展 望

ロボット聴覚の現状と展望

Robot Audition: Present and Future

奥乃 博* *京都大学大学院情報学研究科
Hiroshi G. Okuno* *Kyoto University, Graduate School of Informatics

本特集では，**ロボットが自分自身の耳で聞く**というロボット聴覚研究が広範囲にサーベイされているので，個別技術の詳細はそちらに譲り，本稿では筆者が共同研究者の中臺一博氏（現 HRI-JP）とともにこの 10 年間に従事してきた研究の動機を説明し，今後の展望や夢について考えてみたい．

1. ロボット聴覚は聞き分ける技術がベース

鉄腕アトムには「スイッチひとつで聴力が千倍になり，遠くの人の声もよく聞こえ，さらに 2 千万ヘルツの超音波も聞きとる」サウンドロケータが装備されているという†．サウンドロケータは，1953 年に Cherry が発見した選択的に音声を聞き分ける「カクテルパーティ効果」を実現するスーパーデバイスなのであろう．聴覚障害者や耳の聞こえが悪くなった高齢者には「スーパーデバイスでなくても，常時同時発話が聞き分けられる機能じゃだめなの」という素朴な疑問がわく．

日本書紀推古紀には，「一聞十人訴，以勿失能辨」とあり，同時に 10 人の訴えを聞き分けて裁いたという「聖徳太子」の逸話が紹介されている．動物や草木の言葉が聞こえるという「聞き耳頭巾」の昔話は子供たちの想像力をかき立てる．このような聞き分け機能をロボットに持たせることができれば，人との共生が大きく前進すると期待される．

日常生活で最も重要なコミュニケーション手段が話声や歌声などを含めた音声であることは論を俟たない．音声コミュニケーションは，言語獲得，非音声によるバックチャネルなどを包含し，その機能は極めて多彩である．実際，自動音声認識（ASR, Automatic Speech Recognition）研究の重要性は高く認識され，過去 20 年以上に渡り膨大な資金と労力が投入された．一方，ロボット自身に装着されたマイクロフォンで音を聞き分け，音声認識をするシステムの研究は麻生らの仕事を除き，ほとんど取り組まれてこなかった．

筆者らの研究スタンスは事前知識最小の音の処理方式を開発することであった．そのために，音声だけでなく，音楽，環境音，さらにはそれらの混合音の処理を通じて音環境を分析理解する音環境理解の研究が重要であると考えた．この立場から，単一音声入力を仮定する現行の ASR がロボット学で重要な役割を果たせ切れていないことへの説明が付く．

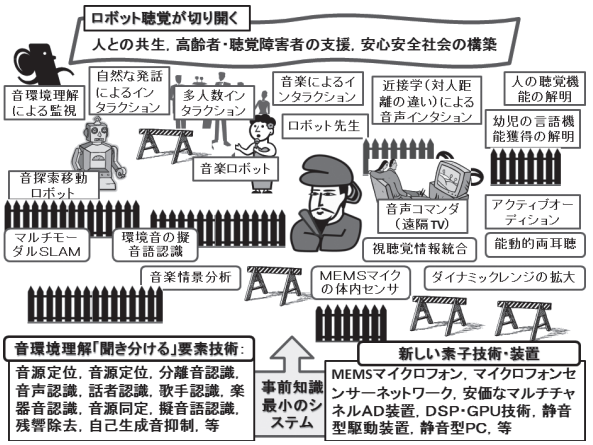

図 1 音環境理解をベースとしたロボット聴覚の展開

2. 音環境理解をベースにしたロボット聴覚

音声に加えて音楽や環境音さらには混合音を含めた音一般を扱う必要があるという立場から，音環境理解（Computational Auditory Scene Analysis）[1] 研究を進めてきた．音環境理解研究での重要な課題は，混合音の処理である．話者の口元に設置した接話型マイクロフォンを使用して混合音の問題を回避するのではなく，入力は混合音との立場から，混合音処理に直球で立ち向かうのが音環境理解である（**図 1** 参照）．

音環境理解の主たる課題は，音源方向認識の**音源定位**（sound source localization），**音源分離**（sound source separation）**分離音の音声認識**（automatic speech recognition）の 3 つである．個々の課題に対してはこれまでに多種多様な技術が研究開発されている．しかし，いずれの技術もその能力を最大限発揮するためには何らかの条件を前提としている．ロボット聴覚でこれらの技術を組み合わ

原稿受付 2009 年 11 月 30 日
キーワード：Robot Audition, Active Audition, Computational Auditory Scene Analysis, Microphone Array, Binaural Hearing
*〒 606–8501 京都市左京区吉田本町
*Yoshida-Honmachi, Sakyo, Kyoto 606–8501, Japan

†http://www31.ocn.ne.jp/˜goodold60net/atm_gum3.htm

せ，能力を最大限発揮させるためには，個別技術のインタフェース，すなわち，前提条件をうまく揃えて，システム化することが不可欠である．このためには，ドベネックの桶（リービッヒの最小律）ではないが，バランスのよい組み合わせを効率よく提供できるミドルウエアも重要となる．

本号解説で筆者が紹介しているロボット聴覚ソフトウエア **HARK** [2] は，FlowDesigner というミドルウエアの上に構築されており，8 本のマイクロフォンを前提として，音環境理解の機能を提供している．HARK は，事前知識を極力減らすという原則で設計されており，“音響処理の OpenCV” を目指したシステムである．実際，3 人の料理の注文を聞き分けるロボットや口によるじゃんけんの審判ロボットなどが複数のロボットで実現されている．

一般には画像や映像が主たる環境センサとなっているものの，見え隠れや暗い場所には対応できず，必ずしも万能というわけではない．音情報を使って，画像や映像での曖昧性を解消し，逆に，音響情報での曖昧性を画像情報を使って解消する必要がある．例えば，2 本のマイクロフォンによる音源定位では，音源が前か後ろかの判断は極めて難しい．

3. 人のように 2 本のマイクロフォンで聞き分ける

人や哺乳類は 2 つの耳で聞き分けを行っている．ただし，頭を固定した実験ではたかだか 2 音しか聞き分けられないことが報告されている．人の音源定位機能のモデルとしては，両耳入力に遅延フィルタをかけて和を取る Jeffress モデルと，両耳間相互相関関数によるモデルがよく知られている．中臺と筆者らは，ステレオビジョンにヒントを得て，調波構造を両耳で抽出し，同じ基本周波数の音に対して，両耳間位相差と両耳間強度差を求めて，音源定位を行っている [3]．一対の参照点を求めるのに，ステレオビジョンではエピポーラ幾何を使用し，我々の方法は調波構造を使用する．

2 本のマイクロフォンによる混合音からの音源定位では，定位が安定せず大きくぶれることが少なからずあり，また，前後問題，特に，真正面と真後ろにある音源を区別するのが難しい．中臺らは視聴覚情報統合により安定した音源定位を実現するとともに，SIG というロボットで呼びかけられたら振り向くロボットを実現している [3]．前後問題の曖昧性解消は百聞は一見に如かず，というわけである．

金と筆者は，SIG2 というロボットに頭を動かすことにより音源定位の曖昧性を解消するシステムを実現している．単純に頭を左右に 10° 動かすだけでなく，音源が 70〜80° にあるときには，下向きに 10° 頷きを入れるとよい．実際，正面の音源同定では 97.6%と 1.1%の性能向上に過ぎないのに対して，後ろの音源同定では 75.6%と 10%大幅に性能が向上する（**図 2**）．これは，Blauert が “Spatial Hearing”（MIT Press）で報告している人の前後問題の解消時の頭の動きとよく一致している．曖昧性の解消のために挙動を用いる方法は**アクティブオーディション**の 1 形態である．

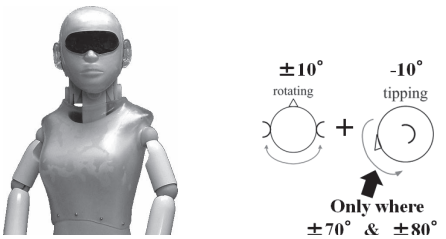

図 2 SIG2 のアクティブオーディション：周辺部の音に対しては首を左右と下に動かして前後問題の曖昧性を解消する

公文のグループや中島のグループは，様々な耳介を用いて頭や耳介自身を動かすことで音源定位の性能向上に取り組んでいる [3]．ちょうど，ウサギの耳が通常は垂れ下がって広範囲な音を聞いており，異常音がすると耳が立ち上がり，特定方向の音を聞くために指向性を高める．このようなアクティブオーディションの実現法の基礎研究である．これが，ロボットだけでなく，様々な動物の聴覚機能の構成的解明に応用できると，新たなロボットの耳の設計開発につながっていくと期待される．特に，両耳聴は，ステレオ入力装置がそのまま使えるので，高性能の両耳聴機能が実現できると，工学的な貢献が大きいと考えられる．

4. 自己生成音抑制機能

アクティブオーディションでは，モータが動くことにより発生するモータ自身の音に加えてロボット自身の体の軋みから音が発生することがある．ロボットの動きに伴って発生する音は，小さい音であっても音源がマイクロフォンの近くにあるので，逆 2 乗則から外部の音源と比較して相対的に大きな音となる．

4.1 モデルベースによる自己生成音抑制

中臺らはロボット SIG の頭部内部にマイクロフォンを 2 本設置し，自己生成音の抑制を試みている．モータ音や機械音について簡単なテンプレートを持ち，モータの稼働中にテンプレート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．本手法を用いた理由は，FIR フィルタに基づくアクティブノイズキャンセラでは，左右の耳が別々に処理されるので両耳間位相差を正しく求めることができないからであり，さらに，バースト性雑音の抑制に FIR フィルタがあまり効果がなかったからである．なお，SIG2 では，マイクロフォンが人の外耳道モデルに埋め込まれており，モータも静音型なので，雑音抑制処理は行っていない．ソニーの QRIO でも体内に 1 本マイクロフォンを設置し，外部を向いた 6 本のマイクロフォンを使用して自分の出す雑音を抑制している．

Ince らは，自分の動きから生じる自己生成雑音を，関節角の情報から予測し，スペクトルサブトラクション法により削減する方法を開発している [3]．中臺らは，特定の方向

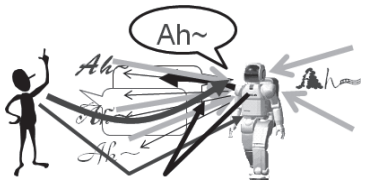

図 3 自分の話声が残響を伴って自分の耳に入り，さらに，相手の割り込み発話（バージイン）も聞こえる

からのモータ雑音を棄却する機能を HARK に組み込んでいる [3]．Even らは，体内に設置した 3 個の振動センサを使って，体表から放射される音の方向を推定し，その放射音方向と話者方向が一致しないように線形マイクロフォンアレイの角度を調節し，自己生成音の抑制を行っている [3]．

ロボットが人とインタラクションを取るときには，自己生成音の影響，環境による音への影響を勘案して，最もよく聞こえる位置に移動したり，体の向きを変えるといった「よりよく聞くための戦略」の開発が不可欠である．

4.2 セミブラインド分離による自己生成音抑制機能

ロボット聴覚では，自己発話信号がロボット自身に既知である点を活用した自己生成音抑制が可能である．武田らは，**図 3** に示した状況において，自己発話を既知として，その残響成分を推定し，入力混合音から自己発話を抑制し，相手の発話を抽出する自己生成音抑制機能を独立成分分析（ICA）に基づいたセミブラインド分離技術より開発している [3]．本技術の応用のプロトタイプとしてバージイン許容発話認識と音楽ロボット（後述）が開発されている．

バージイン許容発話とは，ロボットの発話中でも人が自由に発話ができる機能である．ロボットが項目を列挙して情報提供を行っているときに，ユーザが割り込んで「それ」「2 番目の」「アトム」と発話すると，本技術を応用して，発話内容や発話タイミングからどの項目が指定されたか従来よりは高性能で判定することができる．人とロボットが共生していくためには，交互に話すのではなく，いついかなるときでもお互いに自由に話すことができる混合主導型のインタラクションが不可欠であり，本自己生成音抑制機能によってそのような機能が容易に実現できる．

セミブラインド分離技術は，自己生成音が耳まで入るが，分離されると捨てられ，高次処理の対象となっていない．本庄の『言葉をきく脳しゃべる脳』によると，成人では自分の声が側頭葉の 1 次聴覚野までは入るが，大脳皮質の連合聴覚野には送られず，聞き流していることが観測されている．上述のセミブラインド分離による自己生成音抑制は 1 次聴覚野止まりの処理の工学的実現ととらえることもできよう．

5. 視聴覚情報統合による曖昧性解消

ロボット聴覚は**要素技術ではなく，プロセスであり，複数のシステムから構成される**．構成部品となる要素技術は多数あり，構成部品の性能にはばらつきがあるので，プロセスではすべてがうまくかみ合って機能する必要がある．しかも，このかみ合わせがしっかりするほど，プロセスはうまく機能する．音響処理だけでは曖昧性が解消できないので，視聴覚情報統合がかみ合わせの重要な鍵となる．

情報統合のレベルには，時間的，空間的，メディア間，システム間があり，さらに，各レベル内でも，レベル間でも階層的な情報統合が必要である．中臺らは次のような視聴覚情報統合を提案している．最下位レベルでは音声信号と唇の動きから話者を検出する．その上のレベルでは，音素（phoneme）認識と口形素（viseme）認識とを統合する．その上位レベルは，話者位置と顔の 3D 位置との統合である．最上位は，話者同定・検証と顔同定・検証との統合である．もちろん，同一レベルの情報統合だけでなく，ボトムアップ処理やトップダウン処理の相互作用が考えられる．

一般に混合音処理は不良設定問題であり，より完全な解を得るためには，何らかの前提，例えばスパースネスの仮定が必要となる．時間領域でのスパースネス，周波数領域でのスパースネス，3D 空間でのスパースネス，さらには特徴空間でのスパースネスなどが考えられる．情報統合の成否は，スパースネスの設計だけでなく，個々の要素技術の性能にも依存することに注意する必要がある．

6. ロボット聴覚が切り開くキラーアプリケーション

ロボット聴覚機能が充実しても，それは，個々の信号処理モジュールの統合であり，それからどのような応用が見えてくるのかは明らかでない．実際，音声認識は IT 事業のなかでも非常に低い地位しか与えられていない．そのような現状から，本当に不可欠な応用を見つけるためには，まず，使えるシステムを構築し，経験を積んでいく必要があろう．

6.1 近接学によるインタラクション

インタラクションの基本原理として，対人距離に基づく近接学（Proxemics）が知られている．すなわち，親密距離（〜0.5 [m]），個人距離（0.5〜1.2 [m]），社会距離（1.2〜3.6 [m]），公共距離（3.6 [m]〜）に分け，各距離ごとにインタラクションの質が変わってくる．

近接学に対するロボット聴覚の課題は，マイクロフォンのダイナミックレンジが拡大することである．複数人インタラクションにおいて，個々の話者が同じ音量で話すとすると，遠方の話者の声は逆 2 乗則に従って小さくなる．従来の 16 ビット入力では不足し，24 ビット入力で対応することが不可欠である．システム全体を 24 ビット化するのは，計算資源や既存ソフトウェアとの整合性から難しい．荒井らは，情報欠損の少ない 16 ビットへのダウンサンプリング法を提案している [3]．また，マルチチャネル AD 装置や携帯電話用 MEMS マイクロフォンなど，新しい装置の出現にも対応していく必要もある．

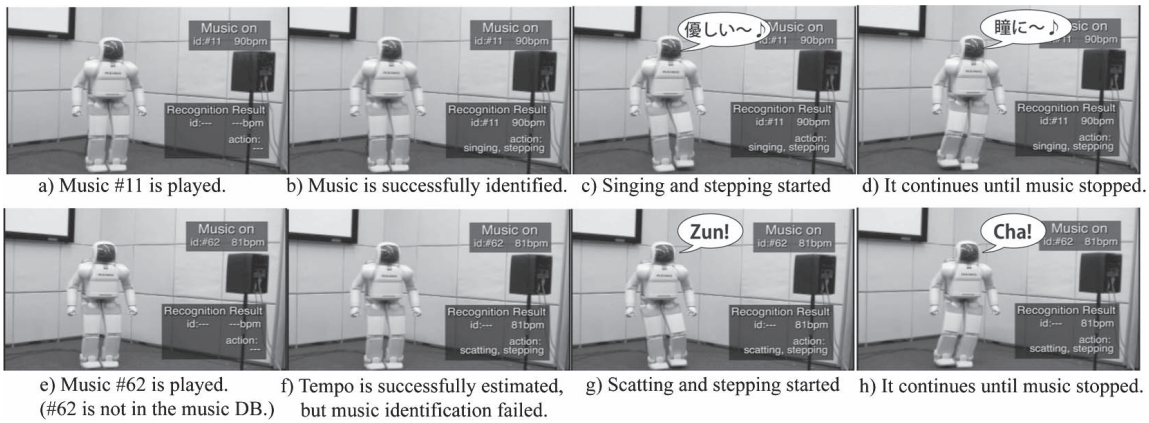

図 4 1 [m] 離れたスピーカから流れる音楽を自分の耳で聞き分け，曲に合わせて歌い，足踏みをするロボット．曲が終わると動きを止め，曲が始まるとまた動き始める

6.2 音楽ロボット

音楽を聴けば自然と体が動き，インタラクションが円滑になるので，音楽インタラクションへの期待は大きい．ロボットが音楽を扱えるようになるには，「聞き分ける」機能が不可欠である．**図 4** に示した ASIMO をテストベッドとして開発した音楽ロボット処理の流れを示す．

(1) 自己生成音を入力音（混合音）から抑制あるいは分離，
(2) 分離音のビート追跡からテンポ認識と次テンポ推定，
(3) テンポに合わせて挙動（歌を歌う，動作）を実行．

ロボットは，スピーカから音楽が鳴るとすぐにテンポに合わせて足踏みを始め，音楽が鳴り終わると足踏みを終える．

自分の歌声を残響の影響を含めて入力混合音から分離するのために自己生成音抑制機能を使用している．ビート追跡やテンポ推定では誤りが避けられない．音楽ロボットでは，テンポ推定誤りから生ずる楽譜追跡時の迷子からいかに早く，かつ，スマートに合奏や合唱に復帰するかが重要であり，人とのインタラクションで不可欠な機能となっている．

6.3 視聴覚統合型 SLAM

佐々木・加賀美（産総研）らは，32 チャネルマイクロフォンアレイを装着した移動ロボットを開発し，室内の音環境理解の研究開発に取り組んでいる．事前に与えられたマップを使い，いくつかのランドマークをたどりながら定位とマップ作成を同時に行う SLAM（Simultaneous Localization And Mapping）の音響版である [2]．従来の SLAM では，画像センサ，レーザレンジセンサ，超音波センサなどが使われるものの，マイクロフォン，つまり，可聴帯域の音響信号は使用されてこなかった．佐々木らの仕事は，従来の SLAM では扱えていなかった音響信号を SLAM に組み込む研究であり，重要な先駆的な研究である．これにより，見えないけれども音がする場合にも，SLAM あるいは音源探索が可能となり，真の情景理解（Scene analysis）や環境理解への道筋が開かれたことになると考えられる．

7. おわりに

ロボットが自分自身の耳で聞くという**ロボット聴覚研究**についての筆者の考え方を述べるとともに，今後の展開への期待を述べた．ロボット聴覚研究は，ほとんど 0 からの立ち上げであったために，自分たちの研究だけでなく，当該研究の振興を図るべく浅野（産総研，以下敬称略），小林（早大），猿渡（奈良先端大）らのアカデミア，NEC，日立，東芝，HRI-JP などのロボット聴覚を展開する企業，さらには，カナダ Sherbrooke 大学，韓国 KIST，フランス LAAS，ドイツ HRI-EU などの海外研究機関からの協力を得て，IEEE/RSJ IROS でこれまでに 6 年間ロボット聴覚 organized session を組み，本学会学術講演会でも 4 年間特別セッションを組んでいる．さらに，2009 年には IEEE 信号処理部門の国際会議 ICASSP-2009 でロボット聴覚スペシャルセッションを開催した．このような研究コミュニティの育成により，世界的に徐々に研究者が増加し，そのなかでも日本のロボット聴覚研究のレベルの高さが輝いている．今後斯学のますますの発展を通じ，聖徳太子ロボットが聴覚障害者や高齢者の支援，安心できる社会の構築に寄与していくことを期待したい．

参考文献

[1] Rosenthal, D. and Okuno, H.G. (eds.): Computational Auditory Scene Analysis. Lawrence Erlbaum Associates, 1998.
[2] 中臺，宮下（編）：特集「ロボット聴覚」，日本ロボット学会誌，本号．
[3] 人工知能学会 AI チャレンジ研究会資料．Web より入手可能：http://winnie.kuis.kyoto-u.ac.jp/AI-Challenge/

奥乃 博（Hiroshi G. Okuno）
京都大学大学院情報学研究科教授，博士（工学）．1972 年年東京大学教養学部基礎科学科卒業．NTT 研究所，JST，東京理科大学を経て，現職．ロボット聴覚，音環境理解，音楽情報処理，人工知能研究に従事．六十而耳順（「論語・為政」）．（日本ロボット学会正会員）