本周波数とその整数倍の倍音から構成されているので，分離のための多くの手がかり情報が得られる．しかし，混合音では，情報が失われたり歪んだりするので，音源分離で曖昧性が大きくなり，他の特徴の活用が必要である．そのような手がかりとして音源方向を考えよう．

人は２つの耳で，複数の音を聞き分けている．そのアナロジーとして，２本のマイクロフォンによる音源定位・分離を考えよう[4]．Jeffress の cross-correlationモデルは，２本のマイクロフォンから得られた入力音を遅延させながら相関をとり，最大の相関を与える遅延から両耳間時間差（Interaural Time Difference, ITD）を求め，定位を行うものである．

音源の定位には，まず，視覚処理におけるステレオマッチングと同様に左右のチャネルごとに抽出された倍音構造のペアを見つけ出す．次に，抽出された倍音構造のペアから，両耳間位相差（Interaural Phase Difference, IPD），および両耳間強度差（Interaural Intensity Difference, IID）を計算し，それらの値から方向情報を求める．IPDやIIDと方向情報との対応関係は，あらかじめ無響室で測定した頭部伝達関数（HRTF）で与えられる．HRTFとは，頭部形状によってもたらされる音響特性の変化を表す関数であり，方向ごとに測定する．通常は，無響室に設置したダミーヘッド等に対するインパルス応答測定によって得られる（**図-3**）．

一般に無響室で測定されたHRTFを実際の環境で使用するには，音源からマイクロフォンまでの伝達特性を畳み込むことが必要になる．そのため，反響も含めて使用する環境で部屋の伝達特性まで含めてHRTFを測定することが行われる．HRTFの測定は多大な労力を要するだけでなく，部屋の家具の配置が変わったり，ロボットの位置や向きが変わると，音響環境が変わるので，再測定が必要となる．したがって，HRTFを使用しない代替手法が実環境では不可欠である．

アクティブパーセプション

２本のマイクロフォンでの音源定位では，正面からの

音と背後からの音を区別することは難しい．Huangらは，ロボットに搭載した3つのマイクにより，前後問題の曖昧性の解消を行っている．人の場合には，音源方向に頭を向けると音源がどちらに移動するかを見たり，あるいは音源を発見することにより，そのような曖昧性を解消している．つまり，挙動を組み合わせたアクティブパーセプションや画像情報との情報統合が曖昧性解消の重要な鍵となっている．

視覚でのアクティブパーセプションであるアクティブビジョンは，焦点距離，ズームイン・アウト，解像度，虹彩といったカメラパラメータをアクティブに制御し，視覚情報を効率よく正確に取得することを目的としている[5]．つまり，アクティブビジョンは，視覚と動作を統合することによって，よりよい視覚的情景分析（Visual Scene Analysis）を行う枠組みとなっている．

聴覚の分野でもアクティブパーセプションは重要な概念である（アクティブオーディションと呼ぶ）．たとえば，一般に人は，複数の音源からの音声や雑音が混在している環境下であっても，これらを分離し，特定の音源からの音に注意を傾け続けることができる．これは，カクテルパーティ効果として知られている．人間の聴覚は常にアクティブな動作を伴い，これによって聴覚的情景分析（Auditory Scene Analysis）を向上させている．

しかし，アクティブパーセプションは聴覚にとっては両刃の剣である．実環境で動き回るロボット，あるいは，行動と知覚が結び付いたアクティブパーセプションでは，マイクロフォン自体が動き，それに伴って，モータ雑音や機械音が発生する．このような音は，たとえ小さくてもマイクロフォンに近いので，相対的に大きな雑音となり，外部からの音の信号雑音比が低下する．つまり，よく聞こうとして動いたところ，自分の出す音が災いして，かえって聞こえにくくなるという皮肉な結果にもなりかねない．

ロボットやシステムが発生する内部雑音を軽減する最も簡便な方法は，動作を中断してから，聞くことである．このような方法を“stop-perceive-act”法と我々は呼んでいる．この方法は，マイクロフォンを搭載した大部分のロボットが採用している．アクティブオーディションでは，内部雑音抑制，あるいは，自己生成音の抑制はきわめて重要である．また，ロボットヒューマンインタラクションでは，自分の発話を削除し，相手の発話の信号雑音比を向上させることも必要である．

ロボット聴覚でのシンボルグラウンディング

本稿の最初に引用したように，人は音声によって言語を学習していく．岡ノ谷の解説記事[6]によれば，音声学習をするのは人，鯨，鳥の一部だけという．産声をあげるというのは，外敵があるような環境ではきわめて危険な行為であるが，そのような危険さを天秤にかけても音声コミュニケーションは重要というわけである．現在の音声認識システムは書き言葉に基づいており「知っていることしか聞けない」．知らない言葉は未定義語（Out of Vocabulary, OOV）という．しかし，人は初めての音でも擬音語として表現し，伝えることができる．つまり，音響信号を記号に落とすというシンボルグラウンディングが言語獲得の基本的な機能となっていると我々は考えている．

ロボット聴覚がロボットの知覚における重要な役割を果たすようになるためには，シンボルグラウンディングが不可欠である．音声，楽音，環境音のシンボルグラウンディングはそれぞれ音声認識，楽音認識，擬音語認識になるのではないであろうか．画像では，物体認識，擬態語認識になる．このような記号レベルでの認識がより高位の情報統合に不可欠であろう．

実環境でのロボット聴覚の例

後半では，2つの耳でどのような機能を実現できるかを実証するために，図-1に示したSIGの上に実現されている複数話者実時間追跡システムおよび複数話者同時発話認識を紹介する．詳細は下記のURLを参照していただきたい：

http://winnie.kuis.kyoto-u.ac.jp/SIG/

音響と画像を統合した実時間複数話者追跡

本システムのアイディアは，以下の通りである[7]：

1．HRTFの解析的な近似—ステレオ画像の3D位置を求めるエピポーラ幾何と同様に解析的にIPDとIIDを求める聴覚エピポーラ幾何を使用する．SIGの頭の形に対しては，1m以上離れた音源は無限遠にあると見なすことができ，ステレオ画像から3D位置が分かっているときには，正確な値が求まる．また，強度差や周辺での位相差をよりよく近似するために，球体に対する散乱理論により近似精度を高めている．

2．ヒューリスティクスによるモータノイズ抑制—SIGでは，体内に有する1対のマイクロフォンから取得されるモータ音や機械音については簡単なモデルを持っており，モータが稼働中でモデルに合うような音が発生すると，ヒューリスティクスを用いて，破壊されやすいサブバンドを破棄するようにしている．FIRフィルタを応用したアクティブノイズキャンセルは，IPDを計算するために必要な左右の位相特性の線形性が実データでは成立しない．

3．イベントレベルによる実時間視聴覚情報統合—画像

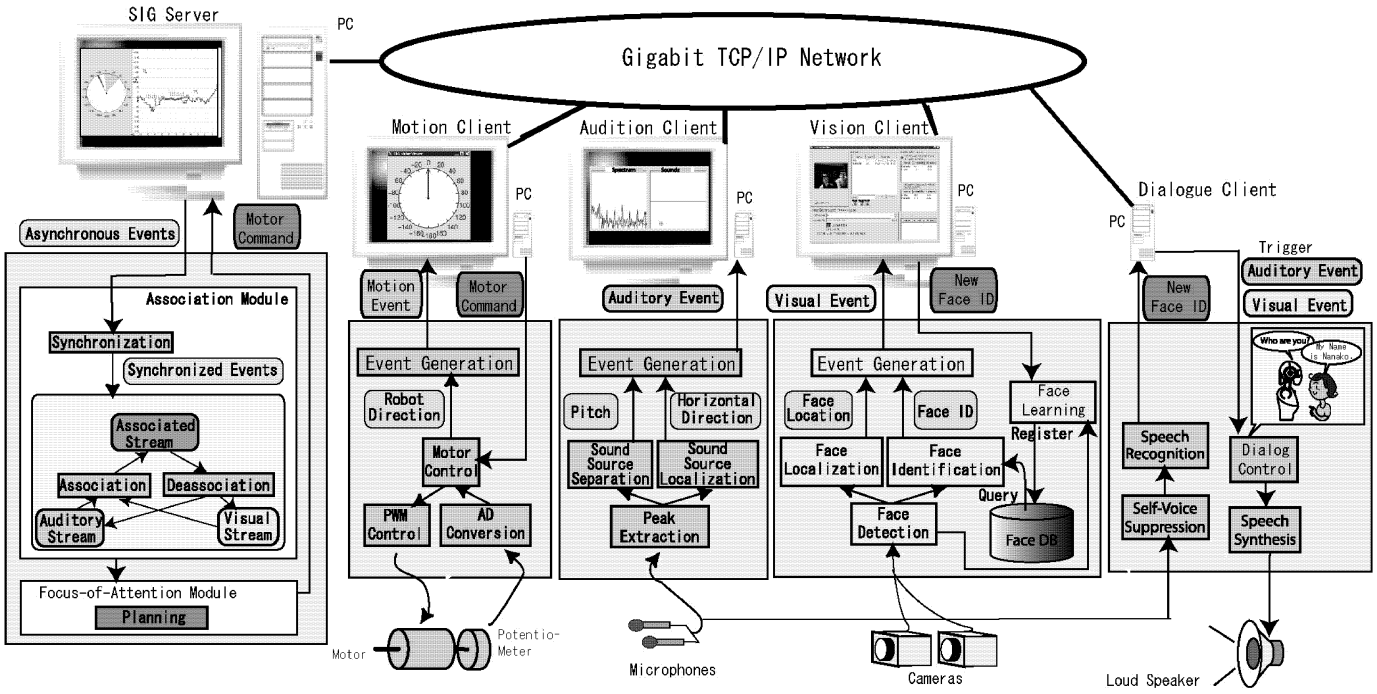


図-4　SIG上での視聴覚統合による実時間話者追跡システム

と音響の統合では，ステレオ画像による3D位置認識と音源定位の統合，顔認識と話者認識の統合話者認識，読唇術と音声認識の統合などさまざまなレベルがあるので，信号レベルで統合するのではなく，各モジュールから送付されるイベントをストリームとして抽出し，統合する．また，各モジュールは分散処理による実時間統合を目指す．

システム全体の構成を**図-4**に示す．システムは，音響処理部，画像処理部，モータ駆動部，アソシエーション部，対話管理部，注意制御部，および，サーバから構成されている．これらは5つのPC上に配置されており，Gigabit EthernetとFast Ethernetで接続されている．音響処理部は各サブバンドごと（FFTのポイントごと）にIPDとIIDを求めて，音源方向を抽出する．すべての方向について仮説を生成し，画像からの顔位置を統合することにより音源方向を抽出する．抽出した方向情報は，確信度とともにアソシエーション部に送られる．

画像処理部は，肌色抽出により顔を発見し，ステレオ画像処理により，3D情報を取得する．さらに，抽出した顔ごとに顔認識を行い，3次元情報と顔ID情報をアソシエーション部に送る．モータ駆動部からは，現在のSIGの体の向きについての情報がアソシエーション部に送られる．

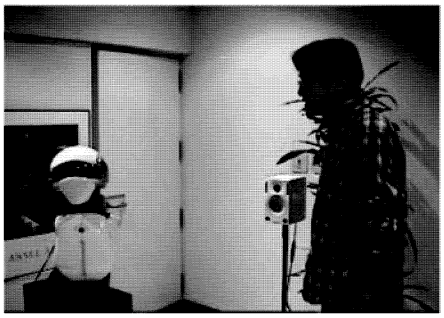
アソシエーション部では，各モジュールから得られる情報（方向や顔）を同期させ，音響ストリームと画像ストリームを構築する．次に，音響ストリームと画像ストリームを，時間的継続性や距離的近さをもとにグルーピングをし，アソシエーションストリームを構成する．一定時間，音響ストリームや画像ストリームが消失すると，アソシエーションは解除される．

注意制御部

注意制御部は，システムが保持する話者や人物情報をもとに，どの対象に注意を向け，正対するかを制御する．この部分は完全にプログラム可能であるので，いくつかのシナリオで，SIGの挙動を紹介する．

1．タスク指向型―たとえば，受付嬢のタスクでは，話している人に注意を向けて対応するのが第一目的である．そのようなストリームがなければ，新たな音ストリームに，すなわち音のする方向に振り向く．具体的な挙動を**図-5**に示す．この例では，来客が知らない人なので，「どちらさまですか」という応答をしている．既知の人の場合には，「こんにちは，XXさんでいらっしゃいますか」とその人の名前を呼び，確認を行う．このように，対話管理部は音声認識と音声合成を行う．

2．ソーシャル指向型―たとえば，コンパニオンの仕事を気配りと定義すると，新たな音のする方にまず注意を払うというように設計する．つまり，新たに生成された音ストリームに最優先で注意を振り向ける．4人

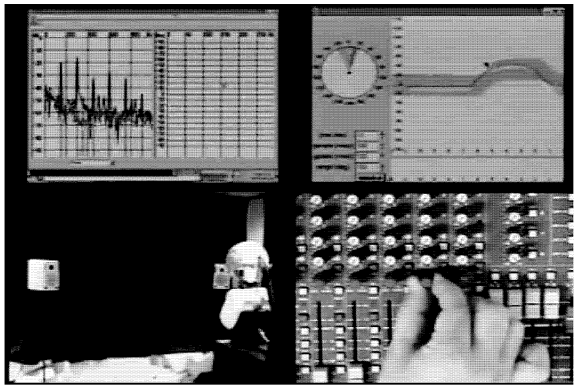


(a) 声をかけられると，声のする方に振り向き，顔認識をする．

(b) 顔認識に基づいて応答し，来客者の確認をする．

図-5　受付嬢としてのSIGの振る舞い

(a) 4人の中で声のする方を向く．



(b) ステレオバランスコントロールの変化に追従．

図-6　SIGの音源追跡のさまざまな評価実験

の人が声をかけると声をかけた中の1人の方を振り向くというデモを**図-6**（a）に示す．

3．追跡するものは，人物だけではなく，仮想的な音源であってもよい．基本周波数が100Hzの調波構造を持つ音を左右のスピーカから流し，そのバランスコントロールを変化させることで，仮想的に音源を左右にふる．すると，SIGは仮想的な音源を正しく追跡する（**図-6**（b））．もし，顔が発見されると，音ストリームと顔ストリームとがアソシエーションされ，SIGはその顔を見続ける．しかし，音源がさらに動き続けると，アソシエーションが破棄され，再び音源を追跡し出す．このような一種の音と画像による錯覚はこれまでほとんど報告されていない．

話者同時発話認識

SIG上で実現されている方向通過型フィルタ（Direction-Pass Filter, DPF）による分離音の品質は，正面の場合には高く，周辺になるに従って劣化する．そのために，分離音の品質が悪く，音声認識が失敗した場合には話者の方を向いて聞き直すというより高位のレベルでのアクティブオーディションが必要となる．**図-7**に示したような実験環境で三話者が同時に日本語の単語（日本人の英語も）を話し，DPFでそれぞれの音声を分離する．

分離した音声を音声認識するために，方向と話者に依存した音響モデルを作成する．具体的には，前面の180度に対して10度ごとにスピーカを通して150単語の学習データを出力し，各話者17個の音響モデルを三話者について計51個作成する．分離音をこれら51個の音響モデルそれぞれに対して音声認識を行い，音声認識結果の信頼度と方向情報と話者情報からそれぞれの発話の認識結果を得る（**図-8**）．

具体的な挙動としては，たとえば，数字を発声する場合には，認識がうまくいくとその合計をまず応答し，次に，それぞれの話者を向いて言った数を答えるということができる（図-7）．また，音声認識結果が曖昧な話者については，話者の方を向いて，候補を示して，どちらかをたずねる．また，顔認識を使い，特定話者に絞り込

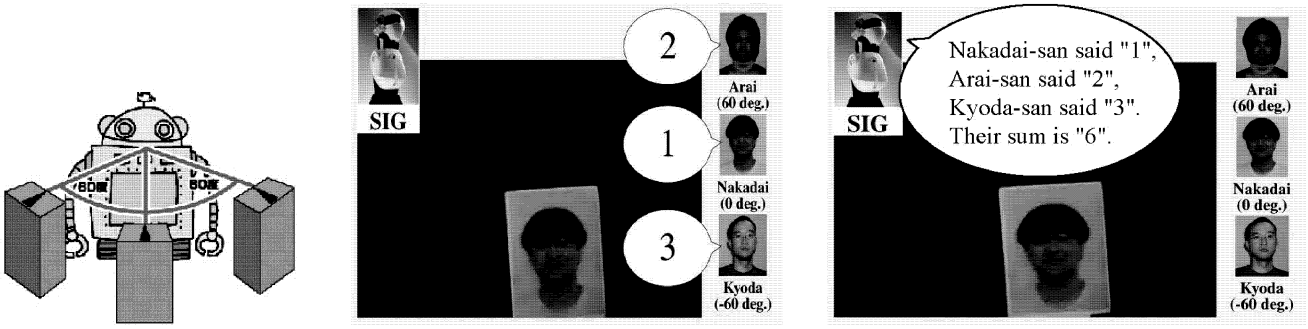


図-7　三話者同時発話認識実験の模式図（稲葉雅幸氏による）と実際の振る舞い

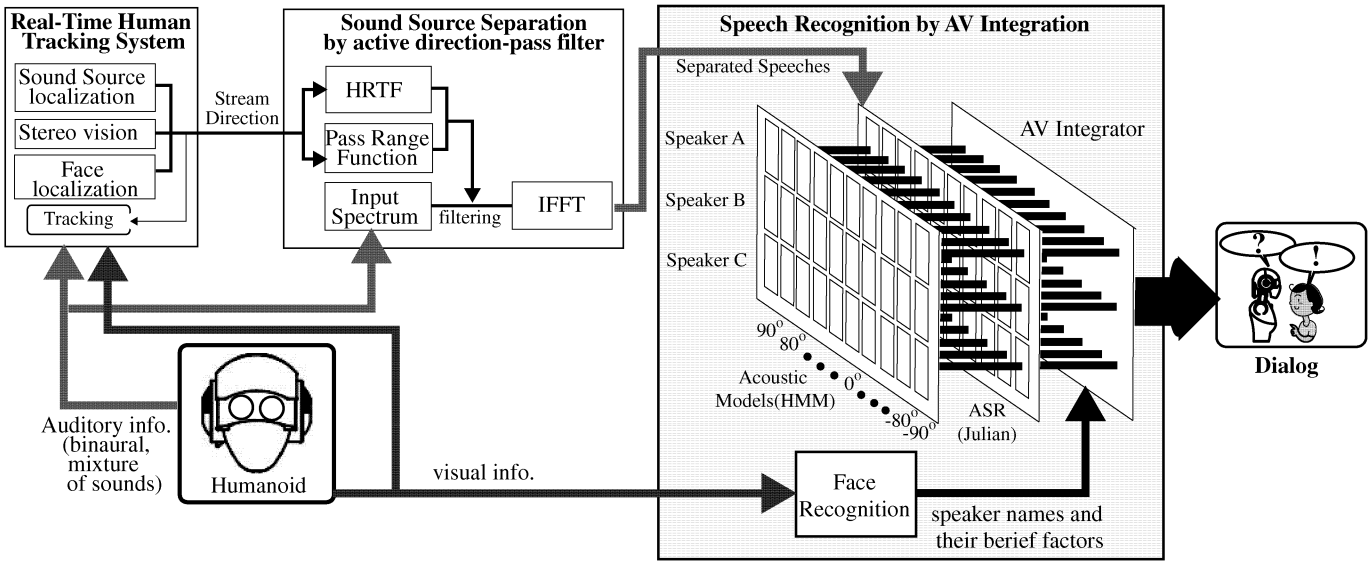


図-8　分離音を複数の音響モデルを用いて音声認識

み17個の音響モデルだけを使うようにすると，計算時間はさらに減少する．

このように小さな語彙の特定話者単語認識ではあるが，三話者同時発話認識ができている．本実験では多数の制約の下で，聖徳太子のように10人とはいかないまでも，簡単な聖徳太子ロボットの実現可能性を実証することができた．

新しい地平の開拓を目指して

米国では，ITバブルの崩壊とともに音声研究のメッカであったLucent Bell研究所やAT&T研究所で音声研究グループの解散や大幅縮小があったと聞いている．音声認識は目標に到達することなく，技術が飽和したと評価されているのではないだろうか．本稿で述べたようなロボット聴覚という実際の応用から，音声や一般の音を含めた音情報処理研究を進めていくことが，ロボットなどの組み込みシステムの知能化やマンマシンインタフェースの高度化に不可欠ではないかと考えている．

最後に，科学技術振興事業団北野共生システムプロジェクト総括責任者北野宏明氏に感謝する．

参考文献

1) Breazeal, C.L.: *Designing Sociable Robots*, MIT Press, Cambridge, MA. (2002).
2) 奥乃　博: 音環境理解－混合音の認識を目指して，情報処理，Vol.40, No.10, pp.1096-1101 (Oct. 2000).
3) Rosenthal, D. and Okuno, H.G. editors.: *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, Mahwah, New Jersey (1998).
4) 奥乃　博，中臺一博: ロボットの耳は二つで十分か，日本音響学会誌，Vol.58, No.3, pp.205-210 (Mar. 2002).
5) Aloimonos, Y., Weiss, I. and Bandyopadhyay, A.: Active Vision, *International Journal of Computer Vision*, Vol.1, No.4, pp.333-356 (1987).
6) 岡ノ谷一夫: 身体的「知」の進化と言語的「知」の創発，人工知能学会誌，Vol.18, No.7, pp.392-398 (July 2003).
7) Nakadai, K.,　Hidai, K.,　Mizoguchi, H.,　Okuno, H.G. and Kitano, H.: Real-time Auditory and Visual Multiple-Object Tracking for Robots, In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (*IJCAI-01*), pp.1425-1432. IJCAI (2001).

（平成15年9月22日受付）