社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

ロボット聴覚

～高雑音下でのハンズフリー音声認識～

中臺　一博[†,††]　奥乃　博[†††]

† (株) ホンダ・リサーチ・インスティチュート・ジャパン　〒 351–0188 埼玉県和光市本町 8–1
†† 東京工業大学大学院情報理工学研究科　〒 152–8552 東京都目黒区大岡山 2–12–1
††† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †nakadai@jp.honda-ri.com, ††okuno@kuis.kyoto-u.ac.jp

あらまし　我々が取り組んでいるロボット聴覚研究について，その位置づけや意義を解説し，これを実現するための高雑音下ハンズフリー音声認識へも適用可能な技術としてマイクロホンアレイを用いた動的環境下の実時間音源分離とその音声認識への適用について紹介する．紹介する技術は，ロボット聴覚ソフトウェア HARK としてオープンソースで公開を行っている．そこで，これらの技術の有効性を，実際のロボットへの HARK の適用事例を通じて示す．
キーワード　ロボット聴覚，HARK，マイクロホンアレイ，動的環境，実時間音源分離，音声認識

Robot Audition

– Hands-Free Automatic Speech Recognition under Highly-Noisy Environemnts –

Kazuhiro NAKADAI[†,††] and Hiroshi G. OKUNO[†††]

† Honda Research Institute Japan Co., Ltd.　Honcho 8–1, Wako-shi, Saitama, 351–0188 Japan
†† Graduate School of Information Science and Engineering, Tokyo Institute of Technology
O-okayama 2–12–1, Meguro-ku, Tokyo, 152–8552, Japan
††† Graduate School of Informatics, Kyoto University Yoshidahonmachi, Sakyou-ku, Kyoto 606–8501, Japan
E-mail: †nakadai@jp.honda-ri.com, ††okuno@kuis.kyoto-u.ac.jp

Abstract　This paper addresses robot audition, which realizes listening capabilities for robots using robot-embedded microphones. For robot audition, we propose real-time sound source separation and automatic speech recognition (ASR) techniques for dynamically changing environments based on microphone array processing, which is applicable to hands-free ASR under highly-noisy environments. Implementation of the proposed techniques is open-sourced as robot audition software called “HARK.” We show the effectiveness of these techniques through applications of HARK to robots.
Key words　robot audition, HARK, microphone array, dynamically-changing environment, real-time sound source separation, automatic speech recognition

1. ま え が き

実環境でロボットが人と自然にコミュニケーションを行う上で，音声認識は最も重要な機能の一つである．我々は，ロボットが自分の耳 (頭部に装着されたマイク) を用いて，実環境で音声を含めた任意の音源の定位，分離，同定，認識などを統合的に行う音環境理解を実現するため，**「ロボット聴覚」**を提案した [1]．これまで，ロボットの音声認識向上という観点から，特に，音源定位・音源分離・音声認識といった機能に着目した研究を行ってきた [2], [3]．研究領域としても，ロボットの最も大きな国際会議の一つ，International Conference on Intelligent Robots and Systems (IROS) で 7 年に渡るオーガナイズドセッション開催や，ICASSP 2009 での，ロボット聴覚オーガナイズドセッション開催を通じて，国内外に広がりを見せている．

同様の研究はハンズフリー音声認識という形で，音声処理コミュニティでも行われてきた．スペクトラルサブトラクション [4] に代表されるシングルチャネルアプローチも多くの報告があるが，近年では，マイクロホンアレイを用いたマルチチャ

ネルアプローチが盛んに研究されており，Hands-free Speech Communication and Microphone Arrays (HSCMA) といったこのトピックにフォーカスした国際学会も開催されるようになってきた．

いずれの研究分野でも，方向性雑音や拡散性雑音を含む加法性雑音，および残響といった雑音を広く扱っているものの，ロボット特有の，動的環境を扱う研究は少ない．ロボットにおける音源分離については，両耳聴の知見を応用したアクティブ方向通過型フィルタ [2] を用いて極めて限定された環境ながら，2 本のマイクロホンで 3 話者の同時発話の分離・認識が報告されている．Valin らは，マイクロホンアレイを用いて，ビームフォーミングとブラインド音源分離のハイブリッドアルゴリズムである Geometric Source Separation (GSS) に基づき，実時間オンライン音源分離 [5] を報告している．さらに，Yamamoto らは，ミッシングフィーチャ理論を用いて，GSS と音声認識を統合し，同時発話の料理注文タスクに適用した [3]．Hara らは，適応ビームフォーマーを用いた音源分離を音声認識と接続し，オフィス環境でテレビの音声制御デモを構築した [6]．猿渡らは，2 本のマイクロホンを用いて，高精度に音源分離を行う SIMO-ICA を提案 [7] し，専用のハードウェアを開発し，生駒市の案内を行うキタちゃん用のロボット対話システムに適用した．しかし，これらのシステムは，基本的に静的環境で用いられることを前提としているという制約があった．本稿では，動的環境の中でも，話者もしくはロボット（マイクロホンアレイ）が移動するような場合を考慮した音源分離および音声強調法を紹介する．

また，我々は，複数のマイクロホン（マイクロホンアレイ）からの入力をもとに，音源定位・追跡・分離，さらに分離音声の認識までを総合的にサポートするロボット聴覚用のオープンソースソフトウェアとして， HARK (HRI-JP Audition for Robots with Kyoto Univ.) の研究開発を行っている．HARK を用いれば，GUI プログラミング環境上で様々なモジュールを配置・接続して，形状やマイクロホンのレイアウトが異なるロボットへの対応や，用途に合わせたロボット聴覚システムの構築が可能である．本稿で紹介する手法も，HARK の新リリース（2010 年 11 月公開）に実装されており，自由にダウンロードして利用することができる．本稿で紹介する手法の有効性を HARK の実ロボットへの適用を通じて示す．

2. 動的環境を扱うための課題

ロボット聴覚で，動的環境を扱うためには，多くの課題がある．図 1 は，基本的なロボット聴覚システムのフローを示しており，音声認識に至るまでに，音源定位, 音源追跡，音源分離・音声強調といったブロックを配置している．動的環境を扱うためには，各ブロック毎に対応を行う必要がある．音源定位や追跡では，これまでにも動作によるロボット聴覚向上，つまり，低レベルのアクティブ聴覚に関する研究が報告されている [2], [8], [9]．我々も，アクティブ聴覚のための要素技術として，マイクロホンアレイを用いて動的な雑音を扱うための音源定位の枠組み [10] や，複数の移動音源を扱う枠組み [11] を報告している．しかし，音源分離・音声強調については，明示的に，動的な環境を扱うための研究は報告されていない．本稿では，動的な環境を扱うために重要である **1) 動的変化への高追従性**, **2) パラメータチューニングの簡便性**を備えた音源分離・音声強調法を紹介する．

3. 動的環境を考慮した音源分離と音声強調

3.1 動的環境を考慮した音源分離 GHDSS-AS

これまで，音源分離法として，*Geometric Source Separation (GSS)* を用いてきた．GSS のブラインド音源分離に対応する規範には出力信号の二次のパワー相互相関を用いた単純なものであったため，独立成分分析のように高次の相関情報を用い，かつ変化に対する追従性の高い *Geometrically constrained High-order Decorrelation based Source Separation with Adaptive Stepsize control (GHDSS-AS)* を新たに提案した [12].

GHDSS-AS の定式化は以下の通りである．M 個の音源と N ($\geqq M$) 個のマイクロホンがあるとする．周波数 ω における M 個の音源に対するスペクトルのベクトルを $\mathbf{s}(\omega) = [s_1(\omega)\ s_2(\omega)\ \cdots\ s_M(\omega)]^T$ とする．同様に，周波数 ω における N 個のマイクロホンで収録した信号のスペクトルのベクトルを ω, $\mathbf{x}(\omega) = [x_1(\omega)\ x_2(\omega)\ \cdots\ x_N(\omega)]^T$ とする．この時，$\mathbf{x}(\omega)$ は，以下の式で表すことができる．

$$\mathbf{x}(\omega) = \mathbf{D}(\omega)\mathbf{s}(\omega), \tag{1}$$

ここで，$\mathbf{D}(\omega)$ は，音源とマイクロホン間の伝達関数行列であり，伝達関数行列の各コンポーネント H_{nm} は m 番目の音源から n 番目のマイクロホンへの伝達関数を表す．この時音源分離は，以下の式で表される．

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega), \tag{2}$$

ここで，$\mathbf{W}(\omega)$ は**分離行列**と呼ばれる．一般的な音源分離問題は，出力信号 $\mathbf{y}(\omega)$ が $\mathbf{s}(\omega)$ となる $\mathbf{W}(\omega)$ を求めることに帰着する．$\mathbf{W}(\omega)$ を求めるため, GHDSS-AS は, ブラインド分離に対応する J_{SS} とビームフォーマに対応する J_{GC} の 2 つのコスト関数を導入する．

$$J_{SS}(\mathbf{W}) = \|\phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H]\|^2 \tag{3}$$

$$J_{GC}(\mathbf{W}) = \|\mathrm{diag}[\mathbf{WD} - \mathbf{I}]\|^{\mathbf{2}} \tag{4}$$

ここで，$\|\cdot\|^2$ はフロベニウスノルム，diag[·] は対角成分，$E[\cdot]$ は期待値，H はエルミート転置をそれぞれ表す演算子を表す．また，$\phi(\mathbf{y})$ は非線形関数であり，以下のように定義される．

$$\phi(\mathbf{y}) = [\phi(y_1), \phi(y_2), \cdots, \phi(y_N)]^T \tag{5}$$

$$\phi(y_i) = -\frac{\partial}{\partial y_i}\log p(y_i).$$

実際には，$\phi(y_i)$ には様々な定義があり，本稿では，澤田らによって提案された以下の関数 [13] を用いた．

$$\phi(y_i) = \tanh(\eta|y_i|)e^{j\cdot\theta(y_i)}, \tag{6}$$

η はスケールパラメータである．

図 1　ロボット聴覚システムの基本的な流れ
Fig. 1　A Basic Flow of Robot Audition Systems

最終的なコスト関数は，以下のように表される．

$$J(\mathbf{W}) = \alpha J_{SS}(\mathbf{W}) + J_{GC}(\mathbf{W}), \tag{7}$$

ここで α は，2つのコスト関数間の重みパラメータを表す．

観測信号 $\mathbf{x}$ が十分に長ければ，オフライン処理で直接，$J(\mathbf{W})$ を最小化すする最適な $\mathbf{W}$ を推定することが可能であるが，ロボットは，リアルタイムシステムであり，周囲の環境が動的に変化すすることから，以下のように $\mathbf{W}$ をインクリメンタルに更新していく必要がある．

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu_{SS}\mathbf{J'}_{SS}(\mathbf{W_t}) + \mu_{GC}\mathbf{J'}_{GC}(\mathbf{W_t}). \tag{8}$$

ここで，$\mathbf{W}_t$ は，時刻 t の $\mathbf{W}$，$\mathbf{J'_{SS}}(\mathbf{W})$ および $\mathbf{J'_{GC}}(\mathbf{W})$ は，$\mathbf{J_{SS}}(\mathbf{W})$ および $\mathbf{J_{GC}}(\mathbf{W})$ の複素勾配 [14] を表す．また，μ_{SS} と μ_{GC} は，ステップサイズパラメータと呼ばれる．一般にこれらのステップサイズパラメータは手動で固定値を設定することが多いが，実際の利用では，最適なステップサイズを手動でいちいち求めていたのでは使い勝手が悪く，また，動的環境下では，最適値が常に変化することから，GHDSS-AS はこれらのステップサイズを常に最適に保つ適応ステップサイズ制御法を包含している．適応ステップサイズ制御法は，エコーキャンセルの分野でよく用いられる手法である [15]．我々は，これを，多次元ニュートン法と複素勾配行列の線形近似を用いて，複素信号の多チャンネル入力に拡張した [16]．この手法を用いれば，音源の移動などにより，分離エラーが大きくなった場合は，大きなステップサイズに，分離行列が収束して，分離エラーが小さくなった場合は小さいステップサイズになり，結果として分離行列が高速に収束，もしくは分離行列の変化に拘束に追従することができる．具体的には，適応ステップサイズ制御を用いて，式 (8) を以下のように書き換えて用いる．

$$\mathbf{W}_{t+1}=\mathbf{W}_t - \mu_{SS}\mathbf{J'}_{SS}(\mathbf{W_t}) - \mu_{GC}\mathbf{J'}_{GC}(\mathbf{W_t}), \tag{9}$$

$$\mu_{SS}=\frac{\|\phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H]\|^2}{8\|\phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H]\tilde{\phi}(\mathbf{y})\mathbf{x}^H\|^2}$$

$$\mu_{GC}=\frac{\|\mathrm{diag}[\mathbf{WD}-\mathbf{I}]\|^{\mathbf{2}}}{8\|\mathrm{diag}[\mathbf{WD}-\mathbf{I}]\mathbf{D^H}\|^{\mathbf{2}}}.$$

3.2　動的環境を考慮した音声強調 HRLE

音声強調は，音源分離の後処理として用い，音源分離で分離しきれない拡散性雑音やチャネル間リークを抑制するサブトラクションベースの非線形な雑音抑圧法である．これまで，MMSE ベースのポストフィルタ [17] を用いてきたが，38 個のパラメータをチューニングする必要があった．また，初期雑音の推定に 1 秒程度の時間がかかっていたため，動的環境での実使用には不向きな面があった．そこで，*Histogram-based Recursive Level Estimation (HRLE)* を新たに提案した．この

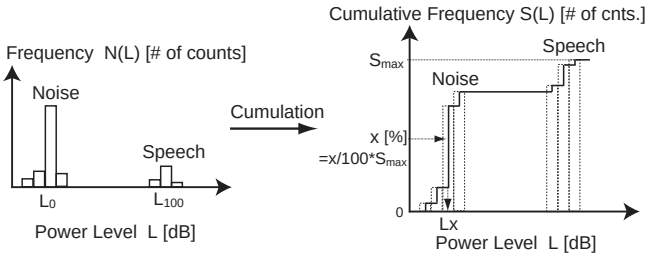


図 2　入力パワーレベルのヒストグラム (左) と累積ヒストグラム (右)
Fig. 2　Histogram (left) and Cumulative histogram (right) of input power level

手法は実質的にチューニングするパラメータ数は 2 つのみであるため，使い勝手が良い．HRLE は，騒音計などで用いられる環境雑音測定で用いられる手法を応用し，入力パワーレベルのヒストグラムを，図 2 に示すように L_x を用いて閾値処理することで入力雑音レベルを推定する．なお，L_x の x は，累積ヒストグラム上の位置を示し，例えば，L_0 は, 最小レベル, L_{100} は最大レベルを示す．

HRLE は再帰的に平均を計算するため，実時間で時変ヒストグラムが取得できる．このため，環境変化に対してスムースかつ高速な雑音レベル推定が可能である．具体的には，HRLE は以下のように定式化できる．

$$L_x(t)=L_{min} + L_{step} \cdot \underset{I}{\mathrm{argmin}} \left[xS(t, I_{max}) - S(t, I)\right], \tag{10}$$

$$S(t,i)=\sum_{k=0}^{i} \alpha N(t-1,k) + (1-\alpha)\delta(k - I_y(t)), \tag{11}$$

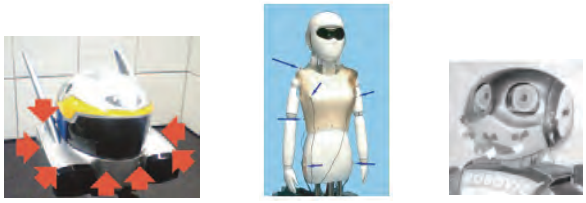
$$I_y(t)=\lfloor (20\log_{10}|y(t)| - L_{min})/L_{step} \rfloor, \tag{12}$$

ここで t は現在の時刻，$y(t)$ は，時間周波数領域での入力信号 L_{min}, L_{step}, I_{max} はそれぞれ，ヒストグラムの最小レベル, bin 一つ分のレベル幅，最大インデックスを示す．x は，累積頻度の位置 (0–1) α は，T_r とサンプリングレート F_s から $\alpha = 1 - 1/(T_r F_s)$ として求められる減衰パラメータである．$L_x(t)$ は，推定した雑音レベル, $\delta(t)$ は，ディラックのデルタ, $\lfloor \cdot \rfloor$ は，フロアを取る演算子である．

この手法ののパラメータ数は 5 で，そのうち 3 つはヒストグラムを規定するために用いる L_{min}, L_{step}, I_{max} であり，雑音レベル推定への影響は小さい．残りの 2 つのパラメータ x, α をチューニングするだけでよい．

4.　ロボット聴覚ソフトウェア HARK

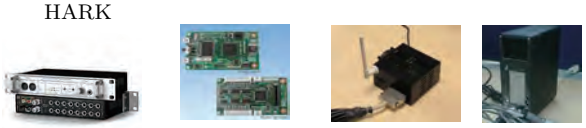
これまでのロボット研究成果の集大成として，ロボット聴覚ソフトウエア HARK (HRI-JP Audition for Robots with Kyoto Univ., hark は listen を意味する中世英語) を『聴覚の



HRP-2 (8ch)　SIG2 (whole body)　Robovie-R2

図 4　HARK で検証したロボットとマイクアレイの例

Fig. 4　Robot's Heads and Microphone arrays verified using HARK



ALSA-supported device (RME 社製など)　東京エレクトロンデバイス社製ボード　システムズインフロンティア社 RASPシリーズ

図 5　HARK で利用可能なマルチチャネル A/D

Fig. 5　multi-channel A/D device avalable for HARK

OpenCV』を目指すべく，音源定位，音源分離，及び，分離音声の音声認識といった機能を中心に 2008 年から研究用のオープンソースソフトウェアとして，無償公開を行っている [17]. 本稿で紹介した動的環境を考慮した機能を含めて，2010 年 11 月 25 日，HARK 1.0.0 として新版を正式に公開し(注1)，同日，HARK 1.0.0 に対応した講習会も開催した.

HARK は，音声認識部 (MFT-Julius) やサポートツールを除き，FlowDesigner [18] をミドルウェアとして用いている. FlowDesigner は，単一コンピュータ内の利用を前提とすることで，高速・軽量なモジュール統合を実現したデータフロー指向の GUI 開発環境を備えたフリー（LGPL/GPL）のミドルウェアである. FlowDesigner では，モジュール間接続は，各クラスの特定メソッドの呼び出し（関数コール）で実現されるため，オーバヘッドが小さい. データは，参照渡しやポインタで受け渡されるため，音響データのようなストリームデータの場合でも，高速にかつ少ないリソースで処理できる. つまり，FlowDesigner の利用によって，モジュール間のデータ通信速度とモジュール再利用性の両立が可能である. 我々は，メモリリーク等のバグ対処，操作性向上（主に属性設定部）を図った FlowDesigner も公開している.

マイクアレイの設置例を図 4 に示す. この例では，いずれも 8 チャネルのマイクアレイを搭載しているが，HARK では，任意のチャネル数のマイクアレイが利用可能である. また，HARK は ALSA ベースの A/D 装置，東京エレクトロンデバイス社製マルチチャネル A/D ボード，システムインフロンティア社製マルチチャネル A/D RASP シリーズといった多様なマルチチャネル A/D 装置をサポートしている（図 5 参照）. マイクは，安価なピンマイクで構わないが，ゲイン不足解消のため，プリアンプがあった方がよい. 東京エレクトロンデバイス社製ボードや RASP シリーズは，プリアンプおよび，プラグインパワー対応の電源供給機能を有しているので，使い勝手がよい. また，RASP シリーズに対応した MEMS マイクロホンも利用可能である. MEMS マイクロホンは，マイクロホン間の個体差が小さい，動作温度範囲が広い，比較的周波数応答がフラットであることなどから，ロボットへの適用には適しているといえる.

HARK を用いた典型的なロボット聴覚に対する FlowDesigner のネットワークを図 3 に示す. ファイル入力によりマルチチャネル音響信号を取得，音源定位・音源分離を行う. 得られた分離音から音響特徴量抽出，ミッシングフィーチャマスク (MFM) 生成を行い，これらを音声認識 (ASR) に送る. 本稿で紹介した GHDSS-AS および HRLE は Sound Source Separataion の中のモジュールとしてそれぞれ GHDSS, HRLE として表示されている.

5. 評　　価

評価実験として，移動話者の音声認識実験，および人ロボット対話シナリオを通じたタスクを紹介する.

5.1　移動話者の音声認識実験

音声認識実験では，2 本のスピーカを用いて，2 話者同時発話の孤立単語音声認識を行った. 音声認識の音響モデルは，JNAS を用いて，分離歪みを吸収できるようマルチコンディション学習を行って構築した. 使用した部屋は，4.0 m × 7.0 m の大きさで，残響時間は RT_{20} で 0.3–0.4 s 程度である. 一方のスピーカは (S_1) はロボットの正面の位置 ($\theta_1 = 0$) で ATR 音素バランス単語 216 語を 1〜2 秒おきに出力した. もう一方のスピーカ (S_2) は，発話毎に位置を変更し，疑似的な移動話者を生成した. 具体的には，$\theta_2 = -90°, -60°, -30°, 30°, 60°, 90°$ から任意の 1 つを選択し，その方向から発話を出力した. スピーカの出力ゲイン G は，$G = -6, -3, 0, 3, 6$ dB から任意に選択した. S_2 は，S_1 と異なる単語を，ほぼ S_1 とオーバラップするように出力した. S_1 と S_2 には，ATR 音素バランス単語セットの女性 (f1-f3) と男性 (m1-m3) 話者の組み合わせから 12 パターンを選択した. また，比較のため，マイクロホンアレイを用いず，1 本のマイク入力を選択して孤立単語認識を行った.

図 7 は，単語正解率を示す. シングルチャネルと比較し，GHDSS の利用により，10–20 ポイント程度向上していることがわかる. また，適応ステップサイズ制御を導入することにより，GHDSS の性能がさらに 20 ポイント以上向上していることがわかる. また，移動音源に対しても静止音源と同様に性能向上が見られることがわかる. 実際の話者を用いた静止話者と移動話者の同時発話認識の実験や，移動話者による数単語からなる短文発話の認識実験も統計言語モデルを用いて，別途行っているが，いずれも GHDSS-AS が最も良い性能を示している.

5.2　人ロボット対話タスク

実際に動的な環境での HARK を用いて構築したロボット聴覚システムの動作例として，1) ユーザが移動しながら行った挨拶への応答タスク，2) ロボット頭部回転中のユーザ質問応答タスク，3) 4 人の同時料理注文の聞き分けタスクを紹介する.

図 8，9 はそれぞれ タスク 1), 2) のスナップショットである. 図 8 では，ロボットは正しくユーザの発話を認識し，挨拶を行

(注1)：プレリリースは 2009 年 11 月に行った. http://winnie.kuis.kyoto-u.ac.jp/HARK/

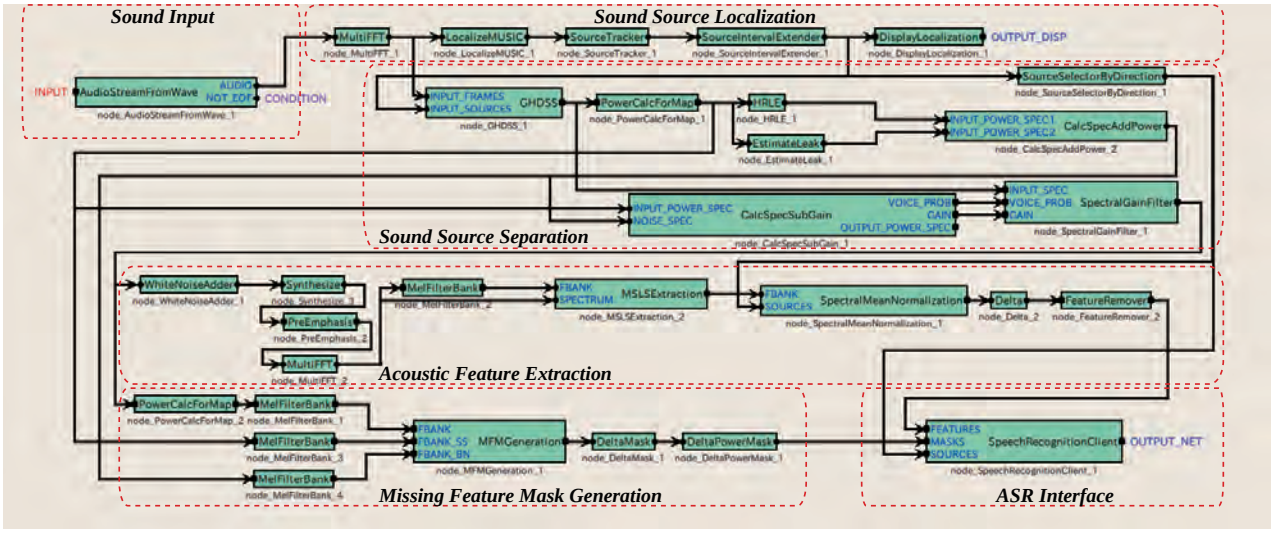


図 3　HARK に基づく実時間ロボット聴覚システム

Fig. 3　HARK-based real-time robot audition system

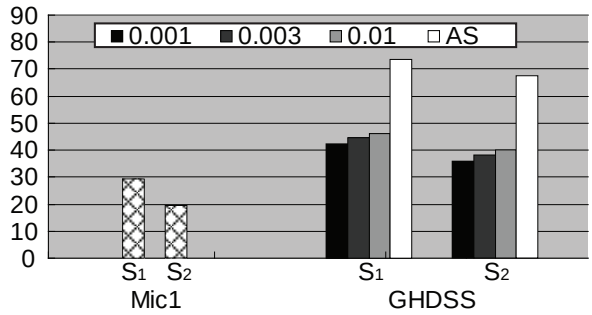


図 6　分離音声に対する単語正解率の向上: S_1 は静止音源, S_2 は疑似移動音源

図 7　Improvement in WCR of separated speech: S_1 is stationary, S_2 is pseudo moving.

うことができた．また，図 9 では，頭部にマイクロホンが搭載されているため，頭部回転中の発話は，マイクロホンアレイからは移動音源とみなすことができる．図は，ロボット頭部回転中でも，正しくユーザの質問を認識し，返答を行うことができることを示している．図中の左下の四角は，システムが検出した音響ストリームが表示されており，縦軸は水平角，横軸が時間を示している．ユーザの発話は，赤いカーブで表されており，これからも頭部回転中の発話は，ロボットからは移動音源として検出されることがわかる．図 10 は，4 人の同時料理注文の聞き分けタスクである．厳密には，このタスクは動的環境下のタスクではないが，動的環境を扱うために導入した手法により，システム全体の性能が向上したため可能になったタスクである．従来は，3 話者同時発話が精々であったが，本稿で紹介した手法の導入により 4 話者同時発話の認識が可能となった．

6. おわりに

本稿では，我々が取り組んでいるロボット聴覚研究について，マイクロホンアレイを用いた動的環境下の実時間音源分離法 GHDSS-AS や音声強調法 HRLE とその音声認識への適用について解説した．これらの技術は，ロボット聴覚ソフトウェア HARK としてオープンソースで公開を行っており，それらの有効性を，実際のロボットへの HARK の適用事例の紹介を通じて示した．本稿で紹介した技術や考え方は，高雑音下ハンズフリー音声認識へも適用可能であろう．こうした技術を用いて，実環境での使用に耐えうるシステムを構築するためには，ロバスト性が鍵になると考えている．これまでの経験から，単一の手法に固執するのではなく，様々な手法を有機的に統合/組み合わせることが，ロバスト性向上に有効であると考えている．HARK は，すでにこうした手法のいくつかがモジュールとして実装されており，統合システム用のベースラインソフトウェアとして有効であると考えている．HARK を様々な場面で研究のツールとして使っていただき，何らかの役に立てれば幸いである．

文　献

[1] K. Nakadai *et al.*, “Active audition for humanoid,” Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000), pp.832–839, AAAI, 2000.

[2] K. Nakadai *et al.*, “Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots,” Speech Communication, vol.44, no.1-4, pp.97–112, 2004.

[3] S. Yamamoto *et al.*, “Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech,” Proc. of the 2007 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-2007), pp.111–116, IEEE, Dec. 2007.

[4] S.F. Boll, “A spectral subtraction algorithm for suppression of acoustic noise in speech,” Proc. of 1979 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79), pp.200–203, IEEE, 1979.

[5] J.-M. Valin *et al.*, “Robust recognition of simultaneous speech by a mobile robot,” IEEE Transactions on Robotics, vol.23, no.4, pp.742–752, 2007.

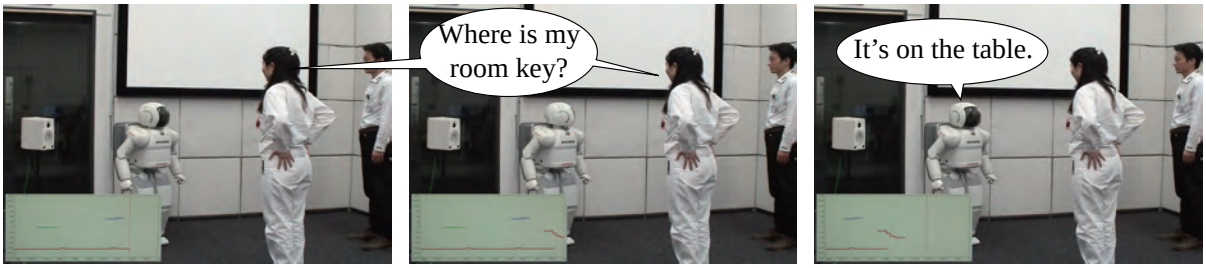
[6] I. Hara *et al.*, Hirukawa, and K. Yamamoto, “Robust speech interface based on audio and video information fusion for humanoid HRP-2,” Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), pp.2404–2410, IEEE, 2004.

[7] H. Saruwatari *et al.*, “Two-stage blind source separation based on ica and binary masking for real-time robot audition system,” Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), pp.209–

a) A user says a greeting to a robot. b) His greeting was made while in motion. c) The robot correctly responded.

図 8　移動話者の音声認識タスク

Fig. 8　Speech recognition of a moving speaker



a) A female starts asking a question while a robot is in motion.

b) Her question is regarded as a moving speech source.

c) The robot correctly answered the question.

図 9　ロボット頭部回転中の音声認識タスク

Fig. 9　Speech recognition while a robot is in motion

a）注文開始　b）同時発話（注文）　c）分離音認識（最右）　d）分離音認識（右中）

e）分離音認識（左中）　f）分離音認識（最左）　g）合計計算　h）注文受付終了

図 10　4 人の同時料理注文の聞き分けタスク

Fig. 10　Meal order taking for Four Simultaneous Speeches

214, IEEE, 2005.

[8] Y. Sasaki *et al.*, "Daily sound recognition using pitch-cluster-maps for mobile robot audition," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.2724–2729, 2009.

[9] H.-D. Kim *et al.*, "Human tracking system integrating sound and face localization using em algorithm in real environments," Advanced Robotics, vol.23, no.6, pp.629–653, 2007.

[10] K. Nakamura *et al.*, "Intelligent sound source localization for dynamic environments," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.664–669, 2009.

[11] 中臺一博 他, "移動型および静止型マイクロホンアレイ統合による複数移動音源追跡," 日本ロボット学会誌, vol.25, no.6, pp.979–989, 2007.

[12] K. Nakadai *et al.*, "Sound source separation and automatic speech recognition for moving sources," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.976–981, 2010.

[13] H. Sawada *et al.*, "Polar coordinate based nonlinear function for frequency-domain blind source separation," 2002 IEEE Int 'l. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2002), pp.1001–1004, 2002.

[14] D.H. Brandwood, "A complex gradient operator and its application in adaptive array theory," IEE Proc., vol.130, no.1, pp.251–276, 1983.

[15] S. Yamamoto and S. Kitayama, "An adaptive echo canceller with variable step gain method," Trans. of the IECE of Japan, vol.E65, no.1, pp.1–8, 1982.

[16] H. Nakajima *et al.*, "Blind source separation with prameter-free adaptive step-size method for robot audition," IEEE trans. ASLP, vol.18, no.6, pp.1476–1484, 2010.

[17] K. Nakadai *et al.*, "Design and implementation of robot audition system "HARK"," Advanced Robotics, vol.24, pp.739–761, 2010.

[18] C. Côté *et al.*, "Reusability tools for programming mobile robots," Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), pp.1820–1825, IEEE, 2004.