ロボット聴覚オープンソースソフトウェアHARKの概要と評価

○中臺 一博 †,‡　奥乃 博 *　中島 弘史 †　長谷川 雄二 †　辻野 広司 †
†(株) 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　* 京都大学

Overview of Open Source Software for Robot Audition and Its Evaluation

Kazuhiro Nakadai†,‡,　Hiroshi G. Okuno*,
Hirofumi Nakajima†,　Yuji Hasegawa†,　Hiroshi Tsujino†
† Honda Research Institute Japan Co., Ltd.　‡Tokyo Insititute of Technology　* Kyoto University

Abstract— Recently, we have released an open source software system for robot audition called HARK (Honda Research Institute Japan Audition for Robots with Kyoto University). HARK consists of a lot of modules including multi-channel audio input, sound source localization, sound source tracking, sound source separation and recognition of separated speech for robot audition based on the data-flow oriented software programming environment, FlowDesigner. By combining these modules, a user can easily build a robot audition system for various types of robots and acoustic environments. This paper describes overview of HARK, and also shows its performance by using Honda ASIMO in detail.

Key Words: robot audition, software architecture, open source, sound source localization, sound source separation, automatic speech recognition

1. はじめに

我々は，より自然な人・ロボットインタラクションを実現するために，ロボット自身に装着されたマイクロホンを用いて，音を聞き分ける「ロボット聴覚」という研究分野を提唱し研究を進めている [1]．これまで，ロボット搭載型マイクロホンアレイを用いた音源定位，音源分離，音声認識をミッシングフィーチャ理論に基づき統合し，３話者が同時に発話する状況下でも，それぞれの話者の音声を認識するロボットを報告した [2]．さらに，さらなるロボット聴覚研究領域の発展やロボット聴覚システムユーザの開拓を目指して，研究用途のオープンソースソフトとしてロボット聴覚システム HARK (Honda Research Institute Japan Audition for Robots with Kyoto University)[1] を公開した [3]．本稿では，HARK の概要について述べるとともに，その性能について詳細な評価を行う．

2. HARK の概要

HARK は，データフロー指向ソフトウェアプログラミング環境 FlowDesigner [4] 上のモジュール群として実現されている．HARK は，多くの多チャンネル音響信号入力デバイスをサポートしており，音源定位，追跡，分離や分離音の認識といったロボット聴覚に必要とされるモジュールが一通り用意されている．これらを使って，GUI ベースのプログラミングだけで自由にロボット聴覚システムの構築やカスタマイズがが可能である．さらに，同等に FlowDesigner 上のモジュール群として公開されている移動ロボット用のロボット聴覚ソフトウェア ManyEars[5][2] と相互互換性を持っている．ManyEars は，音声認識こそ提供していないが，HARK とは異なるアルゴリズムに基づく音源定位，追跡，分離モジュールを提供している．つまり，HARK と ManyEars を組み合わせれば，ユーザは多様なロボット聴覚システムを構築することができる．

2·1 HARK の実装

HARK を構成するモジュールは，大きく多チャンネル音響信号入力，音源定位・追跡，音源分離，音響特徴量抽出，ミッシングフィーチャマスク (MFM) 生成，音声認識インタフェースの 6 つのカテゴリに分類できる．ミッシングフィーチャ理論ベースの音声認識 (MFT-ASR) は，現時点では，FlowDesigner のモジュールとしては実装されていないが，音声認識インタフェースを用いて，FlowDesigner と TCP/IP ベースで接続ができるようになっている．Table 1 に，モジュールの種類をまとめた．

多チャンネル音響信号入力は，WAVE フォーマットのファイルから音響信号の読み込み (AudioStream-FromWave)，サウンドカードなどの音響信号入力デバイスから音響信号の読み込み (AudioStreamFromMic) をサポートしている．入力デバイスは，現時点では，ALSA (Advanced Linux Sound Architecture) がサポートするサウンドカード [6]，RASP シリーズ [7]，TD-BD-8CSUSB[8] の 3 種類のデバイスが利用可能である．

音源定位・追跡用に 6 つのモジュールが用意されている．LocalizeMUSIC は，MUltiple Signal Classification(MUSIC)[9] と呼ばれる適応ビームフォーミングアルゴリズムを用いた音源定位を提供し，実環境でも安定して複数の音源を定位することができる．ConstantLocalization は，常に一定値を出力する音源定位モジュールであり，主に評価実験に用いる．また，ManyEars では LocalizeBeam と呼ばれる steered ビームフォーマを提供している．これらの 3 つのモジュールは互いに置換可能である．SourceTracker は，定位結果から音源追跡を行い，ID 付きの音源方向

[1] http://winnie.kuis.kyoto-u.ac.jp/HARK/
[2] http://manyears.sourceforge.net/

Table 1 Modules provided by HARK

Class Name	Module Name
多チャンネル音響信号入力	AudioStreamFromMic AudioStreamFromWave
音源定位・追跡	LocalizeMUSIC ConstantLocalization SourceTracker DisplayLocalization SaveSourceLocation
音源分離	DSBeamformer Separation (ManyEars のパッチとして) Synthesize
音響特徴量抽出	Delta MelFilterBank FeatureExtraction SpectralMeanNormalization SaveFeatures
MFM生成	DeltaMask MFMGeneration StationaryNoiseEstimator
音声認識インタフェース	SpeechRecognitionClient
MFT 音声認識	Multiband Julius/Julian (non-FlowDesigner module)

を出力する．DisplayLocalization は，LocalizeMUSIC, ConstantLocalization, SourceTracker の出力を可視化するモジュールである．SaveSourceLocation は各フレームの定位結果をテキストファイルに格納する．

音源分離用に，DSBeamformer, Separation，Synthesize という 3 つのモジュールを提供している．DSBeamformer は，遅延和型ビームフォーマを用いた音源分離であり，性能は高くないものの，制御が容易で，環境雑音に対するロバスト性が高い．Separation は，GSS (Geometric Source Separation) [2] を用いたより高精度な音源分離モジュールである (ManyEars の SeparGSS に対するパッチとして提供される)．Synthesize は， DSBeamformer や Separation から得られる分離スペクトルを時間領域波形に変換する．

音響特徴量抽出用に MelFilterBank，FeatureExtraction, SpectralMeanNormalization, Delta, SaveFeatures の 5 つのモジュールを用意した．HARK では，分離歪みや雑音に対する頑健性が高い *Mel-Scale Log Spectrum* (MSLS) [10] を音響特徴量として用いる．MelFilterBank はスペクトルの周波数軸をメルスケールに変換する．FeatureExtraction は，メルスケールスペクトルから，MSLS 特徴量を出力する．SpectralMeanNormalization は，MSLS 特徴量の対雑音ロバスト性を向上させるため，正規化を行う．Delta は，MSLS 特徴量から一次線形回帰を用いて，Δ 特徴量を算出する．SaveFeatures は，発話ごとに，MSLS 特徴量の時系列データをファイルに格納する．

ミッシングフィーチャ理論 (MFT) は，音声認識の対雑音ロバスト性を向上させるアプローチである [11, 12]．MFT は，音声認識時に信頼できない特徴量をマスクし，信頼できる音響特徴量のみを用いて認識を行う．このマスクはミッシングフィーチャマスク (MFM) と呼ばれる時間-周波数領域のマップである．MFM 生成用に 3 つのモジュールが用意されている．MFMGeneration は，メルスケールの分離音スペクトル，ポストフィルタ後のスペクトル，StationaryNoiseEstimator から得られる暗騒音を用いて MFM を生成する．DeltaMask は，MFM を入力としてそのデルタ項を出力する．

SpeechRecognitionClient は，音響特徴量，MFM，音源方向を TCP/IP 経由で Multiband Julius/Julian に送出する．Multiband Julius/Julian [13] は，日本語/英語用オープンソース音声認識システム Julius/Julian[14] を MFT をサポートするように改良したモジュールである．

3. HARK の評価

下記の 4 点の実験を行い，HARK を詳細に評価した．

A: ３種類の異なる雑音下での雑音レベルによる音声認識性能の変化
B: 音源分離・音声認識パラメータの違いによる三話者同時発話認識性能の変化
C: 音源定位手法による音声認識性能の違い
D: 処理時間評価

いずれの実験も図 2 に示すロボット聴覚システムを HARK を用いて構築して用いた．A と D は ASIMO, B,C では SIG2 を用いた．両者は，図 1 に示すように，マイク配置が大きく異なる．HARK では，このように条件が異なるロボットでもマイクの位置を記述したファイルを修正するだけで図 2 に示した構成を変更せずに対応が可能である．なお，ASIMO は，$7\,m \times 4\,m$ の部屋 (残響時間 0.2–0.3 秒) の中央に，SIG2 は，4 m × 5 m の部屋 (残響時間 0.3–0.4 秒) の中央に設置した．

3·1 雑音レベルに対する音声認識性能

認識語彙には，ATR 音素バランス単語 216 語を用いた．この語彙セット 6 名分のデータを用いて，その平均単語誤り率を算出した．音響モデルには，日本語新聞記事読上げコーパス (JNAS) で学習したトライフォン HMM を利用した．3 種類の雑音環境を用いた．1 つ目は，ASIMO のファンやモータ音が含まれるロボット雑音下の実験である．雑音はロボット内部から発生するため，マイクロホンアレイにとっては，拡散性の雑音といえる．2 つ目は，さらに ASIMO の左前方 60 度，1 m の位置にスピーカを設置し，音声雑音源を追加した．拡散性雑音と方向性雑音が混在している雑音環境といえる．3 つ目は，2 つ目と音源が音楽であること以外は同様である．認識用音声データは，実測ではなく，ASIMO 正面 1 m に認識用音声を出力するためのスピーカを設置して測定したインパルス応答と各雑音信号を用いて合成した．データ合成では，S/N を -8 dB から 3 dB まで変化させた．この実験では，マイク入力ではなく音声ファイルを使用したが，マイク入力モジュール AudioStreamFromMic をファイル入力モジュール AudioStreamFromWave に入れ替えるだけで対応できる．

Fig. 3 に実験結果を示す．横軸は S/N， 縦軸は，単語誤り率を示している．ベースライン性能を得るために，1 つ目の雑音環境下で，音声音源に一番近い正面のマイクロホンからの入力を前処理を行わずに Julian で認識した．同じ雑音環境で，ロボット聴覚システムを用いると，性能が大きく向上することがわかる．ロボット雑音に加え，方向性雑音源が存在する場合，ある程度性能劣化が見られるものの，-4 dB でも単語誤り

a) ASIMO

b) *SIG2*

Fig.1 8 ch microphone arrays

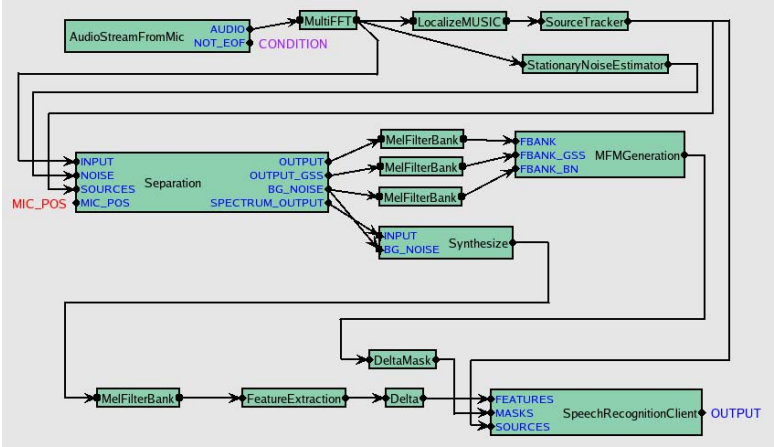

Fig.2 An example of a robot audition system using HARK

率は 20 % 程度と良好な性能を見せている．これらの結果から，HARK ベースのロボット聴覚システムの対雑音ロバスト性，HARK を用いたシステム構成変更の容易さがわかる．

3·2　音源分離・音声認識パラメータと音声認識性能

音源分離・音声認識パラメータを変更して，三話者同時発話の孤立単語認識性能の違いを調べた．音源にはスピーカを用い，ロボットとの距離を 2 m とした．1 本のスピーカは正面に，他の 2 本は対称に ±10° から ±90° まで 10° 毎に変化させた．評価音声データには，ATR 音素バランス単語 216 語を用い，正面からは男性 (m101)，左から女性 (f101), 右から男性 (m102) の音声をランダムに同期させて，ほぼ同じパワーで出力した．

実験の条件は下記の通りである．こうした条件設定も当該モジュールのパラメータを変更することによって簡単に設定することができる．

(1) 正面マイクを選び，クリーン音響モデルを用いて認識．
(2) GSS のみを前処理に用い，クリーン音響モデルを用いて認識．
(3) GSS のポストフィルタ機能を ON にして，クリーン音響モデルを用いて認識．
(4) GSS のポストフィルタ機能を ON にして，マルチコンディション音響モデルを用いて認識．
(5) GSS のポストフィルタ機能を ON にして，クリーン音響モデルを用い，MFT 機能を ON にして認識．
(6) GSS のポストフィルタ機能を ON にして，白色雑音適応音響モデルを用い，MFT 機能を ON にして認識．
(7) GSS のポストフィルタ機能を ON にして，マルチコンディション音響モデルを用い，MFT 機能を ON にして認識．
(8) 基本的に (vi) と同じであるが，MFM として正解マスクを用いた (性能の上限値を知るための実験)

クリーン音響モデルは，評価用のデータを除く ATR 音素バランス単語セット 22 名分で作成した．つまり，話者オープン，語彙クローズドの実験である．マルチコンディション音響モデルは，クリーン音響モデルと同じデータセットから三話者同時発話データを作成し，これを構築したシステムで分離したデータを用いて学習した．白色雑音適応音響モデルは，クリーン音響モデルで用いたデータに白色雑音を重畳して学習した．図 4 にそれぞれの場合の，正面話者の認識率を示す．

図 4 から，マイクロホンアレイによる音源分離，MFT が有効に働いていることがわかる．特に，GSS をポストフィルタ機能付きで用い，MFT を ON にしてマルチコンディション音響モデルで認識した場合，性能が最もよくなっている．白色雑音適応音響モデルもほぼ同等の性能を示しており，音響モデル作成のコストを考慮すると，条件 (6) が最も実用的といえよう．スピーカ間隔が 10 度の場合は，性能があまり上がっていない．しかし，条件 (8) の結果から，マスクがうまく推定できれば，向上の余地があることがわかる．

3·3　音源定位手法と音声認識性能

音源定位手法の変更による音声認識性能の違いを調べた．用いた音源定位は，定位結果所与，MUSIC, steered ビームフォーマの 3 種類である．これらも，音源定位モジュールをそれぞれ，ConstantLocalization, LocalizeMUSIC, LocalizeBeam に変更するだけで設定可能である．音源分離や音声認識のパラメータは，前節の条件 (5),(6) を用いた．

単語正解精度を Table 2 に示す．クリーン音響モデルを用いると, MUSIC はビームフォーマより性能が高いが，白色雑音適応モデルでは，性能差があまり見られない．結果が所与の場合，白色雑音適応モデルによる効果はあまり見られないが，MUSIC やビームフォーマを用いた場合は大きな性能向上が見られる．定位性能の低さは，発話区間検出ミスの増加につながるため，クリーン音響モデルを用いた場合の性能低下を招いていると考えられる．白色雑音適応音響モデルでは，重畳したレベル程度の雑音であれば音源定位誤りによる発話区間検出ミスを吸収でき，認識性能の低下が抑えられているためと考えられる．

3·4　処理時間評価

処理時間の評価を行った．まず，Pentium4 2.4 GHz の PC で 800 秒の音声データを処理したところ，499 秒で音声認識まで終了した．また，音声入力から出力

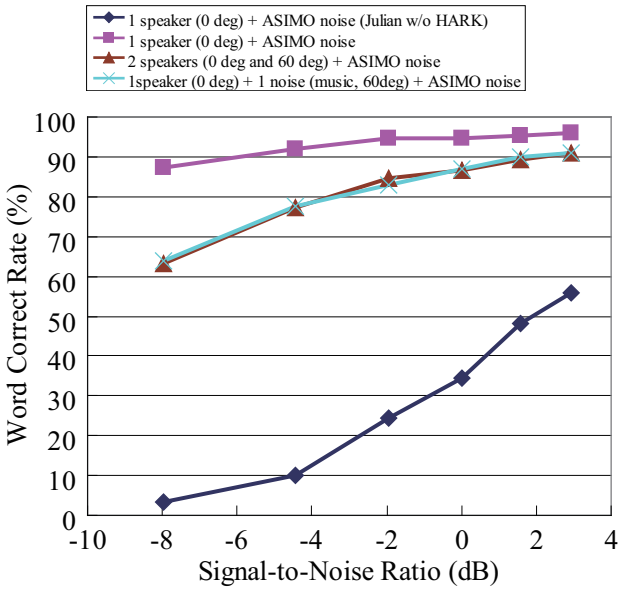

Fig.3 Word error rates of speech recognition

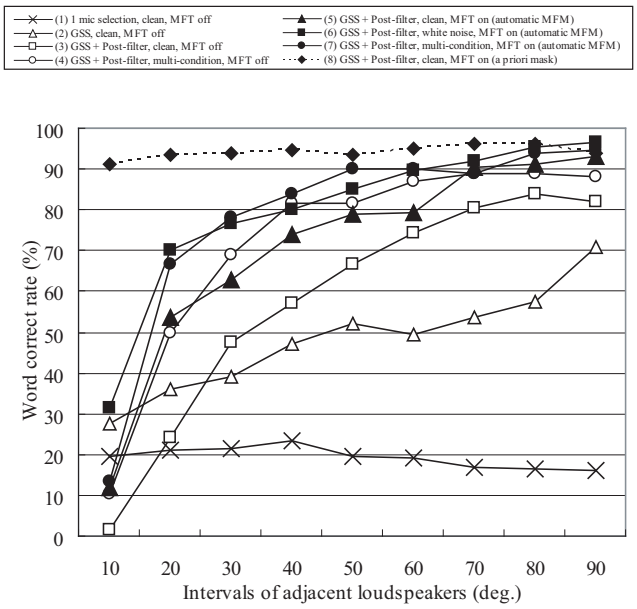

Fig.4 Recognition results with our proposed system (the front speaker)

Table 2 Word correct rate (WCR in %) of the center speaker according to each localization method

Acoustic model	*White noise addition*			*Clean model*		
Interval	30°	60°	90°	30°	60°	90°
given	90.0	88.5	91.0	85.0	84.5	87.0
steered BF	82.3	90.5	89.0	65.5	70.6	72.4
MUSIC	86.0	83.3	86.7	57.0	74.0	64.5

Table 3 CPU Occupacy of Each Process

	localization	separation	feature extraction	misc.
occupacy	44.0%	47.3%	7.8%	0.9%

までの遅延は 0.4 秒であった．これは，実時間で処理ができていることを示している．499 秒のうち，音声認識は 130 秒，FlowDesigner 上の処理に 369 秒を要した．また，FlowDesigner 上の各処理の負荷を Table 3 に示す．定位，分離の処理量が大きい事がわかる．定位は，MUSIC における固有値計算，分離は，GSS における分離行列更新が処理量が大きくなっている原因である．これらは，より高速なアルゴリズムや，ハードウェア命令を有効に利用すればさらに処理時間を削減することができると考えられる．

4. おわりに

ロボット聴覚用ソフトウェア HARK について述べた．HARK は，1) ロバストなロボット聴覚システムの構築支援，2) システム構築のための一通りのモジュールの提供，3) 多くの多チャンネル音響信号入力デバイスサポート，4) GUI を用いてソフトウェア構成を容易に変更可能といった特長を備えており，高いソフトウェアポータビリティーと再利用性を提供することを示した．また，雑音下や同時発話の音声認識といった実験を行いその性能面での有効性を示した．

HARK はロボット聴覚分野だけでなく，対話，視覚処理，ナビゲーションといった研究への貢献も期待できよう．実際に口じゃんけんゲームの審判，食事の注文伺いといったタスクへの応用例を通じてその有効性も示されている [3].

謝辞

Jean-Marc Valin 氏，山本 俊一氏に感謝する．

[1] K. Nakadai *et al.* Active audition for humanoid. *AAAI-2000*, pages 832–839. AAAI, 2000.
[2] S. Yamamoto *et al.* Making a robot recognize three simultaneous sentences in real-time. *IROS 2005*, pages 897–902. IEEE, 2005.
[3] 中臺 一博他. ロボット聴覚用オープンソースソフトウェア hark の概要. In *Robomec 2008, 1P1-G13*, 2008.
[4] C. Côté *et al.* Reusability tools for programming mobile robots. *IROS 2004*, pages 1820–1825. IEEE, 2004.
[5] J.-M. Valin *et al.* Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems Journal*, 55(3):216–228, 2007.
[6] ALSA. http://www.alsa-project.org/.
[7] RASP. http://jeol-st.com/mt/2007/04/rasp2.html.
[8] TD-BD-8CSUSB. http://www.inrevium.jp/eng/news.html.
[9] F. Asano *et al.* Sound source localization and signal separation for office robot “Jijo-2”. In *Proc. of IEEE Int'l Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI-99)*, pages 243–248, 1999.
[10] Y. Nishimura *et al.* Noise-robust speech recognition using multi-band spectral features. In *Proc. of 148th Acous. Soc. of Am. Meetings*, page 1aSC7, 2004.
[11] J. Barker *et al.* Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Proc. of Eurospeech-2001*, pages 213–216. ESCA, 2001.
[12] M. Cooke *et al.* Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285, May 2000.
[13] Multiband Julius. http://www.furui.cs.titech.ac.jp/mband_julius/.
[14] T. Kawahara and A. Lee. Free software toolkit for Japanese large vocabulary continuous speech recognition. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 476–479, 2000.