頭部伝達関数を用いた GSS による 3 話者同時発話認識 ～HARK 1.0.0 の新機能～

○高橋 徹† 中臺 一博‡ 駒谷 和範† 尾形 哲也† 奥乃 博†

Improvement of Simultaneous Listening by Empowering Geometric Source Separation with HRTF – New Capabilities of HARK 1.0.0 –

*Toru TAKAHASHI†, Kazuhiro NAKADAI‡, Kazunori KOMATANI†,
Tetsuya OGATA†, Hiroshi G. OKUNO†
† Graduate School of Informatics, Kyoto University
‡ Honda Research Institute Japan Co., Ltd.

Abstract— This paper presents a newly released open-source robot audition software HARK 1.0.0. First, geometric source separation (GSS) is empowered by HRTF to force more strict constraints on directions to improve the performance of sound source separation Second, GSS suppresses directional noises made by motors given their directions in advance to enhance noise-robustness. Third, acoustic features for automatic speech recognition (ASR) and thus missing feature masks are redesigned. Acoustic models are trained with multi-conditioning data with new acoustic features. HARK 1.0.0 is implemented on HRP-2 and various benchmark shows the performance improvement thanks to all the new features. The word correct rates improve by 5 and 10 points under normal and noisy acoustic environments, respectively.

Key Words: HARK, Robot Audition, Simultaneous Speech Recognition, GSS, HRP-2

1. はじめに

近年, ロボットに人間らしいコミュニケーション能力が求められている. 老若男女によるロボットの広範な利用には, 特別なスキルをユーザに要求しない自然なインターフェースが必要である. ロボットが, 対話能力を持つことでこの要求を満すことができる. これまで, AIBO [1], PaPeRo [2] など多くのロボットが発表され, 我々の生活空間で利用する実験が行われている. これまでは, ユーザがヘッドセットを装着し音声認識するロボットがほとんどであった. これらのロボットは, 音声対話が可能で, 約 1,000 の語彙を自身の耳（マイクロフォン）を用いて認識可能であり, 人間と音声対話を行うことが可能である. これらのロボットは, 基本的に単一音源を対象としている. しかし, 音声認識を生活空間で利用するには, 複数音源を扱う必要がある. 我々は, 単一音源の音声認識だけではなく, 同時発話音声認識も可能なロボット聴覚ソフトウェア HRI-JP Audition for Robotswith Kyoto University (HARK) を開発している.

ロボット自身の耳で聴くと, 音源位置と収音位置が, 1 メートル以上離れる場合がある. 距離が離れると周囲の雑音により目的音声の SN 比が低下し, 残響問題が顕在化する. この問題に対して, HARK は, 収音に, 8 本のマイクロフォンアレーを用い, 目的音を分離・強調し, 認識する. 特定方向の音源分離は, 残響環境でも直接音成分を扱うことが可能であり, 離れた音源の認識が可能となる. 複数音源がある場合は, 定位・分離・強調により個々の音源を認識可能で, 音源が 1 つの場合の自然な拡張になっている.

本稿では, HARK 1.0.0 の新機能を紹介する. これをヒューマノイドロボット HRP-2 に実装し, 新機能による性能改善を示す.

具体的には, 新規モジュールの追加と, これまで他のパッケージの実装を利用していた Geometric Source Separation (GSS) 音源モジュールを新規に実装し, 機能拡張を行った. これらの拡張により, 単語正解精度が, クリーン環境で 5 ポイント, 雑音環境で 10 ポイント向上した.

2. HARK 0.1.17 から 1.0.0 への改良

HARK を 2008 年 4 月に公開して以来, 1 年余りが経ち, 改善に関する数々のフィードバックを頂いた. 我々は, HARK を, モジュール群 (入出力モジュール, 音源定位・追跡モジュール, 音源分離モジュール, 音響特徴量モジュール, 自動ミッシングフィーチャマスク生成モジュール, 音声認識インターフェースモジュール, データ変換・操作モジュール) と, ミッシングフィーチャ理論に基づく音声認識から構成した. 分離性能は, 認識性能に大きな影響を与えるため, 分離モジュールの改良か急務であった.

HARK 0.1.17 では, GSS 部分を ManyEars に依存していたが, HARK 1.0.0 では, すべてのモジュールを独自に提供する. これによりロボット聴覚システムを HARK 1.0.0 単独で構成可能となった. HARK 1.0.0 の最大の変更点は, GSS モジュールを独自に提供した点と GSS モジュールを拡張した点である. これに伴い, ManyEars では GSS モジュール中に統合されていたポストフィルタを独立したモジュールとして分離し, 各種パラメータ設定の変更を可能とすることでユーザビリティを向上させた. また, これに伴い, Minima

Table 1 Modules provided by HARK 1.0.0

Category Name	Module Name
Multi-channel Audio I/O	AudioStreamFromMic AudioStreamFromWave SaveRawPCM
Sound Source Localization and Tracking	LocalizeMUSIC ConstantLocalization SourceTracker DisplayLocalization SaveSourceLocation *LoadSourceLocation *SourceIntervalExtender
Sound Source Separation	DSBeamformer *GSS *Postfilter *BGNEstimator
Acoustic Feature Extraction	MelFilterBank *MFCCExtraction MSLSExtraction SpectralMeanNormalization Delta *FeatureRemover *PreEmphasis SaveFeatures
Automatic Missing Feature Mask Generation	MFMGeneration DeltaMask *DeltaPowerMask
ASR Interface	SpeechRecognitionClient *SpeechRecognitionSMNClient
MFT-ASR	Multiband Julius/Julian (non-FlowDesigner module)
Data Conversion and Operation	MultiFFT Synthesize WhiteNoiseAdder ChannelSelector *SourceSelectorByDirection *SourceSelectorByID *MatrixToMap *PowerCalcForMap *PowerCalcForMatrix

Controlled Recursive Average (MCRA) に基づく背景雑音推定モジュールも新規に実装を行った．その他, 音声認識に有効であるとされる Δ パワー特徴量を音響特徴量とし取扱えるようモジュールを追加した．HARK 1.0.0 のモジュール一覧を表 1 に示す. 表中 ∗ 印が新規モジュールである．なお, 一部のモジュールは, 機能に合致するよう名称変更を行っている．

3. 主要な変更

本稿では紙面スペースの都合上, 主要な変更点として, GSS とポストフィルタについて概説する．

3·1 GSS の機能拡張

GSS は, ブラインド音源分離 (Blind Source Separation： BSS) [3] の一種で, ビームフォーミングとのハイブリッドアルゴリズムである．BSS と異なり, マイクロフォンと音源位置に関する幾何的拘束条件を用い, BSS で問題となるパーミュテーション問題やスケーリング問題を原理的に回避している．HARK 1.0.0 に含む拡張版 GSS は, ロボット聴覚のための 3 つの新しい特徴がある．

特徴 1： 分離性能を向上させるため，GSS の分離行列更新のコスト関数に利用する幾何拘束を柔軟に設定できるようにした．一般に GSS の幾何拘束は, マイクロフォンと音源位置から得る．しかし, ロボット頭部の影響を反映させた方がより正確な幾何拘束が得られるため, 実測したロボット頭部伝達関数を幾何拘束として利用できるよう改良を行った．これにより，分離性能向上と分離行列の収束速度向上が期待できる．

特徴 2： ロボット自身が生成する方向性雑音を考慮可能とした．これにより，ロボットのファン音のような定常的な方向性雑音を予め指定できるため，雑音の分離性能が向上が期待できる．

特徴 3： 分離行列更新に関して，話者とロボットの位置関係の変動を考慮した．GSS の分離行列更新で用いる幾何拘束は話者とマイクロフォン間の幾何的な関係に基いて生成されるため, 分離行列，および用いる幾何拘束は, 話者とロボットの相対位置関係が変化すれば再初期化すべきである．しかし, 頻繁な再初期化は, 分離性能の低下を招く．そこで，初期化の基準とタイミングを調整できるパラメータを導入し，再初期化条件を調整可能とした．

3·2 ポストフィルタの実装

ポストフィルタモジュールは, GSS モジュール [4] の出力音声を強調するために用いる．音源分離問題は不良設定問題であるため，GSS の分離音には, 定常的な背景雑音や, 非定常な分離エラーが含まれる．そこで, HARK 1.0.0 では，ManyEars の SeparGSS に導入されている定常雑音と非定常雑音の両方を抑圧できるスペクトルフィルタを単独モジュール, ポストフィルタモジュールとして実装した．これにより，単独で GSS を用いたり，GSS とポストフィルタモジュールを併用して SeparGSS と同等の機能を実現したりできるようになり，柔軟性が向上した．また，ポストフィルタには多数のパラメータがあり，分離性能を向上させるためには，これらの値を最適にチューニングする必要があるが，SeparGSS では，チューニング可能なパラメータ数が少なかった．HARK 1.0.0 のポストフィルタモジュールでは，ほぼすべてのパラメータが設定可能となっており，柔軟な制御を可能とした．これは，特に, 様々なロボットに HARK を適用する際のパフォーマンスチューニングに効果的である．

4. 実験

HARK 1.0.0 で拡張した機能の中から, 3 つの機能について評価する．HARK は, 本来, 音源定位・音源分離・認識など様々な側面での評価が必要である．しかし紙面の都合があるため, 本稿では, 総合評価として単語正解精度で評価を行う．

評価 1 頭部伝達関数を使った幾何的拘束条件付きの GSS の性能比較．この 2 つの単語認識精度を比較することで頭部伝達関数を制約として用いることの有効性を確認する．

評価 2 ロボットのファン音を対象とした雑音下での単語正解精度を比較する．方向性雑音の方向を予め与える HARK 1.0.0 と与えることのできない HARK 0.1.17 で比較する．

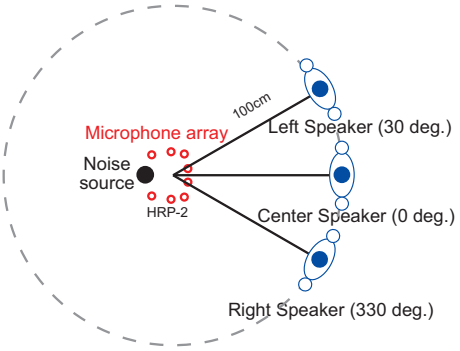

Fig.1 3 話者同時発話認識のロボットと話者の配置.

評価 3 音響特徴量として従来 MSLS 24 次の動的・静的の 48 次元特徴と新モジュールの Δ パワー計算モジュールを使って, MSLS 13 次の動的・静的特徴量と合せた 27 次元特徴量による単語正解精度を比較する.

4·1 実験条件

男性 3 名の同時発話による単語認識実験により評価した. 3 話者同時発話音声は, 計算機シミュレーションにより作成した. 予め HRP-2 頭部モックアップの伝達特性を各方向 1m の地点から計測し, 音声データに実測のインパルス応答を畳み込んだ上で, 3 音を同レベルで混合した. 各話者は, ロボットの正面に 1 名, 左右にそれぞれ 15, 30, 45, 60, 75, 90, 105, 120 度の位置にそえぞれ 1 名配置し同時に発話する条件である. 3 名の発話は, 長時間平均エネルギーが等しい発話となるよう正規化しロボットに呈示じた. 3 話者 30 度間隔条件の話者配置例を図 1 に示す. ロボットの中心から話者までの距離は 1m である. 発話単語は, 話者ごとに異なり, 同じ単語は同時に発声しないものとした.

4·2 評価 1 : 頭部伝達関数を用いた GSS の性能

HARK 1.0.0 の頭部伝達関数を使った幾何的拘束条件付きの GSS と, HARK 0.1.17 のマイクロフォンと音源位置での幾何的拘束条件付きの GSS で単語正解精度を比較する. 伝達関数は, 図 4·2 右側に示す HRP-2 の頭部モックアップで測定した. 実際の HRP-2 ロボットは, 図 4·2 左側に示す. HRP-2 頭部モックアップと実際の HRP-2 の頭部伝達関数は異なっている. HRP-2 の頭部伝達関数は, 胸部以下の身体の影響を受けるため, モックアップから得られる頭部のみの伝達関数を用いることで首の向きに依存しない汎用的な伝達関数となる利点がある.

認識用の音響モデルは, 新聞記事読み上げ音声コーパス (JNAS : Japanese Newspaper Article Sentences) で学習した 3 状態トライフォン HMM である. 出力確率分布は対角共分散型の 16 混合ガウス分布でモデル化した. コンテキストクラスタリングにより総状態数を 2000 状態にしている. 音響特徴量は MSLS 13 次の静的・動的特徴量と Δ パワーから成る 27 次元特徴量である. ATR 音素バランス単語 216 単語中の 200 単語を用いて単語正解精度を比較した. GSS の分離音声を認識した. 結果を図 3 に示す. 縦軸と横軸は, 単語正解精度と, 3 話者の間隔を表している.

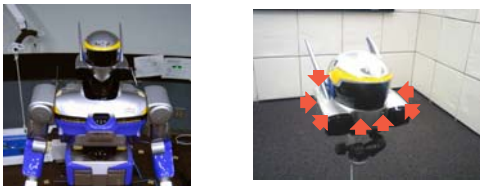

Fig.2 左:8 ch マイクロフォン搭載のヒューマノイドロボット HRP-2W 改の全体図. 右:HRP-2 の頭部モックアップとマイクロフォン位置. 赤矢印はマイクロフォン位置を示す. ただしロボット左後頭部は隠れているので矢印は非表示.

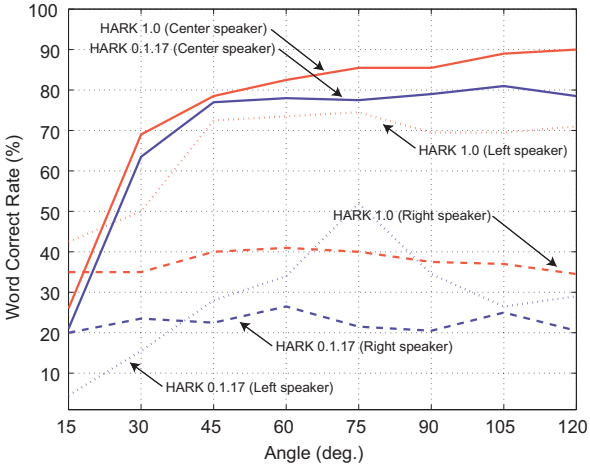

Fig.3 HARK 1.0.0 と HARK 0.1.17 による GSS の単語正解精度の比較.

HARK 1.0.0 が, すべての条件で, HARK 0.1.17 の精度を上回っている. 正面話者に対して 10 ポイント, 周辺話者に対して 15 から 30 ポイント改善している. HRP-2 のマイクロフォン配置が, 前方で密のため, 正面話者の正解精度が常に最大となる. 音源分離は, 空間的なスパースネスに基いているので, 角度が広がるに伴い分離性能が向上し, 正解精度が増加している.

4·3 評価 2 : 方向性雑音除去機能のある GSS

方向性雑音に対する耐性を評価するためにロボットのファン音を対象とした雑音下での単語正解精度を比較する. 方向性雑音として HRP-2 (図 4·2) の待機時のファンノイズを収録し, 計算機上で 180 度方向からノイズを加えた 3 話者同時発話音声を分離・認識する実験を行った.

HARK 1.0.0 と HARK 0.1.17 で, 単語正解精度を比較した結果を図 4 に示す. 縦軸と横軸は, 単語正解精度と, 3 話者の間隔を表している. 図 3 と異なる条件は, 3 つで, 音響モデルの構築条件と認識条件, 音響特徴量を求めるのにポストフィルタ出力を用いた点, ミッシングフィチャマスク処理 [5, 6] を適用した点である. 音響モデルは, JNAS で学習したが, 一度 GSS によって分離した JNAS の音声データベースと合せてマルチコンディションで学習した. 分離は, 0 度方向 1m 条件 1 話者再生で行った. この条件は, HARK 1.0.0 で可能

なすべての処理を適用しており 3 話者同時発話認識の最高パフォーマンスと位置付けられる．この条件で，方向性雑音を予め与える効果の有無による単語正解精度で比較している．

HARK 1.0.0 の単語正解精度は，ほとんど全ての条件 (右話者 75 度条件以外すべての条件) で，HARK 0.1.17 の精度を上回っている．特に，正面話者の正解精度は，45 度間隔より狭い条件では，大きな改善がみられる．周辺話者の正解精度は，75 度条件で，極端に低下している．HRP-2 の頭部にある角部分の影響と考えられる．

単語正解精度は，話者間角度とノイズ方向の両方に依存する．一般に，話者間角度が大きくなると，話者間の分離性能はあがるので，単語正解精度は，角度と伴に単調増加傾向を示す．しかし，今回の実験の様に 0 度方向の話者 1 名と，左右にそれぞれ 1 名ずつの配置で，180 度方向に方向性雑音がある条件では，話者間角度が大きくなると，周辺話者は，ノイズ方向に近付き，周辺話者の音源分離性能が低下する．従って，話者間角度が適度に離れ，かつ周辺話者がノイズ方向と適度に離れている配置で正解精度が最大となる．必ずしも話者間角度に伴って単語正解精度が単調増加せず，話者間角度とノイズ方向の両方に依存していることがわかる．

4·4　評価 3：音響特徴量とマルチコンディション学習

音響特徴量の次元削減によって計算効率の改善や，記憶量の削減が期待できる．我々は，クリーン音声コーパスを用いて音響モデルを構築し，1 話者によるクリーンな孤立単語音声を用いた単語正解精度を比較したところ，27 次元の音響特徴量と従来の 48 次元を用いた場合で同等の単語正解精度が得られることがわかった．分離音声に対する音響特徴量では，従来 MSLS 24 次と Δ MSLS 24 次の合計 48 次元の音響特徴量を用いてきた．今回 13 次 MSLS と 13 次 Δ MSLS と Δ パワーを用いた合計 27 次元の特徴量を用いた実験を行った．

3 話者同時発話認識において，GSS による分離音とミッシングフィーチャマスク処理を行った場合，正解精度を正面話者に対して 5 から 10 ポイントの改善がみられた．次元削減によって，音響モデルのモデルパラメータの自由度が減少し，ロバストなモデルを学習できたためと推測される．

この実験結果を踏まえ，HARK 1.0.0 では音響特徴量に Δ パワーの使用に対応した．具体的には，Δ パワー計算用モジュールと，Δ パワー用のミッシングフィーチャーマスク生成モジュールを追加した．ただし，Δ パワーの信頼度計算が未確立のため，現在の実装では，常に特徴量を信頼するマスク値を生成する．

5.　結論

新しいバージョンである HARK 1.0.0 を紹介し，新モジュールの評価を行った．GSS における音源とマイクロフォンの相対的な位置情報を制約としていた従来方法に比べ，測定した頭部伝達関数を用いた新方法を用いた場合，3 話者同時発話認識における単語正解精度が最大で 30 ポイント向上した．

方向性雑音がある場合には，新モジュールでは，性雑音方向の音源定位結果を音源分離の対象から除外できるため，従来モジュールに比べ，方向性雑音にロバストになった．180 度に方向にロボットのファンノイズが存在する場合を仮定したシミュレーション実験では，従来法と比較して，一部の条件を除き，周辺話者での単語認識精度が 10 から 40 ポイント向上した．正面話者に対しては，マルチコンディション学習と音響特徴量の改善，GSS とポストフィルタ，ミッシングフィーチャマスクをすべて使用する最も良い条件で 75 度以上の話者間隔があれば，方向性雑音がある条件であっても 90 % 以上の単語正解精度が得られた．

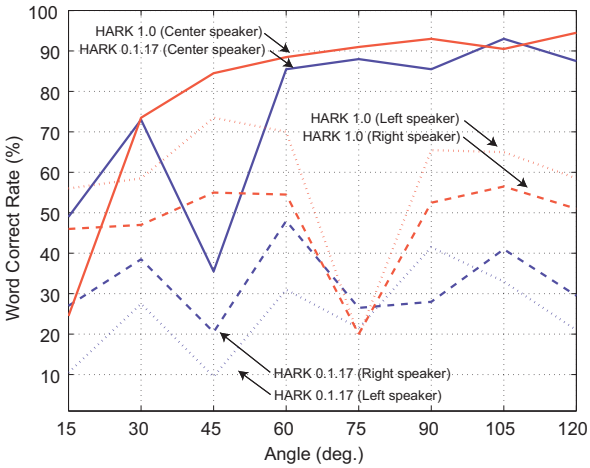

Fig.4 単語正解精度．

音響特徴量について検討を行った．従来 48 次元の特徴量を用いていたが 27 次元の特徴量に変更することで，少い次元数であるにもかかわらず，単語正解精度が，正面話者に対して 10 ポイント，周辺話者に対して最大 15 ポイント改善した．

謝辞

本研究課題は，科学研究補助金と京都大学グローバル COE プログラムの支援により行われた．研究を進める上で，貴重な助言を頂いた HRI-JP の中島弘史博士に感謝する．

参考文献

[1] M. Fujita: “AIBO : Toward the era of digital creatures”, *IJRR*, 20(10):781–794, 2001.

[2] 岩沢 透: “パーソナルロボット PaPeRo の音声認識インタフェース”, 人工知能学会研究会資料, JSAI Tech. Rep., SIG-Challenge-0317-3, pp.17–22,2001.

[3] Parra, L. C., *et al.*: “Geometric source separation: Merging convolutive source separation with geometric beamforming”, *IEEE Trans. SAP* Vol.10, No.6, pp.352–362, 2002.

[4] S. Yamamoto, *et al.* “Making a robot recognize three simultaneous sentences in real-time”, *Proc. IEEE/RSJ IROS*, pp.897–902, 2005.

[5] T. Takahashi, *et al.*: “Soft missing-feature mask generation for simultaneous speech recognition system in robots”, *Proc. Interspeech*, pp.992–995, 2008.

[6] T. Takahashi, *et al.*: “Missing-feature-theory-based robust simultaneous speech recognition system with non-clean speech acoustic model”, *Proc. IEEE/RSJ IROS*, 2009, (to appear).