Coarse-to-Fine チューニングを用いた HARK の音源定位パラメータの最適化

杉山　治[1], **小島 諒介**[1], **中臺 一博**[1,2]

Osamu SUGIYAMA[1], **Ryosuke KOJIMA**[1], **Kazuhiro NAKADAI**[1,2]

1. 東京工業大学 大学院 情報理工学研究科,

2.（株）ホンダ・リサーチ・インスティチュート・ジャパン

1. Graduate School of Information Science and Engineering, Tokyo Institute of Technology,

2. Honda Research Institute Japan Co., Ltd.

{sugiyama.o, kojima, nakadai}@cyb.mei.titech.ac.jp

Abstract

本稿ではオープンソースロボット聴覚ソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) の音源定位におけるパラメータ最適化のためのインタフェースを提案する．HARK でパラメータ調整用のインタフェースは存在するものの，HARK に熟練していてもそのパラメータの最適化には時間を要する．本稿で提案するインタフェースは，HARK のパラメータ最適化における課題を，可視化，操作，最適化における課題に分類し，それぞれを解決する機能を設計・実装した．そして，ユーザ評価において，可視化性・設定の柔軟さの点で，従来のインタフェースを上回るという結果を得た．

1　はじめに

本稿では，オープンソースロボット聴覚ソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) のパラメータ最適化を効率的におこなうことができるよう，HARK の音源定位機能に焦点をあて，インタラクティブなインタフェースを提案する．

HARK は，2008 年にロボット音響における OpenCV を目指しリリースされたオープンソースソフトウェアである [Nakadai 10]．複数のマイクロホンからなるマイクロホンアレイを用いた処理に対応し，音源定位 [Nakamura 09, Ohata 13]，音源分離 [Nakajima 08]，音声認識といった機能を，HARKDesigner と呼ばれるグラフィカルユーザインタフェースを用いて組み合わせることで柔軟なロボット聴覚ソフトウェアを作成することができる．HARK を用いることで，例えば４人のユーザが同時に発話するような状況においても，個々の発話を音声認識するロボットアプリケーションを容易に作成することが可能になる．

HARK の最新版であるバージョン 2.2 でもパラメータを調整するためのインタフェースは存在するが，熟練した作業者がパラメータの最適化を行った場合でも数日を要することもあり，ソフトウェアが安定して使えるようになるまでのオーバヘッドが高い．本研究では，これらの音源定位のパラメータ最適化における課題を，可視化・操作・最適化の３つの観点から整理し，それぞれの課題を解決するためのインタラクティブなインタフェースを設計・開発する．提案するインタフェースでは，音源定位の処理過程を可視化し，マウスジェスチャによる直感的に変数の変更を可能にした．さらに，システムが変数の最適値の予想を示し（Coarse チューニング），それを元にユーザがより正確に変数を最適化する（Fine チューニング）手順を踏む Coarse-to-Fine チューニング [Fujii 11] を取り入れた．これらのインタフェースの機能を利用することで，ユーザは従来のインタフェースより直感的に音源定位のパラメータを設定・最適化することができる．また，ユーザによる定性評価を実施し，提案インタフェースの有効性を検証した．

2　課題とアプローチ

図 1 に HARK における音源定位のプロセスを示す．まず，マイクアレイから多チャンネル音声信号を取得し，短時間フーリエ変換 (Short-Time Fourier Transform, STFT) にかけて周波数スペクトラムへと変換する．その後，MUltiple SIgnal Classification (MUSIC) 法 [Schmidt 86] を用いることで，横軸が時間・縦軸が方位角，色がパワーを示す MUSIC スペクトログラムを得る．最後に，音源追跡により，MUSIC スペクトログラムから音源の位置情報を抽出する．この過程で，ユーザは，以下のパラメータを設定する必要がある．

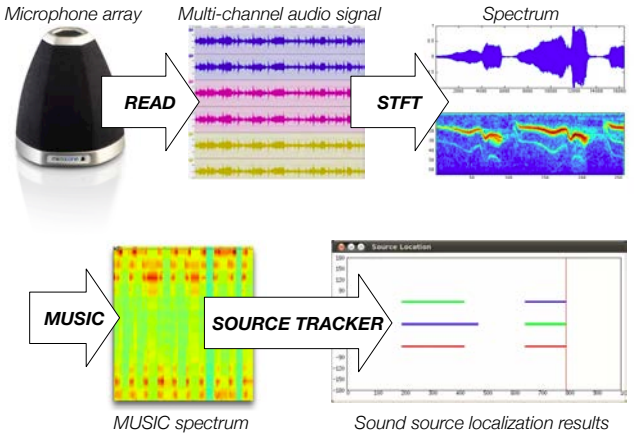


図 1: 音源定位のプロセス

- **num sources**: 音源数
- **thresh:** 音源と雑音を分けるパワーの閾値
- **pause length**: 音源の前区間長
- **preroll length**: 音源の後区間長

これらのパラメータを個々に最適化するには時間がかかり，実験環境における HARK の即時セットアップの障害となっている．本研究では，この問題を以下の 3 つの課題に分類し，それぞれを解決するインタラクティブなインタフェースを設計・開発する．

- **可視化の課題**: 音源定位の途中のプロセスを可視化できていないため，経過を見ながらパラメータの調整ができない
- **操作の課題**: 閾値などのパラメータを直接数値で調整することは非直感的であり，またその結果が即時に反映されない
- **最適化の課題**: システムによる最適化支援機能がない．ユーザは一からパラメータを調整しなければならない

以降の節では，提案するインタフェースがこれらの課題をどのように解決するのかを詳細に述べる．

3 音源定位のためのインタラクティブインタフェースの提案

本稿で提案するインタフェースは，先に述べた可視化・操作・最適化における 3 つの課題の解決を図り，HARK における音源定位のパラメータ調整の時間を短縮することを目的とする．一般に，HARK を用いて音源定位のパラメータを調整する時，ユーザは，1) 適当なパラメータセットを選択し，それを用いて音源定位を行い，定位結果と MUSIC スペクトログラムを得る．2) 得られた音源定位結果と MUSIC スペクトログラムを比較し，雑音部と音源部を推測する．3) 推測した通りにそれらの雑音部と音源部を分けるパラメータセットを導き出す．

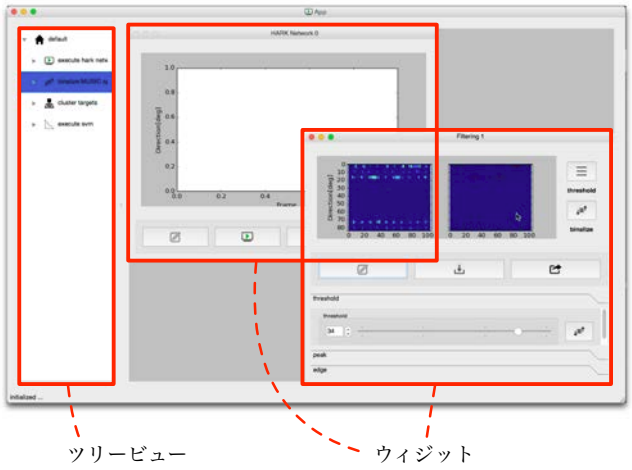


図 2: 提案するインタフェースの概観

3.1 音源定位プロセスの可視化

本稿で提案するインタフェースの概要を図 2 に示す．提案するインタフェースは上記の 3 プロセスを実行する以下の 3 つのウィジットを持つ．

a) 音源定位実行ウィジット

b) 音源のラベル付けウィジット

c) 動的閾値最適化ウィジット

音源定位のパラメータ調整に必要な処理を複数のウィジットに分けることで，ユーザはそれらのプロセスを同時に確認しながら，多面的にパラメータの調整をすることができる．

図 3 に 3 つのウィジットの概観を示す．それぞれのウィジットは共通してチャートボックスとコントロールボックスを持ち，チャートボックスでは，各音源定位過程の可視化を，コントロールボックスでは各定位過程の実行とパラメータ調整を行う．

音源定位の実行　音源定位の実行は音源定位実行ウィジットで行う（図 3a)）．このウィジットでは，HARK による音源定位を実行することができ，コントロールボックスで，解析する多チャネル音声ファイル，チャネル数，伝達関数，音源数，音源時間長をパラメータとして指定することができる．チャートボックスは，音源定位のプロセスの途中で得られる MUSIC スペクトログラムが表示され，得られた定位結果と MUSIC スペクトログラムを音源のラベル付けウィジットに出力することができる．

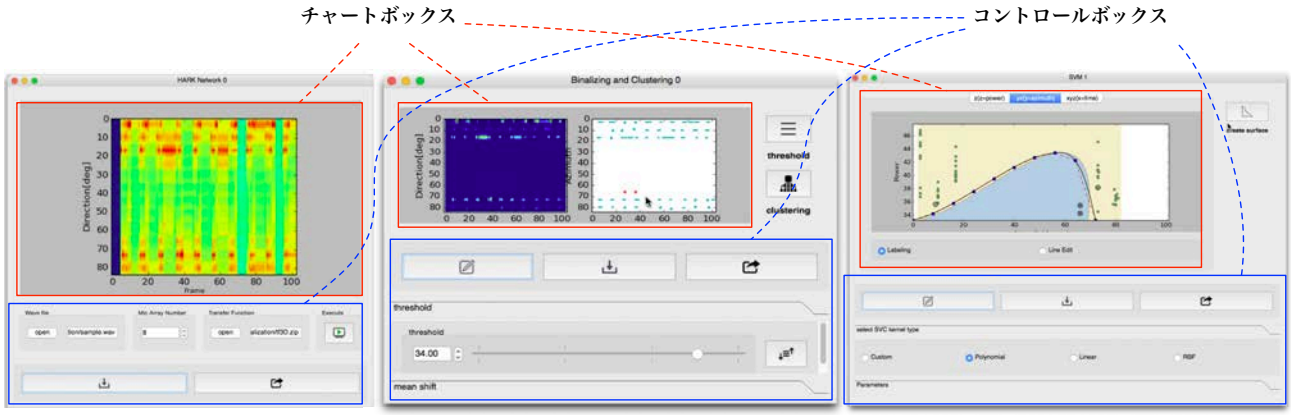


図 3: 音源定位のパラメータ調整のためのウィジット

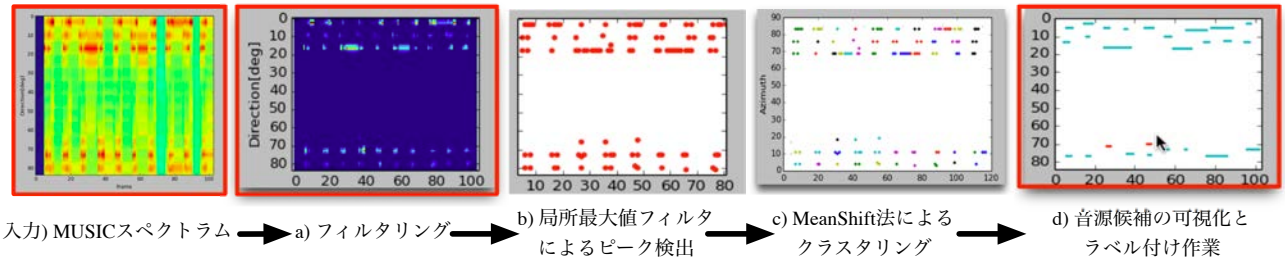


図 4: ラベル付けウィジットのバックグラウンド処理

音源のラベリング　音源ラベル付けウィジットでは，MUSIC スペクトログラムの表示に対して直接，音源のラベル付けを行うことができる（図 3b)）．チャートボックスには 2 つのチャートが表示され，一方には MUSIC スペクトログラムが，もう一方には音源候補が図示される．ユーザは最初のチャートを用いて雑音部を除去することで音源部を，次のチャートで音源の候補を確認，その候補が音源なのか，雑音なのかをラベリングする．これらの操作を実行するため，音源ラベル付けウィジットはバックグラウンドで以下の処理を実行する．

a) 閾値によるフィルタリング

b) 局所最大値フィルタによるピーク検出

c) 検出されたピークをクラスタリングすることによる音源候補の抽出

d) 音源候補の可視化とラベリング

図 4 は，上記のバックグラウンドプロセスの過程を図示したものである．図 4 中，赤い枠線を持つものは処理結果が可視化される処理を表し，それ以外のものはチャート上には図示されずバックグラウンドで処理される．

閾値によるフィルタリングでは，1) ピーク検出にむけて MUSIC スペクトログラムの低パワー部を除去する．2) 局所最大値フィルタ [Nishiguchi 04] によってピーク検出を行う．3) この処理によって得られたピーク群を，Mean-Shift 法 [Okada 08] を用いてクラスタリングし，得られたクラスタを音源候補とする．4) それぞれのクラスタをチャートボックスの右のチャートにレンダリングする．この際，クラスタを構成するピーク時間軸の最大値と最小値の差をそのクラスタ長とする．またこの間の方向軸の平均が縦軸の値としてプロットされる．

表 1: 最適化処理に用いるパラメータ

ウィジット	アルゴリズム	変数名	型	初期値
ラベリング	フィルタリング	power	float	32.0
ラベリング	局所最大値フィルタ	x	int	1
		y	int	2
ラベリング	Mean Shift	kernel_size	float	0.02

これらの過程で必要な局所最大値フィルタのフィルタサイズや Mean-Shift 法のカーネルサイズなどの各パラメータはコントロールボックスのスライダーで調整することができる．また，その値を数値としても確認することができる．各パラメータの初期値を表 1 にまとめる．

動的閾値の最適化　動的閾値最適化ウィジットでは，音源と雑音を分けるパワー閾値を動的に設定することができる（図 3c)）．閾値を複数の視点から設定できるようにす

るためチャートボックスはマルチタブ構成になっており，それぞれのタブでは以下に示す複数の次元で音源候補をプロットする．

1D 縦軸を各音源候補のパワーの平均とし，それぞれの音源候補のパワーを降順に並べたもの

2D 縦軸を各音源候補の各方向軸ごとのパワーの平均とし，横軸を方向として音源候補をプロットしたもの

3D 縦軸を各音源候補のパワーとし，横軸を時間フレーム，奥行きを方向として音源候補をプロットしたもの

音源候補は，ラベル付けウィジットで事前にラベル付けされており，音源とラベル付けされたものは青く，雑音とラベル付けされたものは赤くプロットされる．ユーザはこれらの音源と雑音を切り分ける境界を，サポートベクタマシン (Support Vector Machine, SVM) によってラフに求め (Coarse チューニング)，マウスジェスチャによって閾値を詳細に設定することができる（Fine チューニング）．これらの挙動については，3.3 節で詳しく述べる．一方，コントロールボックスでは，SVM のカーネルの選択，それぞれのパラメータを調整することができる．ユーザは，これらのインタフェースを用いることで直感的に音源と雑音を分ける閾値を設定し，音源定位に反映することができる．図 6 は多項式カーネルを用いた場合の閾値の設定例である．

3.2 ジェスチャ操作によるインタラクティブなインタフェース

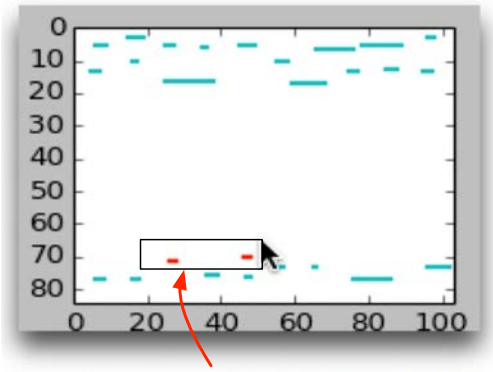
提案インタフェースのジェスチャ操作について述べる．既存の HARK のインタフェースでは，音源定位のパラメータを数値で指定するため，その値がどのように結果に反映されるのかわかりにくいという課題があった．本稿では，この課題を解決するために 2 つの機能をインタフェースに実装した．

3.2.1 マウスジェスチャによる音源候補の選択

図 3b), c) のチャートボックスでは，マウスジェスチャによる音源候補のラベリングをすることができる．ユーザはラベル付けしたい音源候補の周辺の矩型領域を，マウスのドラッグ&リリースジェスチャで指定することでラベル付けを行うことができる（図 5）．この情報は，動的閾値最適化ウィジットで，音源と雑音をわける閾値を設定するときに使われる．

3.2.2 パラメータ変更の即時反映

提案インタフェースのすべてのチャートボックスは，パラメータの変更やマウスジェスチャの結果が即時に反映される．ユーザは，自身のパラメータの変更がどのように音源定位結果の各プロセスに影響を与えるのかをチャートボックスから直感的に読み取ることができるため，それぞれの過程で反映される結果を見ながらパラメータの最適化作業をインタラクティブにすることが可能となる．



ドラッグ&リリースで囲った矩型領域の候補をラベリングする

図 5: ラベル付けのためのマウスジェスチャ

3.3 Coarse-to-Fine チューニング

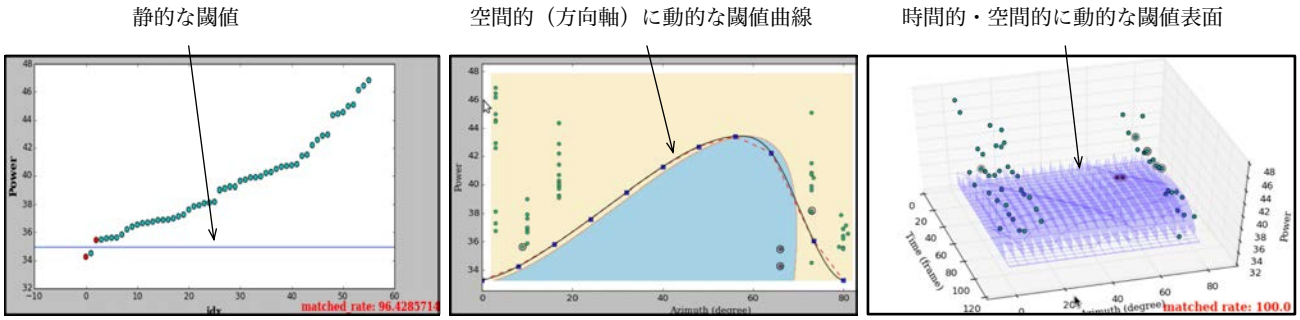
Coarse-to-Fine メカニズムとは，人間の視覚はまず全体を見てから，細部を詳細に見るという動きをするというメカニズムのことである [Menz 03]．このメカニズムは，画像処理における物体認識などに応用されており，本稿では，このメカニズムを組み込んだシステムと人の協調作業の方法を提案する．

環境や状況依存で最適な値が変わってしまうため，機械学習技術を用いても音源定位パラメータの完全な最適化を行うことは困難である．本稿では，機械学習のマシンループにユーザのアドバイスを加えることで短時間で詳細なパラメータチューニングを行うことを目指し，そのためのインタフェースを開発する．

Coarse-to-Fine チューニングの最適化対象は，前述の 3 つのパラメータのうち，音源と雑音を分離する際のパワーの閾値である．HARK の既存のインターフェースでは，この閾値は時間的，空間的に静的にしか設定できなかった．しかし，音源や方向性雑音のパワーに違いがある場合や，ある一定期間，高いパワーのノイズがのってしまった場合には，静的な閾値では対応できないことがある．本稿では，この閾値を空間・時間軸で動的に設定できるようにし，その最適化を Coarse-to-Fine チューニングで行う．

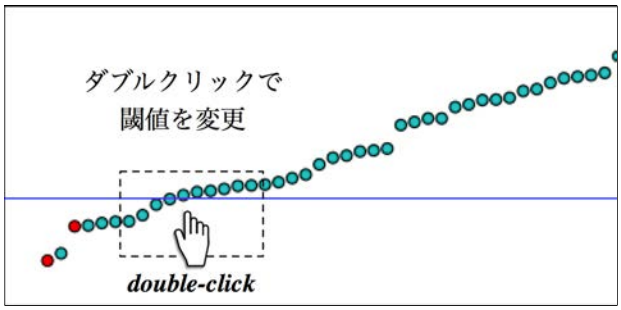
3.3.1 Coarse チューニング

Coarse チューニングでは，システムがラフにパラメータの最適値をユーザに提示する．具体的には，音源と雑音を分けるパワー閾値の動的な変化に対応できるように空間・時間方向に対する閾値曲線（面）として表す．この閾値曲線（面）は SVM を用いて推定する．動的閾値最適化



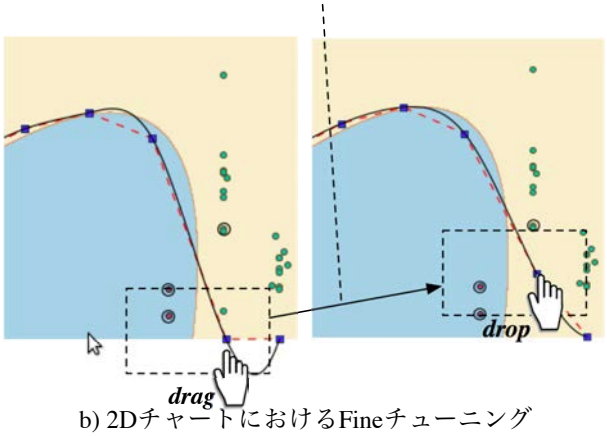
a) 1Dチャート　　b) 2Dチャート　　c) 3Dチャート

図 6: 動的閾値の調整画面



a) 1Dチャートにおける Fine チューニング

ドラッグ&ドロップで閾値曲線を構成するノードを移動



b) 2Dチャートにおける Fine チューニング

図 7: Coarse-to-Fine チューニング

ウィジットは，閾値を設定する 3 つの異なるチャート画面を持つ（図 6）．図 6a) のチャートでは，設定する閾値は静的で既存の HARK と変わらないが，音源フィルタリングウィンドウでラベル付けした音源候補を抽出する閾値を求め，ユーザに提示する．図 6b) のチャートでは空間（方向）軸に沿って，MUSIC スペクトログラム上のパワーの強い領域のピーク座標がプロットで表示される．同時に，ユーザがラベル付けした音源を定位するために最適な閾値の境界曲線を多項式カーネルを用いて求め，提示する．図 6c) のチャートでは，3 次元（時間，空間（方向），パワー軸）空間上に MUSIC スペクトログラムのパワーの強い領域のピーク座標がプロットされており，ユーザがラベル付けした音源を定位するために最適な閾値の境界面を提示する．ユーザはこれらの提示される 3 つの静的閾値，閾値の境界曲線，境界面の中からその状況に最もあったものを選択し，Fine チューニングを行う．

3.3.2 Fine チューニング

Fine チューニングでは，Coarse チューニングで提示されたパラメータの値に基づき，ユーザが詳細にパラメータの最適化を行う．図 7 は，動的閾値最適化ウィンドウにおけるマウスジェスチャ操作を示す．動的閾値最適化ウィジットでは，SVM に基づいてシステムが閾値候補を提示し，その後にユーザが最適値を調整する．その際，図 7a) の 1D チャートでは各音源候補のパワーの平均値が降順にプロットされており，音源を示す青いプロットと雑音を示す赤いプロットをうまく切り分けるように閾値を設定する．閾値の設定は画面をダブルクリックすることで行い，ダブルクリックされた y 軸の値を閾値として採用する．図 7b) の 2D チャートでは境界線をノードをマウスでドラッグ & リリースすることで閾値を自由に変更することができる．Coarse チューニングでシステムから提案された境界曲線は，境界曲線上のノード群とそれらを補完する spline 曲線として，ユーザに提示される．ユーザは提示されたノード群の位置をマウスのドラッグ & ドロップジェスチャで任意の位置に変更することができる．これらのマウスオペレーションは即座にシステムに伝達され，変更された結果が図に反映されるため，ユーザは反映結果を見ながらインタラクティブに閾値の調整を行うことができる．

4 システム評価

提案システムの有効性を評価するために，評価実験を行った．実験では，ロボット実験で収集した多チャンネル音声信号を提案インタフェースでパラメータ調整する様子と，HARK の既存インタフェースで調整する様子をビデオで撮影し，その様子を 8 名の大学院生に見せ，その印象を

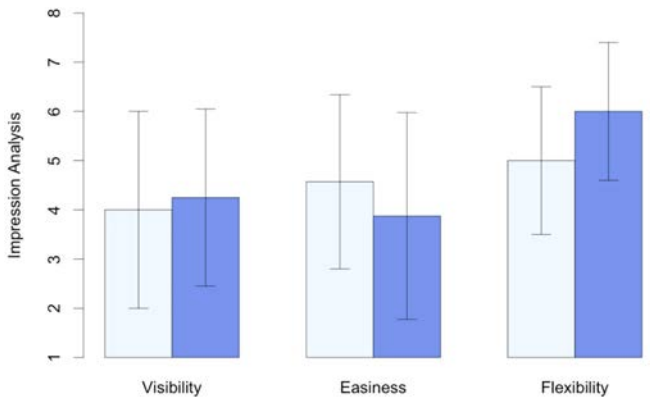


図 8: 定性評価

a) 可視化性，b) 操作性，c) 設定の柔軟性の観点から 7 段階で評価してもらった．なお，実験前に学生はそれぞれのインタフェースの使い方に関するレクチャを 10 分間受けており，その使い方，操作の意味を理解してもらった．

4.1 実験結果

実験結果を 図 8 に示す．図 8 からわかるように，提案インタフェースは，可視化性，設定の柔軟性の 2 つの観点で既存の HARK のインタフェースの評価を上回ることが示された．対して，操作性に関しては既存の HARK インタフェースが上回るという結果になった．

可視化性と設定の柔軟性で既存の HARK インタフェースより良い評価を得たことは本稿の提案するインタフェースが設計の意図通りにユーザの負荷を軽減できていることを示していると考えられる．一方，操作性に関しては，良い評価が得られなかった．実験アンケート後，被験者に実施したインタビューでは，複数の被験者から提案インタフェースは設定する項目が多く，便利だと思われる反面，いろいろと覚えるべきことが多いのではないかという指摘を受けた．これらの懸念が，設定項目が少なく操作できる既存の HARK インタフェースの評価が提案インタフェースより高くなった原因であると考えられる．本稿では，これらのユーザの評価から，それぞれのウィジットでショートカット機能を実装することでシステムによるユーザの補助機能を追加し，操作性においても既存インタフェースを上回る機能を実装する予定である．これらの設計・実装と評価は将来課題である．

5 結論

本稿では，HARK における音源定位のパラメータ最適化のため，インタラクティブなインタフェースを設計・開発した．提案インタフェースは，可視化・操作・最適化における定位パラメータ調整の課題を解決することで，直感的な最適化を行うことができる．そして，ビデオによる評価実験を通じて，可視化性と操作の柔軟性において既存の HARK インタフェースよりも高く評価されることを示した．

謝辞

科研費 24220006 および，JST ImPACT タフロボティクスチャレンジの支援を受けた．

参考文献

[Nakadai 10] K. Nakadai *et al.*: “Design and Implementation of Robot Audition System “HARK”,” Advanced Robotics, Vol.24, pp.739-761, VSP and RSJ, 2010.

[Nakamura 09] K. Nakamura *et al.*, “Intelligent sound source localization for dynamic environments,” IROS 2009, pp. 664-669.

[Ohata 13] 大畑 他, “クワドロコプタを用いた屋外環境音源探索，” SICE SI2013, pp. 360-363.

[Nakajima 08] H. Nakajima *et al.*, “Adaptive step-size parameter control for real-world blind source separation,” IEEE ICASSP 2008, pp. 149-152.

[Fujii 11] 藤井 他, “ロボット聴覚ソフトウェア HARK における音源定位パラメータチューニングの検討,” SICE SI-2011, pp. 202-205.

[Menz 03] M.D. Menz *et al.*, “Stereoscopic depth processing in the visual cortex: a coarse-to-fine mechanism,” Nature neuroscience Vol.6, No.1, pp. 59-65, 2003.

[Schmidt 86] R.O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. on Antennas and Propagation, Vol.34, No.3, pp. 276-280, 1986.

[Carle 04] C. Carle, *et al.* “Code reusability tools for programming mobile robots,” IEEE/RSJ IROS 2004, pp.1820-1825.

[Nishiguchi 04] 西口 他, “スターセンサ画像の暗い星検出への繰り返し型最大値フィルタの応用,” 計測自動制御学会論文集, Vol.40, No.5, pp.573-581, 2004

[Okada 08] 岡田, “ミーンシフトの原理と応用,” 信学技報, Vol. 107, No. 539, PRMU2007-308, pp. 308-346, 2008.