Vol.2015-MUS-107 No.33
2015/5/23

両耳聴ロボット聴覚ソフトウェアHARK-Binauralと Raspberry Pi 2を用いたヒューマノイドロボットへの適用

坂東 宜昭[1,a)]　金 宜鉉[1]　糸山 克寿[1]　吉井 和佳[1]　中臺 一博[2]　奥乃 博[3]

概要：ロボットが環境認識や人間とのコミュニケーションを行うには，聴覚や視覚機能が不可欠であり，ロボット聴覚やロボットビジョンの研究が広く行われている．我々はこれまで，ロボットに装着した複数個のマイクロホン (マイクロホンアレイ) を用いて音源定位や音源分離を行うロボット聴覚の研究を行い，その成果をオープンソース・ソフトウェア HARK として公開してきた．一般に多数のマイクロホンを用いるマイクロホンアレイは高い性能を発揮するが，比較的高価で複雑な音響システムを要す，計算量が比較的大きいといった問題がある．本稿では，現在公開中の 2 つのマイクロホンのみを用いる両耳聴ロボット聴覚ソフトウェア HARK-Binaural の機能を述べる．両耳聴ロボット聴覚は，ステレオ入録の A/D 変換器で構成でき，計算量が比較的少ないことが特徴で，組み込みシステム上でのリアルタイム動作が可能である．実際に Raspberry Pi 2 上に本ソフトウェアを用いて実装したヒューマノイドロボットの適用例を示す．

1. はじめに

ロボットが環境認識や人間とのコミュニケーションを行うには，聴覚や視覚機能が不可欠であり，ロボット聴覚やロボットビジョンの研究が広く行われている [1,2]．これらの手法の一部はオープンソース・ソフトウェアとして公開され，その研究分野に詳しくない人でも容易にロボット・システムの構築が可能となっている [2–5]．

我々は，これまでロボットに装着した複数個のマイクロホン (マイクロホンアレイ) を用いて音環境認識を行うロボット聴覚の研究を進め，その成果をオープンソース・ソフトウェア HARK (Honda Research Institute Japan audition for robots with Kyoto University) として公開してきた [*1]．本ソフトウェアでは，信号処理の初心者でも容易にロボット聴覚機能を構成できるように，GUI インターフェースを用いたプログラミング (図 1) を提供している．また，Python やロボット用ミドルウェア ROS [5] へのインターフェースも用意しており，幅広いユーザのニーズへの対応を目指している．

本稿では，HARK のプラグインの一つである両耳聴ロボット聴覚パッケージ HARK-Binaural の機能と，その応用例を述べる．一般に多数のマイクロホンを用いるマイクロホンアレイは高い性能を発揮するが，比較的高価で複雑な音響システムを要す，計算量が比較的大きいといった問題がある [6]．両耳聴ロボット聴覚では，ステレオ入録の A/D 変換器で構成可能でき，計算量が比較的少ないことが特徴で，組み込みシステム上でのリアルタイム動作が可能である．以降では，関連研究および HARK について概観し，HARK-Binaural について述べる．さらに，組み込みプラットフォームの一つである Raspberry Pi 2 上で本ソフトウェアが動作することを示し，教育・ホビー向けヒューマノイドロボット Rapiro への応用例を述べる．

2. 関連研究

HARK と同様にロボットの聴覚を構成するためのソフトウェア rospeex および，両耳聴ロボット聴覚ソフトウェアを公開しているプロジェクト Two!Ears の概要を述べる．

rospeex [4] はクラウド型音声コミュニケーションツールキットとして開発されている．音響モデルや言語モデルなどの大規模な資源をロボット上に搭載する必要がなく，ハードウェアを簡略化できる．本システムは，ROS 上で動作し，1) 雑音抑圧・発話区間検出，2) 音声認識，3) 音声合成の 3 つの機能が提供されている．標準の音声認識・合成エンジンだけでなく他のエンジンとの接続インターフェースが実装されており，Google 音声認識・合成の API が標準で提供されている．また，rospeex で提供される機能は ROS を通じて HARK とも接続でき，HARK の音源定位・

[1] 京都大学, Kyoto Univ., Sakyo, Kyoto, 606–8501, Japan
[2] ホンダ・リサーチ・インスティチュート・ジャパン, 東京工業大学, HRI-JP, Wako, Saitama, 351–0188, Japan
[3] 早稲田大学, Waseda Univ., Shinjuku, Tokyo, 169–0072, Japan
[a)] yoshiaki@kuis.kyoto-u.ac.jp
[*1] http://www.hark.jp/

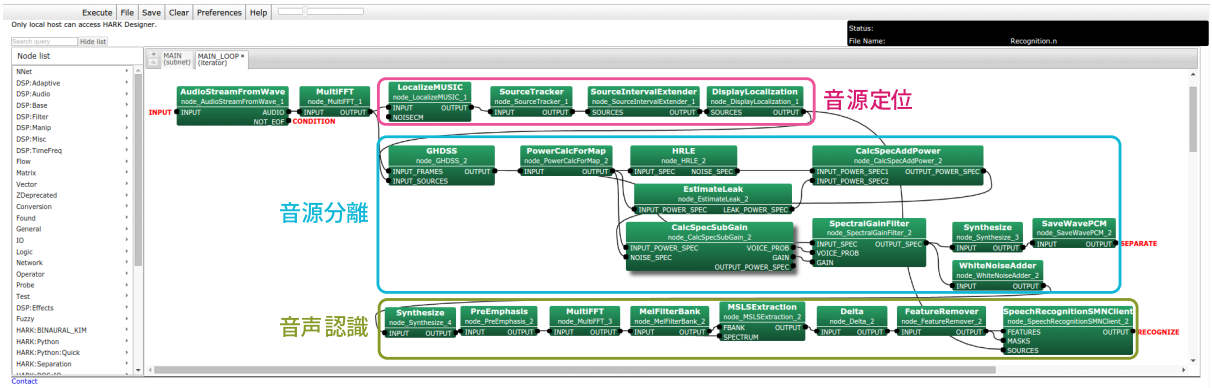

図 1: HARK のプログラミング例．信号処理の各機能を持つノードをつなぎ合わせてプログラムを構成する．

分離結果を利用できる．

Two!Ears [*2] は，両耳聴ロボット聴覚研究プロジェクトの一つで，プロジェクトの成果として両耳聴ロボット聴覚モデルやシミュレータを公開している．また，物理実験と音響測定結果のデータベースを公開しており，無響室を含む屋内での両耳聴のインパルス応答セットや環境音セット等を提供している．

3. HARK

HARK は，「音を聞き分ける」という立場でロボット聴覚機能を実現する．HARK 全体の説明については文献 [3] が詳しいので，以下では HARK の概要についてのみ述べる．本システムでは，ロボット聴覚の基本機能として音源定位と音源分離および音声認識の機能を提供している．環境には複数の音源が存在すると考え，個別に音源を分離・認識することで，雑音存在下や複数話者が同時に会話する環境でも頑健な動作を実現する [1].

3.1 特徴

HARK は，1) 机上だけでなく実ロボットで利用可能，2) 信号処理や音声処理に十分な知識がない人でも出来るだけ簡単に利用可能の 2 点を重視して開発を行っている [3].

実ロボットではリアルタイム動作が要求される．HARK では，各モジュールを共有ライブラリとして実装しオーバーヘッドをできるだけ小さくするフレームワークである Flowdesiner [7] に含まれる batchflow で開発され，リアルタイム動作を実現している．また，様々な形状のロボットに対応するために，マイクロホンの本数やレイアウトに応じてキャリブレーションを行うツールキット wios および harktool を公開している．Microsoft Kinect や Playstation Eye といった市販のマイクロホンアレイについては事前キャリブレーションデータも用意している．

信号処理の初心者でも容易に HARK を使用できるように，図 1 に示すような GUI インターフェース (HARK Desiner) を用いたプログラミングを採用している．また，linux では apt によるインストールを，windows も専用のインストーラを配布しており，簡単にインストールできるようになっている．さらに，ほぼ毎年行っているソフトウェアのアップデートに合わせて，無料の講習会を開催している．

3.2 基本機能

音源定位 MUSIC (MUltiple SIgnal Classification) 法 [8] による音源定位を提供している．MUSIC 法は固有値分解に基づく手法で，一般にビームフォーミングに基づく手法より空間解像度が高いという特徴がある．実際のロボットでは，ロボット自身等から発生する強い雑音によって，定位性能が劣化することがあるが，HARK では一般固有値展開や一般特異値分解に基づく GEVD-MUSIC および GSVD-MUSIC 法 [9] に基づく，雑音による性能劣化の軽減を提供する．

音源分離 GHDSS-AS (Geometric High-order Decorrelation Source Separation with Adaptive Stepsize) 法 [10] による音源分離を提供している．GHDSS-AS は，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手法である．一般にビームフォーミングはモデル誤差による性能劣化が発生しやすく，ブラインド分離では一意な分離が困難となるパーミュテーション問題が発生する．GHDSS-AS では双方の欠点を補完するように設計されている．さらに，移動音源に対応するために適応ステップサイズ (Adaptive Stepsize) による拡張を行っている．

音声認識 音源分離や音声強調を行った後の歪んだ音声に対して頑健な音声認識を行うために，分離音の各時間・周波数ビンに対して信頼度を付与する MFT (Missing Future Theory) [11] を用いた音声認識を提供している．音声認識でよく用いられる特徴量である MFCC の他に，分離歪みの影響を低減できる MSLS 特徴量を提供する．

この他にも音楽情報処理を行うパッケージ HARK-MUSIC，Kinect や OpenCV を使用するための HARK-

[*2] http://twoears.aipa.tu-berlin.de/

表 1: Hark-Binaural に実装されているノード一覧

	ノード名	機能
信号処理	BinauralMultisourceLocalization	音源定位
	BinauralMultisourceTracker	音源追跡
	SpeechEnhanvement	音声強調
	VoiceActivityDetection	音声区間検出
	SourceSeparation	音源分離
可視化	SSLVisualization	定位結果可視化
	SpectrumVisualization	スペクトログラム表示
	VADVisualization	音声区間可視化
	WaveVisualization	波形表示

Kinect, HARK-OpenCV などを提供している．

4. HARK-Binaural パッケージ

HARK-Binaural パッケージでは，2 つのマイクロホンのみで信号処理を行う機能が提供されている．表 1 に HARK-Binaural で実装されているノードを示す．

4.1 特徴

HARK-Binaural の特徴は，1) 計算量が少ない・聴覚システムの簡素化が可能であること 2) 他の HARK ノードと相互利用可能という点である．1) HARK の標準パッケージに含まれる音源定位法 MUSIC は，空間解像度や雑音頑健性など実環境での高い性能を発揮する [9] が，比較的高価な多チャンネル A/D 変換器や豊富な計算資源を必要とする．両耳聴信号処理ではステレオ A/D 変換器でよく，市販のステレオ・オーディオレコーダ等の比較的安価な A/D 変換器でも聴覚システムを構築できる．さらに本システムの実装は，線形代数ライブラリ Armadillo に基づいており，計算の高速化が図られている．2) また，HARK-Binaural は HARK の標準パッケージと型互換性があるので相互利用が可能となっており，これまでの HARK のソフトウェア資源を活用することができる．例えば，BinauralMultisourceTracker での定位結果を HARK-SSS パッケージの Beamforming による音源分離 (音声強調) と組み合わせることが可能である．

4.2 基本機能

BinauralMultisourceLocalization ノード 対雑音・残響性に優れる GCC-PHAT [12] に基づく両耳聴での音源定位機能を提供する．HARK の標準パッケージに含まれる MUSIC 法では，固有値空間に最低 1 つ以上の雑音空間が必要なために，両耳聴処理では最大で 1 つの音源しか定位できない．本ノードでは，各時間周波数ビンを占める音源は 1 つであると仮定し，各ビンごとの到達時間差を dynamic K-mans 法でクラスタリングすることで，複数音源の定位に対応する [6]．

BinauralMultisourceTracker ノード 音源追跡機能を提供する．BinauralMultisourceLocalization ノードの結果は，各時刻独立にどの方向に音源が存在するかの情報のみなので，本ノードにより連続する 1 発話として各音源の追跡結果を出力する．

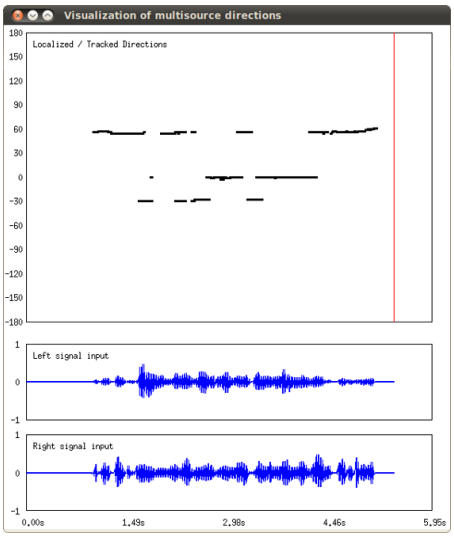

図 2: SSLVisualization ノードによる可視化例．3 話者が同時に発話している音声に対する定位結果．

VoiceActivityDetection ノード マイクロホンからの入力音響信号に対して音声区間検出機能を提供する．低パワーの雑音下での検出精度が高い，ガウス分布に従うノイズ成分を推定する手法 [13] に基づいている．

各種可視化ノード HARK-Binaural パッケージでは，各ノードの処理結果を可視化するノード群 (表 1 下段) を提供している．図 2 に，-30, 0, 60 度に音源がある場合の BinauralMultisourceLocalization ノードの定位結果の SSLVisualization ノードによる可視化例を示す．

5. Raspberry Pi 2 を用いた応用例

Raspberry Pi 2 を搭載した教育・ホビー用ヒューマノイドロボット Rapiro (図 3) への，HARK-Binaural の応用例を示す．Raspberry Pi 2 は 900MHz 動作の ARM Cortex-A7 を 4 コア搭載する教育用シングルボードコンピュータで，Debian ベースの Raspbian が動作し，HARK も Raspbian 上でコンパイルすることで動作する．HARK 標準パッケージの MUSIC 法は，本システムでのリアルタイム動作は困難だが，HARK-Binaural パッケージの BinauralMultisourceLocalization は計算量が小さくリアルタイム動作が可能である．実際に，60 秒のステレオ録音を Raspberry Pi 2 上で定位した場合の計算時間は 17.6 秒となり十分高速である．本システムを構成するために表 2 のノードを使用したネットワーク (プログラム) を構築した．動作の様子を示したデモ動画を `http://winnie.kuis.kyoto-u.ac.jp/members/yoshiaki/demo/sigmus107/` にアップロードした．

図 3: HARK-Binaural の応用例として使用したヒューマノイド．頭部量側面にマイクロホンを装着した．

表 2: 応用例で使用した主要ノード

パッケージ名	ノード名
HARK 標準パッケージ	AudioStreamFromMic
	MultiFFT
	SourceTracker
	SourceIntervalExtender
HARK-Binaural	BinauralMultisourceLocalization
	BinauralMultisourceTracker
HARK-ROS	HarkMsgsStreamFromRos
	HarkMsgsSubscriber
	RosHarkMsgsPuclisher
HARK-SSS	Beamforming

5.1 ハードウェア構成

図 4 に本システムのハードウェア構成を示す．本ロボットには，カメラ，2-ch マイクロホンアレイおよびスピーカが装着されている．図 3 に示す 2 つのマイクロホンは，市販の USB ステレオ・オーディオキャプチャを通して，Raspberry Pi 2 へ接続されている．距離センサは地面との距離を測り，落下防止に使用される．

5.2 ソフトウェア構成

図 5 に本システムのソフトウェア構成を示す．限られた計算資源を有効活用するために，出来るだけ並列計算を行うことと，ローカルとクラウドで計算すべき処理の切り分けを原則として設計した．

並列処理と既存ソフトウェア資源の有効活用のために ROS を使用した．音源定位だけでなく音声強調も行う場合，全ての処理を単一コアで実時間動作することは困難である．HARK 単体では，複数コアを使用した並列処理に対応していないが，HARK-ROS パッケージを用いて，ROS ノードとして機能を分割することで並列処理が可能となる．各ノードはプロセスが独立のため，特定のノードが異常終了してもシステム全体を停止することなく復帰できる．また，CCD カメラのように ROS の既存のソフトウェア資源を活用できる．HARK-ROS を用いて作成したノードは，1) 録音ノード，2) 音源定位ノード，および 3) 音源分離 (強調) ノードの 3 つである．

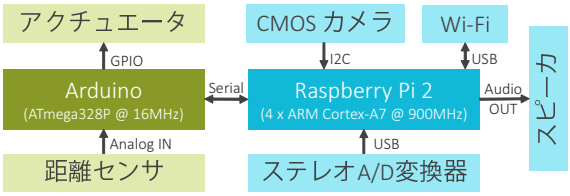

図 4: ハードウェア構成

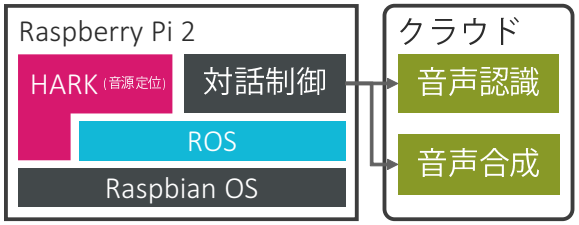

図 5: ソフトウェア構成

音声認識と音声合成は，クラウドサービスを使用した．音源定位や音声強調に比べて，音声認識・合成は計算量が大きい．特にこれらは，発話が発生した時のみの処理でよいので，音響信号のクラウドへの転送コストも小さい．逆に常に計算が必要で比較的計算量の小さい音源定位や強調はローカルで実行する．

6. おわりに

本稿では，両耳聴ロボット聴覚ソフトウェア HARK-Binaural および，その Raspberry Pi 2 への応用例を述べた．両耳聴ロボット聴覚では，ステレオ入録の A/D 変換器で構成可能でき，計算量が比較的少ないことが特徴で，HARK-Binaural ではさらに線形代数ライブラリ Armadillo を使用し，組み込みシステム上でのリアルタイム動作が可能である．実際に Raspberry Pi 2 で録音時間の半分以下の時間で音源定位が動作することを確認した．また，限られた計算資源を有効活用するために，並列計算とクラウド処理を組み合わせた応用例を述べた．

謝辞 本研究の一部は科研費基盤 (S) No.24220006 の支援を受けた．

参考文献

[1] Rosenthal, D. F. et al.: *Computational Auditory Scene Analysis*, Lawrence Erlbaum (1998).
[2] Bradski, G. et al.: *Learning OpenCV: Computer vision with the OpenCV library*, " O'Reilly Media, Inc." (2008).
[3] 中臺一博ら：ロボット聴覚オープンソースソフトウェア HARK の紹介，第 15 回計測自動制御学会 システムインテグレーション部門講演会，pp. 1712–1716 (2014).
[4] 杉浦孔明ら：rospeex：クラウド型音声コミュニケーションを実現する ROS 向けツールキット，信学技報 (CNR2013-10)， Vol. 113, pp. 7–10 (2013).

[5] Quigley, M. et al.: ROS: an open-source Robot Operating System, *ICRA workshop on open source software*, Vol. 3, No. 3.2, p. 5 (2009).
[6] Kim, U.-H. et al.: Improved binaural sound localization and tracking for unknown time-varying number of speakers, *Advanced Robotics*, Vol. 27, No. 15, pp. 1161–1173 (2013).
[7] Cote, C. et al.: Code reusability tools for programming mobile robots, *Proc. of IEEE/RSJ IROS 2004*, Vol. 2, pp. 1820–1825 vol.2 (2004).
[8] Asano, F. et al.: Real-time sound source localization and separation system and its application to automatic speech recognition., *Proc. of Interspeech 2001*, pp. 1013–1016 (2001).
[9] Nakamura, K. et al.: Intelligent sound source localization and its application to multimodal human tracking, *IEEE/RSJ IROS 2011*, pp. 143–148 (2011).
[10] Nakajima, H. et al.: Blind Source Separation with Parameter-Free Adaptive Step-Size Method for Robot Audition, *IEEE TASLP*, Vol. 18, No. 6, pp. 1476–1485 (2010).
[11] Okuno, H. G. et al.: Robot Audition: Missing Feature Theory Approach and Active Audition, *Robotics Research*, Springer, pp. 227–244 (2011).
[12] Knapp, C. et al.: The generalized correlation method for estimation of time delay, *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)*, Vol. 24, No. 4, pp. 320–327 (1976).
[13] Sohn, J. et al.: A statistical model-based voice activity detection, *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1–3 (1999).