人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402

ＡＩチャレンジ研究会 (第４１回)

Proceedings of the 41st Meeting of Special Interest Group on AI Challenges

CONTENTS

◇ 【基調講演】機械学習のこれから：汎用的なデータ解析を目指して 1
杉山 将 （東京大学）
◇ 相関行列スケーリングを用いた屋外音源探索手法の解析 7
大畑 琢磨（東京工業大学），長峰 諒英（東京工業大学），中村 圭佑（HRI-JP），石崎 孝幸（東京工業大学），水本 武志（HRI-JP），中臺 一博（東京工業大学，HRI-JP）
◇ 屋外音環境理解における音源検出の性能評価と可視化 13
長峰 諒英，大畑 琢磨，上村 知史，小島 諒介，杉山 治（東京工業大学），中村 圭佑（HRI-JP），中臺 一博（東京工業大学，HRI-JP）
◇ 深度センサとマイクロフォンアレイを用いた聴覚アウェアネスの提示 20
井山 貴裕（京都大学），杉山 治（東京工業大学），坂東 宜昭，糸山 克寿，吉井 和佳（京都大学），奥乃 博（早稲田大学）
◇ 臨場感の伝わる遠隔操作システムのデザイン 26
劉 超然，石井 寿憲カルロス，石黒 浩，萩田 紀博（ATR）
◇ 【基調講演】非同期分散マイクロフォンアレーによる音源定位・音源分離 33
小野 順貴（国立情報学研究所）
◇ マイクアレイ伝達関数のオンライン校正とそのロボットへの適用 39
中村 圭佑，中臺 一博（HRI-JP）
◇ マイクロホンアレイとスピーカをもつ柔軟索状ロボットのための動的スピーカ選択による姿勢推定の高速化 45
坂東 宜昭，糸山 克寿（京都大学），昆陽 雅司，田所 諭（東北大学），中臺 一博（東京工業大学），吉井 和佳（京都大学），奥乃 博（早稲田大学）
◇ Robust Hands-free Human-Robot Communication in Reverberant Environments ... 51
Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto, Kazuhiro Nakadai (HRI-JP)
◇ 音源定位における能動耳介での動作の影響について 58
尾堂 航，公文 誠（熊本大学）

日　時　2014 年 11 月 21 日　　場　所　慶應義塾大学 日吉キャンパス 来往舎　シンポジウムスペース
Keio University, Kanagawa, Nov. 21, 2014

社団法人 人工知能学会
Japanese Society for Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-01

機械学習のこれから：汎用的なデータ解析を目指して

Machine Learning in Future: Towards Versatile Data Analysis

杉山将
Masashi Sugiyama
東京大学 複雑理工学専攻
Department of Complexity Science and Engineering, The University of Tokyo
sugi@k.u-tokyo.ac.jp http://www.ms.k.u-tokyo.ac.jp

産業界や基礎科学の様々な分野において，大量のデータの山から新たな価値を創造する機械学習技術の重要性が増している．しかし，解析すべきデータの量・次元・複雑さが爆発的に増加しているため，データ解析手法の研究・開発が社会的なニーズに追いつかなくなりつつある．また，最先端のデータ解析アルゴリズムは極めて高度な確率論・統計学・最適化理論等を駆使して設計されているため，技術修得が著しく困難であるという問題もある．

そこで我々は，データ解析に「データ解析コア技術」という独自の抽象的な階層を導入することを提案している．これは，分類，回帰，特徴選択，異常検出などの主要なデータ解析タスクからなる部分集合を考えるものであり，これらのタスク群に共通して適用できるデータ解析基盤技術を開発してきた．本講演では，確率分布間の距離の推定や情報量などを用いた汎用的な機械学習技術を紹介するとともに，それらの応用例や最新の研究成果についても述べる．

参考文献

[1] 杉山 将. 密度比推定によるビッグデータ解析. 電子情報通信学会誌, vol.97, no.5, pp.353-358, 2014. http://www.ms.k.u-tokyo.ac.jp/2014/IEICE-DensityRatioReview-jp.pdf

[2] 杉山 将. 確率分布間の距離推定：機械学習分野における最新動向. 日本応用数理学会論文誌, vol.23, no.3, pp.439-452, 2013. http://www.ms.k.u-tokyo.ac.jp/2013/DivergenceReview-jp.pdf

[3] Sugiyama, M., Suzuki, T., & Kanamori, T. Density Ratio Estimation in Machine Learning, Cambridge University Press, Cambridge, UK, 2012.

機械学習 1

- 機械学習：データの背後に潜む知識を学習する
- 様々な応用例：
  - 音声・画像・動画の認識
  - ウェブやSNSからの情報抽出
  - 商品やサービスの推薦
  - 工業製品の品質管理
  - ロボットシステムの制御
- ビッグデータ時代の到来に伴い，機械学習技術の重要性は益々高まりつつある

機械学習のタスク 2

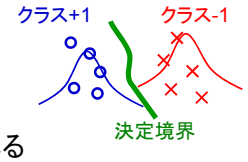
- 機械学習には様々なタスクがある：
  - 非定常環境下での適応学習，ドメイン適応，マルチタスク学習
  - 二標本検定，異常値検出，変化点検知，クラスバランス推定
  - 相互情報量推定，独立性検定，特徴選択，十分次元削減，独立成分分析，因果推論，クラスタリング，オブジェクト適合
  - 条件付き確率推定，確率的パターン認識

最も汎用的なアプローチ 3

- データを生成する規則（確率分布）を推定すれば，あらゆる機械学習タスクが解決できる！

データの生成規則を知る ⇔ データの全てを知る

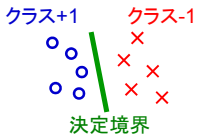
- 例：各クラスのデータの生成分布がわかれば，パターン認識ができる



- 生成的アプローチとよばれる

各タスクに特化したアプローチ 4

- しかし，確率分布の推定は困難であるため，生成モデル推定に基づくアプローチによって，必ずしも高い学習精度が得られるとは限らない
- 確率分布の推定を行わず，各タスクを直接解く
  - 例：サポートベクトルマシンでは，各クラスのデータ生成分布を推定せず，パターン認識に必要な決定境界のみを学習
  - パターン認識に対しては，識別的アプローチとよばれる

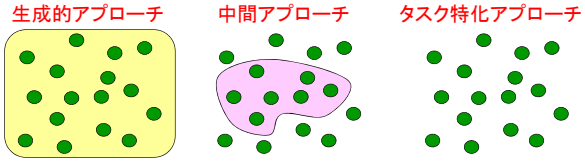


各タスクに特化したアプローチ 5

- 各タスクに特化したアルゴリズムを開発した方が原理的には生成的アプローチよりも性能が良い
- しかし，様々なタスクに対して個別に研究開発を行うのは大変：
  - アルゴリズム考案
  - 理論的性能評価
  - 高速かつメモリ効率の良い実装
  - エンジニアの技術習得

本日紹介するアプローチ 6

- 中間的なアプローチ：あるクラスのタスク群に対して，研究開発を行う
  - 確率密度比，確率密度差，距離，情報量，確率密度微分などの抽象的な量の推定を通して，データ解析を行う



確率密度比に基づく機械学習

7

- 前述の機械学習タスク群は複数の確率分布を含む

$$p(\boldsymbol{x}), q(\boldsymbol{x})$$

- しかし，これらのタスクを解くのに，それぞれの確率分布そのものは必要ない
- 確率密度関数の比が分かれば十分である

$$r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$$

- 各確率分布は推定せず，密度比を直接推定することにする

直感的な正当化

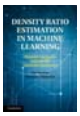
8

バプニックの原理 Vapnik (1998)

ある問題を解くとき，それより一般的な問題を途中段階で解くべきでない

$p(\boldsymbol{x}), q(\boldsymbol{x})$ が分かる ⇒ $r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$ が分かる

- 密度を求めるよりも，密度比を求めるほうが易しい



Sugiyama, Suzuki & Kanamori,
Density Ratio Estimation
in Machine Learning,
Cambridge University Press, 2012

発表の流れ

9

1. 密度比推定に基づく機械学習の枠組み
2. 密度比推定法
3. 密度比推定の応用事例
4. 発展的な話題

最小二乗密度比適合

10

Kanamori, Hido & Sugiyama (JMLR2009)

- データ: $\{\boldsymbol{x}_i^p\}_{i=1}^{n_p} \overset{i.i.d.}{\sim} p(\boldsymbol{x})$, $\{\boldsymbol{x}_j^q\}_{j=1}^{n_q} \overset{i.i.d.}{\sim} q(\boldsymbol{x})$
- 真の密度比$r(\boldsymbol{x})$との二乗誤差を最小にするように密度比モデル $r_{\boldsymbol{\alpha}}(\boldsymbol{x})$ を学習:

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\int \left(r_{\boldsymbol{\alpha}}(\boldsymbol{x}) - r(\boldsymbol{x})\right)^2 q(\boldsymbol{x}) d\boldsymbol{x} \qquad r(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$$

$$= \frac{1}{2}\int r_{\boldsymbol{\alpha}}(\boldsymbol{x})^2 q(\boldsymbol{x}) d\boldsymbol{x} - \int r_{\boldsymbol{\alpha}}(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} + C$$

$$\approx \frac{1}{2n_q}\sum_{j=1}^{n_q} r_{\boldsymbol{\alpha}}(\boldsymbol{x}_j^q)^2 - \frac{1}{n_p}\sum_{i=1}^{n_p} r_{\boldsymbol{\alpha}}(\boldsymbol{x}_i^p) + C$$

アルゴリズム

11

- 密度比モデル: $r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{\ell=1}^{n_p} \alpha_\ell \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_\ell^p\|^2}{2\sigma^2}\right)$
- 最適化規準: $\min_{\boldsymbol{\alpha}}\left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{G}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}\right]$

$$\widehat{G}_{\ell,\ell'} = \frac{1}{n_q}\sum_{j=1}^{n_q} \exp\left(-\frac{\|\boldsymbol{x}_j^q - \boldsymbol{x}_\ell^p\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\boldsymbol{x}_j^q - \boldsymbol{x}_{\ell'}^p\|^2}{2\sigma^2}\right)$$

$$\widehat{h}_\ell = \frac{1}{n_p}\sum_{i=1}^{n_p} \exp\left(-\frac{\|\boldsymbol{x}_i^p - \boldsymbol{x}_\ell^p\|^2}{2\sigma^2}\right)$$

- 大域的最適解が解析的に計算可能:

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{G}} + \lambda \boldsymbol{I})^{-1}\widehat{\boldsymbol{h}}$$

最小二乗密度比適合のMATLABによる実装

12

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{G}} + \lambda \boldsymbol{I})^{-1}\widehat{\boldsymbol{h}}$$

$$\widehat{G}_{\ell,\ell'} = \frac{1}{n_q}\sum_{j=1}^{n_q} \exp\left(-\frac{\|\boldsymbol{x}_j^q - \boldsymbol{x}_\ell^p\|^2}{2\sigma^2}\right)\exp\left(-\frac{\|\boldsymbol{x}_j^q - \boldsymbol{x}_{\ell'}^p\|^2}{2\sigma^2}\right)$$

$$\widehat{h}_\ell = \frac{1}{n_p}\sum_{i=1}^{n_p} \exp\left(-\frac{\|\boldsymbol{x}_i^p - \boldsymbol{x}_\ell^p\|^2}{2\sigma^2}\right)$$

```
%人工データの生成
n=300; x=randn(n,1); y=randn(n,1)+0.5;

%密度比の推定
x2=x.^2; xx=repmat(x2,1,n)+repmat(x2',n,1)-2*x*x';
y2=y.^2; yx=repmat(y2,1,n)+repmat(x2',n,1)-2*y*x';
r=exp(-yx); s=r*((r'*r+eye(n))¥(mean(exp(-xx),2))); plot(y,s,'rx');
```

理論解析

13

■ パラメトリックモデルの場合：$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \phi_\ell(\boldsymbol{x})$

- 学習したパラメータは $n^{-\frac{1}{2}}$ の速さで最適値に収束
- 最適な収束率を達成している $n = \min(n_p, n_q)$

Kanamori, Hido & Sugiyama (JMLR2009)

■ ノンパラメトリックモデルの場合：$r_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \sum_{\ell=1}^{n_p} \alpha_\ell \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}_\ell^p\|^2}{2\sigma^2}\right)$

- 学習した関数は $n^{-\frac{1}{2+\gamma}}$ の速さで真の関数に収束（関数空間のブラケットエントロピーに依存）
- 最適な収束率を達成している $0 < \gamma < 2$

Kanamori, Suzuki & Sugiyama (ML2012)

発表の流れ

14

1. 密度比推定に基づく機械学習の枠組み
2. 密度比推定法
3. 密度比推定の応用事例
4. 発展的な話題

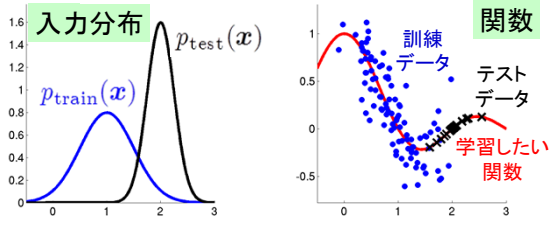
共変量シフト適応

15

■ 共変量とは入力変数の別名

■ 共変量シフト：訓練時とテスト時で入力分布が変化するが，入出力関数は変わらない

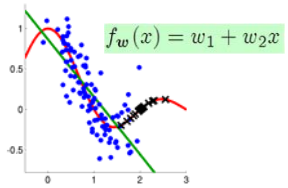
■ 外挿問題が典型的な例



重要度重み付き最小二乗学習

16

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} \left(f_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i\right)^2 \qquad \min_{\boldsymbol{w}} \sum_{i=1}^{n} \frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)} \left(f_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i\right)^2$$



■ 共変量シフト下では，通常の最小二乗学習は一致性を持たない（$n \to \infty$ でも最適解に収束しない）

■ 共変量シフト下でも一致性を持つ

■ 様々な学習法に適用可能：

- サポートベクトルマシン，ロジスティック回帰，条件付き確率場など

実世界応用例

17

■ 顔画像からの年齢予測：

- 照明環境の変化

Ueki, Sugiyama & Ihara (IEICE-ED2011)

■ 話者認識：

- 声質の変化

Yamada, Sugiyama & Matsui (SigPro2010)

■ テキスト分割：

- ドメイン適応

Tsuboi, Kashima, Hido, Bickel & Sugiyama (JIP2008)

■ ブレイン・コンピュータインターフェース：

- 心理状態の変化

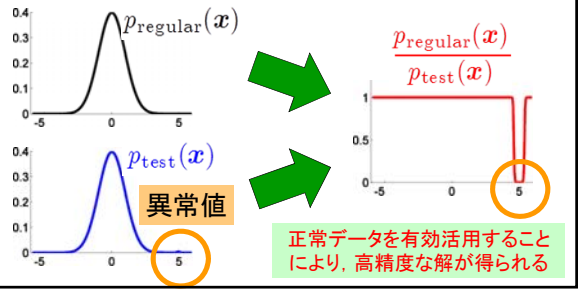
Sugiyama, Krauledat & Müller (JMLR2007)
Li, Kambara, Koike & Sugiyama (IEEE-TBE2010)

正常値に基づく異常値検出

18

Hido, Tsuboi, Kashima, Sugiyama & Kanamori (KAIS2011)

■ 正常データと傾向が異なるテストデータを異常値とみなす．



正常データを有効活用することにより，高精度な解が得られる

実世界応用例

19

■製鉄プロセスの異常診断

Hirata, Kawahara & Sugiyama (Patent2010)

■光学部品の品質検査

Takimoto, Matsugu & Sugiyama (DMSS2009)

■ローン顧客の審査

Hido, Tsuboi, Kashima, Sugiyama & Kanamori (KAIS2011)

二標本検定

20

Sugiyama, Suzuki, Ito, Kanamori & Kimura (NN2011)

■目的：二つのデータセットの背後の確率分布が同じかどうかを検定する

$$\{\boldsymbol{x}_i^p\}_{i=1}^{n_p} \overset{i.i.d.}{\sim} p(\boldsymbol{x})$$
$$\{\boldsymbol{x}_j^q\}_{j=1}^{n_q} \overset{i.i.d.}{\sim} q(\boldsymbol{x})$$

■アプローチ：密度比を用いて分布間の距離を推定する

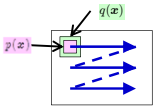
- カルバック・ライブラー距離：$\int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$
- ピアソン距離：$\int q(\boldsymbol{x}) \left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1 \right)^2 d\boldsymbol{x}$

実世界応用例

21

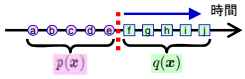
■画像中の注目領域抽出

Yamanaka, Matsugu & Sugiyama (IPSJ-TOM2013)



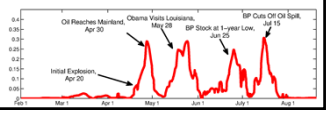
■動画からのイベント検出

Yamanaka, Matsugu & Sugiyama (IPSJ-TOM2013)



■ツイッターデータ解析

Liu, Yamada & Sugiyama (NN2013)



相互情報量推定

22

Suzuki, Sugiyama, Sese & Kanamori (FSDM2008), Sugiyama (Entropy2013)

■相互情報量：$\mathrm{MI} = \int p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x}) p(\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{y}$

$\mathrm{MI} = 0$ ⇔ $\boldsymbol{x}$ と $\boldsymbol{y}$ は統計的に独立

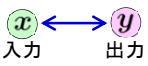
■相互情報量は密度比を用いて計算できる

■最小二乗密度比推定には，二乗損失相互情報量が自然：

$$\mathrm{SMI} = \int p(\boldsymbol{x}) p(\boldsymbol{y}) \left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x}) p(\boldsymbol{y})} - 1 \right)^2 d\boldsymbol{x} d\boldsymbol{y}$$

相互情報量に基づく機械学習

23



■入出力間の独立性判定：

- 特徴選択
- クラスタリング

Suzuki, Sugiyama, Sese & Kanamori (BMC-Bioinfo2009)

Suzuki & Sugiyama (NeCo2012)

Sugiyama, Niu, Yamada, Kimura & Hachiya (NeCo2013)

■実世界応用例：

- 遺伝子解析
- 画像認識
- 音響認識

相互情報量に基づく機械学習

24

■入力間の独立性判定：

- 独立成分分析
- オブジェクト適合

Suzuki & Sugiyama (NeCo2011)

Karasuyama & Sugiyama (NN2012)

Yamada & Sugiyama (AISTATS2011)

■実世界応用例：

- モーションキャプチャデータの解析
- 医療画像の位置合わせ
- 写真の自動レイアウト

条件付き確率密度の推定

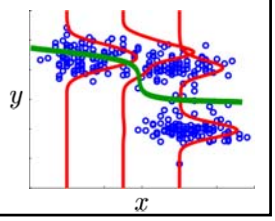
25

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})}$$

Sugiyama, Takeuchi, Suzuki, Kanamori, Hachiya & Okanohara (IEICE-ED2010)

- 回帰分析：条件付き期待値の推定
- 非対称なノイズや多峰性を持つようなデータに対しては，回帰分析では不十分
- 実世界応用例：
  - ヒューマノイドロボット制御

Sugimoto, Tangkaratt, Wensveen, Zhao, Sugiyama & Morimoto (HUMANOIDS2014)



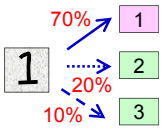
確率的パターン認識

26

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})}$$

Sugiyama (IEICE-ED2010)

- 出力 y がカテゴリのとき，条件付き確率の推定は確率的なパターン認識に対応



- 実世界応用例：
  - 顔画像からの年齢推定

    Ueki, Sugiyama, Ihara & Fujita (ACPR2011)
  - 加速度データからの行動認識

    Hachiya, Sugiyama & Ueda (Neurocomputing2012)

発表の流れ

27

1. 密度比推定に基づく機械学習の枠組み
2. 密度比推定法
3. 密度比推定の応用事例
4. 発展的な話題

発展的な話題

28

- ブレグマン距離を用いた密度比推定の統一理論

  Sugiyama, Suzuki & Kanamori (AISM2012)
- 次元削減付き密度比推定

  Sugiyama, Kawanabe & Chui (NN2010)
  Sugiyama, Yamada, von Bünau, Suzuki, Kanamori & Kawanabe (NN2011)
- 相対密度比推定

  Yamada, Suzuki, Kanamori, Hachiya & Sugiyama (NIPS2011, NeCo2013)

$$\frac{p(\boldsymbol{x})}{\beta p(\boldsymbol{x}) + (1-\beta) q(\boldsymbol{x})} < \frac{1}{\beta} \qquad 0 < \beta < 1$$

- 密度差推定

  Sugiyama, Suzuki, Kanamori, du Plessis, Liu & Takeuchi (NIPS2012, NeCo2013)

$$p(\boldsymbol{x}) - q(\boldsymbol{x})$$

密度比の世界

29

実問題応用例：
ブレイン・コンピュータインターフェース，ロボット制御，音声認識，画像認識，自然言語処理，バイオインフォマティクス，データマイニング

機械学習アルゴリズム：
重点サンプリング（共変量シフト適応，ドメイン適応，多タスク学習），
二標本問題（二標本検定，外れ値検出，変化点検知），
相互情報量推定（独立性検定，変数選択，独立成分分析，
次元削減，因果推定，クラスタリング，オブジェクト適合）
条件付き確率推定（可視化，状態遷移推定，確率的パターン認識），

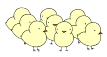
密度比推定法：
基本アルゴリズム（LR，KMM，KLIEP，LSIF），
大規模対応，高次元対応，安定化，ロバスト化，統一化

理論解析：
収束性解析（確率論），情報量規準（統計学），安定性解析（最適化）

まとめ

30

- 密度比は，単純な最小二乗法で精度・効率良く推定できる
- 多くの学習タスクが実は最小二乗法で解ける：
  - 重点サンプリング：$\sum_{i=1}^{n} \frac{p_{\mathrm{test}}(\boldsymbol{x}_i)}{p_{\mathrm{train}}(\boldsymbol{x}_i)} \mathrm{loss}(\boldsymbol{x}_i)$



  - ダイバージェンス推定：$\int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$

  - 相互情報量推定：$\iint p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x}) p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$
  - 条件付き確率推定：$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})}$

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-02

相関行列スケーリングを用いた屋外音源探索手法の解析

Analysis of Outdoor Sound Detection Using Correlation Matrix Scaling

大畑琢磨 [1]，長峰諒英 [2]，中村圭佑 [3]，石崎孝幸 [1]，水本武志 [3]，中臺一博 [1,3]

Takuma OHATA, Akihide NAGAMINE, Keisuke NAKAMURA,

Takayuki ISHIZAKI, Takeshi MIZUMOTO, Kazuhiro NAKADAI

1 東京工業大学 大学院 情報理工学研究科，2 東京工業大学 工学部 電気電子工学科,

3 (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 Department of Electric and Electrial Engineering, Tokyo Insistute of Technology,

3 Honda Research Institute Japan Co., Ltd.

ohhata@cyb.mei.titech.ac.jp, nakadai@jp.honda-ri.com

Abstract

我々は，屋外でクアドロコプタに搭載したマイクロホンアレイを用いて，雑音下でもロバストに音源定位を行うことができる MUSIC (MUltiple SIgnal Classification) ベースの手法について研究を行っている．これまでに，雑音相関行列の逐次推定や，相関行列のスケーリングといった拡張を施した CMS 付 iGSVD-MUSIC 法を提案し，良好な音源定位性能が得られることを示した．この手法は，理論的に雑音にロバストであることは知られているものの，実環境での挙動の解析が十分ではなく，どのような条件でロバストに動作するのか，パラメータ値の最適性についての議論することが難しかった．本稿では，提案手法の挙動をシュミレーション実験によって解析し，雑音環境の変化に対する最適パラメータ値の傾向について議論する．

1 はじめに

屋外で，音源の位置，種類，発生時刻といった音源に関する情報を抽出し，構造化する「屋外音環境理解」研究は，学術的な側面だけでなく，災害地での人命救助にも応用が可能な重要な研究領域である．特にクアドロコプタは，被災地でも広範囲に移動することが可能であり，制御の容易さから近年商用化も進んでいる．このため，クアドロコプタにマイクロホンアレイを搭載し，音源探索を行うことができれば，上述の場面での有用性が高いといえる．

従来，飛行体から音源探索を行う試みは，軍事用途を中心に行われてきたが，Acoustic Vector Sensor (AVS) などの高価なセンサが必要であったり，戦車や飛行機などパワーの大きな音源を対象にしていた [1]．我々は，マイクロホンアレイを用いた音源定位手法の中でも雑音に頑健であるとされる MUSIC (MUltiple SIgnal Classification) 法 [3] をベースにクアドロコプタのプロペラ音や風切り音が存在する屋外雑音下で，ロバストに音源定位ができる手法を報告した [4, 6, 7, 2]．例えば，奥谷らは，小型で軽量なマイクおよびマイク収録デバイスを用いて，コンシューマ向けのクアドロコプタである AR.Drone にマイクロホンアレイを搭載した [4]．また，プロペラ音が変化する雑音下でも雑音を適応的に白色化しながら，音源定位を行うことができる *Multiple Signal Classification based on incremental Generalized EigenValue Decomposition (iGEVD-MUSIC)* 法を提案し，その有効性を示した．さらに，iGEVD-MUSIC 法の計算量を削減するため，特異値展開に基づく MUSIC 法である GSVD-MUSIC [5] に対して，雑音の逐次推定機能を追加した *MUltiple SIgnal Classification based on incremental Generalized Sigular Value Decomposition (iGSVD-MUSIC)* 法を提案し，性能劣化を抑えつつ，計算量を劇的に削減できることを報告した [6]．また，iGSVD-MUSIC 法は，iGEVD-MUSIC 法と比較して，誤差項（クロスターム）が存在するため，性能劣化が起こりやすい，特に雑音相関行列の推定が正確でない場合，過抑圧が発生して検出性能が低下する場合があるという問題があった．そこで，相関行列スケーリング（Correlation Matrix Scalingm, CMS) 法を合わせて用いる iGSVD-MUSIC-CMS 法を提案し，この問題の解決を図った [2].

iGSVD-MUSIC-CMS 法により，過剰な雑音抑圧を防ぐことができ，定位性能は飛躍的に向上したものの，その挙動については未解明な部分も多く，このため，ロバストに定位を行うための条件や実験的に求めた最適なパラメータ値の妥当性を検証することは難しかった．そこで，本稿では，iGSVD-MUSIC-CMS 法と，その未解明な部

分について述べ，その挙動をシュミレーション実験によって解析し，雑音環境の変化に対する最適パラメータ値の傾向を議論する．

2 iGSVD-MUSIC-CMS 法

iGSVD-MUSIC-CMS 法の挙動を解析を行う前に，iGSVD-MUSIC-CMS 法の説明と課題の整理を行う．

2.1 iGSVD-MUSIC 法

iGSVD-MUSIC 法は，GSVD-MUSIC 法の雑音相関行列推定を逐次的に行うことができるように改良した手法である．これによって，少ない計算量で，動的な雑音が存在する環境でも頑健に音源定位を行うことが可能となる．以下に，そのアルゴリズムを説明する．

f フレーム目の M チャネル入力音響信号をフーリエ変換して得られる $\boldsymbol{X}(\omega, f) \in \mathbb{C}^M$ から，以下のように相関行列 $\boldsymbol{R}(\omega, f) \in \mathbb{C}^{M\times M}$ を定義する．

$$\boldsymbol{R}(\omega, f) = \frac{1}{T_R} \sum_{\tau=f}^{f+T_R-1} \boldsymbol{X}(\omega, \tau)\boldsymbol{X}^*(\omega, \tau) \qquad (1)$$

ただし，ω は周波数ビン番号，T_R は相関行列の計算に用いるフレーム数である．

MUSIC 法 [3] では，式 (1) の $\boldsymbol{R}(\omega, f)$ を以下のように標準固有値展開 (*Standard EigenValue Decomposition (SEVD)*) して，その固有ベクトルを音源定位に用いていた．

$$\boldsymbol{R}(\omega, f) = \boldsymbol{E}(\omega, f)\boldsymbol{\Lambda}(\omega, f)\boldsymbol{E}^*(\omega, f) \qquad (2)$$

ここで，$\boldsymbol{\Lambda}(\omega, f)$ は降順に並んだ固有値を対角成分に持つ行列であり，$\boldsymbol{E}(\omega, f)$ は固有ベクトルを並べた行列である ($\boldsymbol{E}(\omega, f) = [\boldsymbol{e}_1(\omega, \psi), \ldots, \boldsymbol{e}_M(\omega, \psi)]$)．しかし，この手法は目的音よりも大きな雑音がある場合は性能が著しく劣化する問題があった [5]（本手法をこれより SEVD-MUSIC 法と呼ぶこととする）．

そこで，GSVD-MUSIC 法では，f 番目のフレームに対して，f_s 前のフレームから，T_N フレーム分の信号は雑音区間であると仮定して，雑音の相関行列 $\boldsymbol{K}(\omega, f)$ を求める．

$$\boldsymbol{K}(\omega, f) = \frac{1}{\mathrm{T}_N} \sum_{\tau=f-f_s-\mathrm{T}_N}^{f-f_s} \boldsymbol{X}(\omega, \tau)\boldsymbol{X}^*(\omega, \tau)\ , \qquad (3)$$

GSVD-MUSIC 法は，雑音の相関行列には，与えられた雑音区間から事前に計算したものを使用しており，動的な雑音の変化に対応できないという問題があった．iGSVD-MUSIC 法では，フレームごとに（逐次的に）雑音が推定できるため，iGEVD-MUSIC 法と同様，動的な雑音変化に対応できることが期待できる．

$\boldsymbol{K}$ の逆行列を，左から $\boldsymbol{R}$ に掛けることで，雑音成分を白色化することが出来る．こうして得られた $\boldsymbol{K}^{-1}(\omega, f)\boldsymbol{R}(\omega, f)$ を一般化特異値展開し，左特異ベクトルを計算する．

$$\boldsymbol{K}^{-1}(\omega, f)\boldsymbol{R}(\omega, f) = \boldsymbol{E}_l(\omega, f)\boldsymbol{\Lambda}(\omega, f)\boldsymbol{E}_r^*(\omega, f) \qquad (4)$$

ただし，$\boldsymbol{\Lambda}(\omega, f)$ は降順に並んだ特異値を対角成分に持つ行列である．$\boldsymbol{E}_l(\omega, f)$, $\boldsymbol{E}_r(\omega, f)$ は，特異ベクトルを並べた行列である．

これと音源方向 ψ に対応した伝達関数 $\boldsymbol{G}(\omega, \psi)$ を用いて MUSIC スペクトル $P(\omega, \psi, f)$ を計算する．

$$P(\omega, \psi, f) = \frac{|\boldsymbol{G}^*(\omega, \psi)\boldsymbol{G}(\omega, \psi)|}{\sum_{m=L+1}^{M} |\boldsymbol{G}^*(\omega, \psi)\boldsymbol{e}_m(\omega, \psi)|} \qquad (5)$$

ただし，L は目的音源数，M はマイク数である．$\boldsymbol{e}_m$ は，$\boldsymbol{E}_l$ に含まれる m 番目の特異ベクトルを表す．音源方向を推定するために $P(\omega, \psi, f)$ を以下のように ω 方向に平均する．

$$\bar{P}(\psi, f) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} P(\omega, \psi, f) \qquad (6)$$

なお ω_H，ω_L は使用する周波数ビンの上限と下限に対応したインデックスである．

最後に，$\bar{P}(\psi, f)$ に対してピーク検出と閾値処理を行い，得られたピークに対する ψ を音源方向として検出する．

SEVD-MUSIC 法の拡張である GEVD-MUSIC 法では，式 (4) において，一般化特異値展開の代わりに一般化固有値展開を用いていた．しかし，$\boldsymbol{K}^{-1}(\omega, f)\boldsymbol{R}(\omega, f)$ は一般にエルミート行列ではないため，固有値ベクトル同士が直交するとはかぎらない．SEVD-MUSIC 法では式 (5) に示すようにベクトル同士が直交していることを利用したアルゴリズムであるため，性能劣化が生じる．そこで，GEVD-MUSIC 法では，この問題を解決するために，$\boldsymbol{K}^{-1}(\omega, f)\boldsymbol{R}(\omega, f)$ の代わりに，$\boldsymbol{K}^{\frac{1}{2}}(\omega, f)\boldsymbol{R}(\omega, f)\boldsymbol{K}^{\frac{1}{2}}(\omega, f)$ を用いている．しかし，この計算にかかる計算量が大きく，実時間処理が困難であった．

一方，GSVD-MUSIC 法では，非エルミート行列に対しても，特異ベクトル同士が直交することが保証されているため，この問題は生じない．このため，$\boldsymbol{K}^{\frac{1}{2}}$ を計算する必要がないこと，一般化特異値展開の計算量が一般化固有値展開のそれに比べて小さいことから，雑音ロバスト性能の劣化を抑えつつ，計算量を大きく削減できることが期待できる．

ここで，式 (1) の入力音響信号を次のように定義する（簡単のため，ω, f は省略する）．

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{N} \qquad (7)$$

$\boldsymbol{A} \in \mathbb{C}^{M\times L}$ は L 個の音源と M 個のマイクロホンアレイ間の伝達関数 ($\boldsymbol{A} = [\boldsymbol{A}_1(\psi_1), \ldots, \boldsymbol{A}_L(\psi_L)]$)，$\boldsymbol{S} \in \mathbb{C}^L$ は L 個の音源信号 ($\boldsymbol{S} = [S_1, \ldots, S_L]^T$)，$\boldsymbol{N} \in \mathbb{C}^M$ は雑音信

号を表している．$\boldsymbol{N}$ と $\boldsymbol{S}$ は無相関であると仮定すると $\boldsymbol{R}$ は以下のように変換できる．

$$\boldsymbol{R} = \boldsymbol{X}\boldsymbol{X}^* = \boldsymbol{A}\boldsymbol{S}\boldsymbol{S}^*\boldsymbol{A}^* + \boldsymbol{N}\boldsymbol{N}^* = \boldsymbol{\Gamma} + \boldsymbol{K} \tag{8}$$

iGEVD-MUSIC 法では，以下のように雑音が白色化されて $\boldsymbol{I}$ となる．

$$\begin{aligned} \boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{R}\boldsymbol{K}^{-\frac{1}{2}} &= \boldsymbol{K}^{-\frac{1}{2}}\left(\boldsymbol{\Gamma}+\boldsymbol{K}\right)\boldsymbol{K}^{-\frac{1}{2}} \\ &= \boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{\Gamma}\boldsymbol{K}^{-\frac{1}{2}} + \boldsymbol{I}. \end{aligned} \tag{9}$$

iGSVD-MUSIC 法は，式 (4) より，以下のように $\boldsymbol{R}^2$ と $\boldsymbol{K}^2$ を用いた一般化固有値問題とみなせる．

$$\begin{aligned} & \boldsymbol{K}^{-1}\boldsymbol{R} = \boldsymbol{E}_l\boldsymbol{\Lambda}\boldsymbol{E}_r^* \\ \Leftrightarrow\ & \boldsymbol{K}^{-1}\boldsymbol{R}(\boldsymbol{K}^{-1}\boldsymbol{R})^* = \boldsymbol{E}_l\boldsymbol{\Lambda}\boldsymbol{E}_r^*(\boldsymbol{E}_l\boldsymbol{\Lambda}\boldsymbol{E}_r^*)^*, \\ \Leftrightarrow\ & \boldsymbol{K}^{-1}\boldsymbol{R}^2\boldsymbol{K}^{-1} = \boldsymbol{E}_l\boldsymbol{\Lambda}^2\boldsymbol{E}_l^*. \end{aligned} \tag{10}$$

ここで，$\boldsymbol{E}_l$ が固有ベクトルとなっていることがわかる．

式 (10) は，式 (8) を用いて以下のように表せる．

$$\begin{aligned} \boldsymbol{K}^{-1}\boldsymbol{R}^2\boldsymbol{K}^{-1} &= \boldsymbol{K}^{-1}\left(\boldsymbol{\Gamma}+\boldsymbol{K}\right)\left(\boldsymbol{\Gamma}+\boldsymbol{K}\right)^*\boldsymbol{K}^{-1} \\ &= \boldsymbol{K}^{-1}\boldsymbol{\Gamma}^2\boldsymbol{K}^{-1} + \boldsymbol{K}^{-1}\boldsymbol{\Gamma} + \boldsymbol{\Gamma}\boldsymbol{K}^{-1} + \boldsymbol{I} \end{aligned} \tag{11}$$

式 (11) から雑音相関行列 $\boldsymbol{K}$ による白色化が実現されている（右辺第 4 項）ものの，式 (9) の白色化と比較すると，iGSVD-MUSIC 法の白色化は，右辺第 2，3 項が残ってしまい，完全な白色化が達成されない問題がある．

2.2 CMS

iGSVD-MUSIC 法では，雑音相関行列の推定に過去の入力音響信号を用いるため，実際に抑圧したい現時刻の雑音相関行列を完全に予測することは不可能である．実際の雑音相関行列と適合しない雑音相関行列を用いた場合，過抑圧が生じ，結果として定位性能が劣化する．CMS 法は，雑音相関行列が適合しない場合でも抑圧の程度を制御することにより，過抑圧を防ぐことができる．具体的には，雑音相関行列の値を固定し，雑音抑圧率のみを変化させるように雑音部分空間を制御する．iGSVD-MUSIC 法における式 (3) の $\boldsymbol{K}$ に対し，次のように固有値展開を行う．

$$\boldsymbol{K} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^* \tag{12}$$

ここで，$\boldsymbol{\Lambda}$ は固有値を含む対角行列，$\boldsymbol{E}$ は固有ベクトルを表す．$\boldsymbol{\Lambda}$ は各固有ベクトルのパワーを表し，$\boldsymbol{E}$ は雑音部分空間における各固有ベクトルの方向を表す．$\boldsymbol{\Lambda}$ を制御すれば，雑音部分空間の大きさのみを制御できることから，$\boldsymbol{K}^\alpha$ を以下のように定義する．

$$\boldsymbol{K}^\alpha = \boldsymbol{E}\boldsymbol{\Lambda}^\alpha\boldsymbol{E}^*, \tag{13}$$

$$\boldsymbol{\Lambda}^\alpha = \mathrm{diag}(\lambda_1^\alpha, ..., \lambda_M^\alpha) \tag{14}$$

図 1: マイクアレイ配置

ここで，α は CMS 法におけるスケーリングパラメータとする．CMS 法を用いた iGSVD-MUSIC 法では，式 (4) における $\boldsymbol{K}$ を $\boldsymbol{K}^\alpha$ とする．α が 1 のとき，$\boldsymbol{K}^\alpha$ は $\boldsymbol{K}$ となり，CMS 法を用いない iGSVD-MUSIC 法と同等となる．また，α が 0 のとき，$\boldsymbol{K}^\alpha$ は $\boldsymbol{I}$ となり，雑音抑圧を行わない SEVD-MUSIC 法と一致する．

我々はこれまで，iGSVD-MUSIC-CMS 法において，α は実験的に 0.5 付近が最適であるとの結果を得た [7]．しかし，$\boldsymbol{K}$ の推定誤差に対する α の最適値との関係はこれまで解析していなかった．

3 iGSVD-MUSIC-CMS 法の白色化性能解析

本稿では，2 章で述べた iGSVD-MUSIC 法の課題による性能への影響と，CMS 法の導入による効果を解析する．

2.1 章では，iGSVD-MUSIC 法での白色化（式 (11)）は，iGEVD-MUSIC 法での白色化（式 (9)）と比較して，クロスターム項が残るため，完全な白色化を達成するためには目的音源と雑音が無相関であることが求められることを述べた．また，2.2 章では，一般的に雑音相関行列は正しく推定することができないため，その誤差が定位性能を劣化してしまう問題について述べた．目的音源と雑音が無相関，かつ雑音相関行列が正しく推定されていれば，CMS での α は 1 であるべきであり，0.5 程度が最も性能が良いとする結果 [7] は，この仮定が成り立たなくなったためであると考えられる．

そこで，本稿では，以下をシミュレーション実験によって調べることで白色化性能解析を行う．

- 式 (11) のクロスターム項と白色化性能の評価
  1) 拡散性雑音：空間的白色雑音のみが存在する場合
  2) 方向性雑音：空間的有色雑音が存在する場合
- 雑音推定誤差と白色化性能の評価
  3) パワー誤差：雑音相関行列生成時の雑音源の大きさが，観測信号のそれと異なる場合
  4) 方向誤差：雑音相関行列生成時の雑音源の方向が，観測信号のそれと異なる場合

シミュレーション実験では，図 1 に示されるクアドロコプタに搭載された 16 チャネルのマイクアレイ（半径 0.37m）を想定し，伝搬波モデルを用いた幾何計算による伝達関数を生成して仮想的な目的音（白色雑音）と雑音（白色雑音）を使用することで評価した．入力音響信号は 16kHz, 16 ビットとし，音響信号処理のフレーム長とシフト長はそれぞれ，512, 160 サンプルとした．

評価では，0° 方向に目的音 (白色雑音 S_0) があるとし，上述の雑音や誤差を加えた．0° 方向に 1 つの目的音のみがある場合，式 (7) は以下で表される．

$$\boldsymbol{X} = \boldsymbol{A}_0(\psi_0 = 0^\circ)S_0 \tag{15}$$

この場合，雑音が存在しないため，この $\boldsymbol{X}$ から得られる相関行列 $\boldsymbol{R}$ を用いた SEVD-MUSIC 法 [3] は白色化処理を行わなくても，信号の部分空間が式 (2) の $\boldsymbol{e}_1$ として得られ，目的音方向に正しく定位することができる．この時の $\boldsymbol{e}_1$ を $\tilde{\boldsymbol{e}}_1$ とする．

式 (15) に雑音 $\boldsymbol{N}$ を加え，

$$\boldsymbol{X} = \boldsymbol{A}_0(\psi_0 = 0^\circ)S_0 + \boldsymbol{N} \tag{16}$$

とした相関行列を固有値展開すると，$\boldsymbol{e}_1 = \tilde{\boldsymbol{e}}_1$ となるとは限らないため，音源定位性能が劣化する．従って，式 (16) から得られる第一固有ベクトル $\boldsymbol{e}_1$ と $\tilde{\boldsymbol{e}}_1$ の内積を評価することで音源定位性能を評価できる．

iGEVD-MUSIC 法，iGSVD-MUSIC 法，iGSVD-MUSIC-CMS 法は，式 (16) の観測信号から得られる相関行列の第一固有（特異）ベクトル $\boldsymbol{e}_1$ が $\tilde{\boldsymbol{e}}_1$ となるように，雑音情報 $\boldsymbol{N}$ を用いて白色化を行う（式 (9)，式 (10)）．従って，以下から得られる第一固有（特異）ベクトル $\boldsymbol{e}_1$ と $\tilde{\boldsymbol{e}}_1$ の内積を評価することで各手法の白色化性能を評価できる．

- SEVD-MUSIC : $\boldsymbol{R}$ の第一固有ベクトル（白色化なし）
- iGEVD-MUSIC : $\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{R}\boldsymbol{K}^{-\frac{1}{2}}$ の第一固有ベクトル
- iGSVD-MUSIC : $\boldsymbol{K}^{-1}\boldsymbol{R}$ の第一特異ベクトル
- iGSVD-MUSIC-CMS : $\boldsymbol{K}^{-\alpha}\boldsymbol{R}$ の第一特異ベクトル

ここで，$\alpha = \{0.1, 0.2, \ldots, 0.9\}$ とした．

相関行列計算のための式 (1),(3) のパラメータ $T_R = T_N = 50$ とした．また，内積は各周波数毎に算出されるため，以下のように 500Hz $\leq \omega \leq$ 2800Hz の周波数帯で平均を取った．

$$\xi = \frac{1}{\omega_H - \omega_L + 1}\sum_{\omega=\omega_L}^{\omega_H} |\boldsymbol{e}_1^*(\omega)\tilde{\boldsymbol{e}}_1(\omega)| \tag{17}$$

S_0 と $\boldsymbol{N}$ の信号対雑音比 (SNR) を変化させた時に ξ が 1 に近い方が白色化性能が高いと言える．

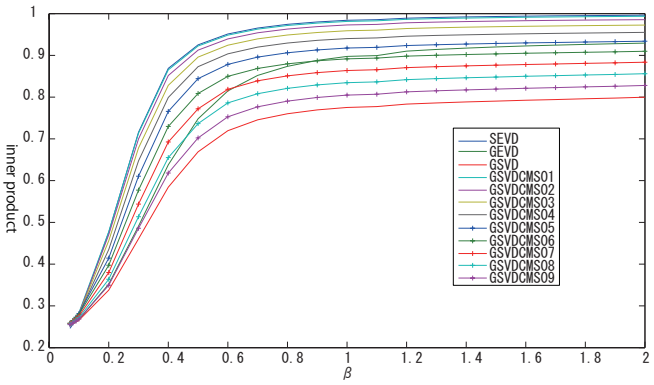


図 2: 拡散性雑音と ξ との関係

3.1 拡散性雑音に対する解析

式 (16) の $\boldsymbol{N}$ として，以下のように，空間的に白色な雑音を入力して評価する．

$$\boldsymbol{N} = \beta[S_1, \ldots, S_M]^T \tag{18}$$

ここで，$S_1, \ldots, S_M$ は互いに異なる白色雑音，β は S_0 と $\boldsymbol{N}$ の SNR を変化させるパラメータである．本稿では，$0.07 \leq \beta \leq 2.0$ とした．

空間的に白色な雑音のみが存在する場合，式 (11) の課題であったクロスターム項が白色化によって残らないことから，iGEVD-MUSIC 法と iGSVD-MUSIC 法の差異がないことが見込まれる．雑音相関行列 $\boldsymbol{K}$ は，

$$\hat{\boldsymbol{N}} = \beta[\hat{S}_1, \ldots, \hat{S}_M]^T \tag{19}$$

から生成し，$\boldsymbol{N}$ とは異なる白色雑音を用いた．

図 2 に，β の変化に対する，各手法から得られた式 (17) の ξ の変化を示す．凡例の **SEVD** は SEVD-MUSIC 法を，**GEVD** は iGEVD-MUSIC 法を，**GSVD** は iGSVD-MUSIC 法を，**GSVDCMS**** は iGSVD-MUSIC-CMS 法を表し，** は α の値を表す．

図より，全ての手法において，ξ の挙動が SEVD-MUSIC 法と類似していることがわかる．これは，S_0 と $\boldsymbol{N}$ が無相関であることと，$\boldsymbol{K} \approx \delta\boldsymbol{I}$（ただし，$\delta$ はスカラ）となっているからだと考えられる．この場合は 式 (8) の SEVD を用いた場合でも，

$$\boldsymbol{R} = \boldsymbol{\Gamma} + \boldsymbol{K} = \boldsymbol{\Gamma} + \delta\boldsymbol{I} \tag{20}$$

となり，$\boldsymbol{R}$ を固有値展開して得られる固有ベクトルが $\boldsymbol{\Gamma}$ を固有値展開して得られる固有ベクトルと等しい．

したがって，$\boldsymbol{N}$ が空間的に白色な場合の最適な α は 0 と結論づけられた．

3.2 方向性雑音に対する解析

次に，$\boldsymbol{N}$ が方向性雑音の場合を考える．具体的には，式 (16) に対して以下の雑音を考える．

- 単独雑音：120° 方向に白色雑音 S_1 が存在する

$$\boldsymbol{N} = \beta \boldsymbol{A}_1(120^\circ) S_1 \tag{21}$$

- 二雑音源：±90° 方向に白色雑音 S_1, S_2 が存在する

$$\boldsymbol{N} = \beta \boldsymbol{A}_1(90^\circ) S_1 + \beta \boldsymbol{A}_2(-90^\circ) S_2 \tag{22}$$

- 四雑音源：±45°, ±135° に白色雑音が存在する

$$\begin{aligned} \boldsymbol{N} \quad = \quad & \beta \boldsymbol{A}_1(135^\circ) S_1 + \beta \boldsymbol{A}_2(45^\circ) S_2 \\ & + \beta \boldsymbol{A}_3(-45^\circ) S_3 + \beta \boldsymbol{A}_4(-135^\circ) S_4 \end{aligned} \tag{23}$$

$0.07 \leq \beta \leq 2.0$ とした．

方向性雑音は空間的に有色な雑音であることから，式 (11) の課題であったクロスターム項の影響があると考えられ，CMS 法によってその誤差を吸収できるかを評価することができる．雑音相関行列には，式 (21), (22), (23) の $S_1, \ldots, S_4$ を $\hat{S}_1, \ldots, \hat{S}_4$ として相関行列を生成し，最後に逆行列が不安定とならないように $\delta \boldsymbol{I}$ を加えたものを用いた (δ は十分に小さい 10^{-4} とした)．

図 3, 4, 5 に，それぞれ単独雑音の場合，二雑音源の場合，四雑音源の場合の結果を示す．単独雑音の場合を見ると，方向性雑音のパワーが小さい $1 \leq \beta \leq 2$ では，$\alpha = 0.1$ が最も良い性能を示しており，パワーが大きくなるにつれ，$0.5 \leq \beta \leq 1$ では $\alpha = 0.2$ が，$0.3 \leq \beta \leq 0.5$ では $\alpha = 0.3$ が最も性能が良いことが確認できる．いずれも iGEVD-MUSIC 法や iGSVD-MUSIC 法よりも高い性能を示していることから，CMS 法を導入したことの有効性を確認することができた．また，方向性雑音のパワーが大きくなるにつれて最適な α が大きくなっていることから，雑音の空間的有色度を推定することで動的に α を変化させる適応的 CMS の可能性を確認できる．適応的 CMS については今後の課題とする．

次に二雑音源や四雑音源の場合を見ると，α が 0.4 や 0.5 の場合に最適な場合があることがわかる．このように環境の雑音有色度が増すほど，大きな α が最適であることがわかった．実環境下のクアドロコプタの場合，プロペラが 4 つあることから，四雑音源の場合に類似した環境であると考えられる．本稿の評価からも，クアドロコプタの環境において α が 0.4〜0.5 で最適であることの妥当性が示された．

3.3 パワー誤差に対する解析

2.2 章で述べた雑音相関行列の推定誤差について評価するため，雑音源のパワーに対する誤差について考える．雑音源は，3.2 章の単独方向性雑音と同じものを考えるが，雑音相関行列として，実際の雑音の 0.1 倍の雑音を以下のように考えた．

$$\hat{\boldsymbol{N}} = 0.1 \beta \boldsymbol{A}_1(120^\circ) \hat{S}_1 \tag{24}$$

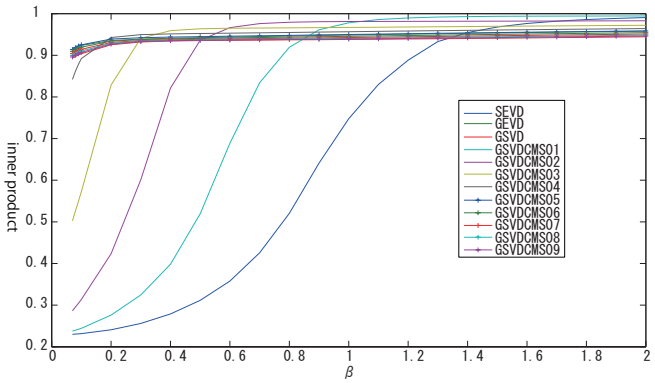


図 3: 方向性雑音と ξ との関係（単独雑音）

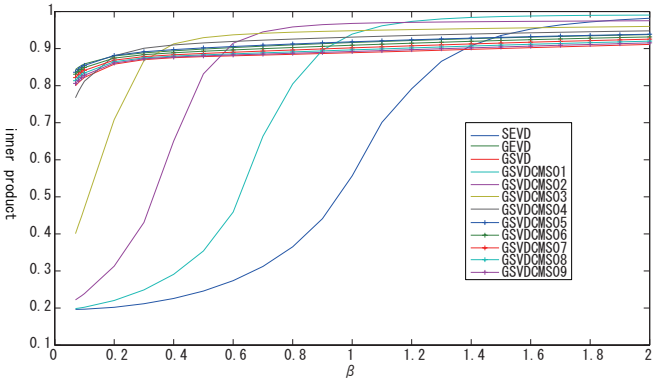


図 4: 方向性雑音と ξ との関係（二雑音源）

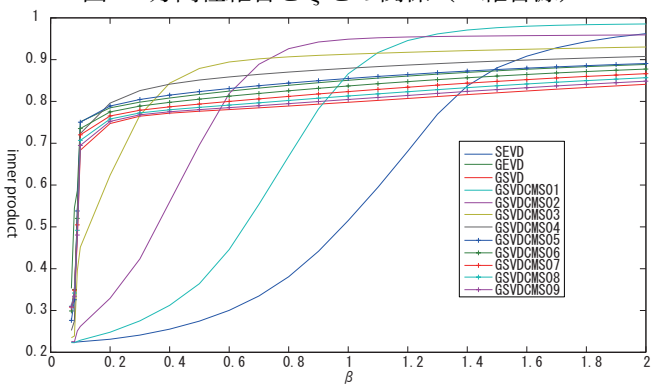


図 5: 方向性雑音と ξ との関係（四雑音源）

図 6 に結果を示す．図 3 と比較すると，最適な α がより大きい方向にシフトしていることがわかる．このように，雑音のパワーに対する推定誤差は α によって吸収できることがわかる．

図 7 は図 6 の $0.2 \leq \beta \leq 0.9$ 付近を拡大した図である．図より，SNR によって，最適な α が変化していること，また iGEVD-MUSIC 法や iGSVD-MUSIC 法よりもそれらが性能が高いことがわかる．従って，雑音のパワー推定誤差を含める範囲で iGSVD-MUSIC-CMS 法が有効であることがわかった．

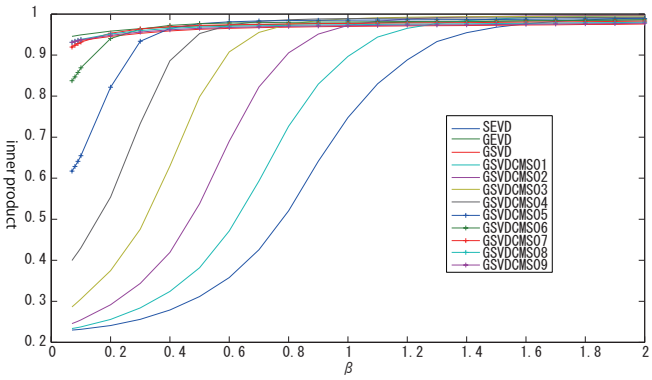


図 6: パワー誤差と ξ との関係

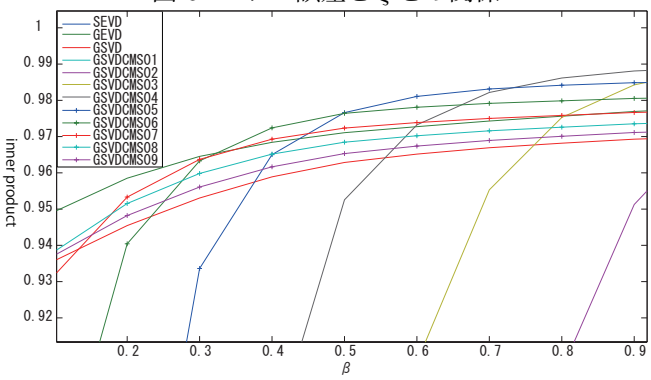


図 7: パワー誤差と ξ との関係（図 6 の拡大）

3.4 方向誤差に対する解析

最後に，雑音の方向に対する推定誤差について考える．雑音源は，3.2 章の単独方向性雑音と同じものを考えるが，雑音相関行列として，実際の雑音とは 5° 誤差のある雑音を以下のように考えた．

$$\hat{\boldsymbol{N}} = 0.1\beta \boldsymbol{A}_1(115°)\hat{S}_1 \tag{25}$$

図 8 に結果を示す．図より，全ての α について同様の白色化性能であることから，方向誤差は α で吸収できないことがわかる．しかし，これは雑音相関行列が 5° よりも細かな解像度であることを示唆しており，ターゲットとなる雑音方向に対してより急峻な白色化が達成できるといえる．一方，iGEVD-MUSIC 法は，5° の誤差に対して iGSVD-MUSIC-CMS 法よりも白色化性能が高かったことから，方向誤差に対するロバスト性が高いが，ターゲット方向に対して急峻な白色化は難しいことがわかった．このように，目的に応じた iGEVD-MUSIC 法と iGSVD-MUSIC 法の使い分けも興味深い今後の課題であると考える．

4 おわりに

本稿では，クアドロコプタのプロペラ音や風切り音が存在する屋外雑音下で，ロバストに音源定位ができる手法として提案していた CMS 付 iGSVD-MUSIC 法について，これまで未解明であったロバストに定位を行うための条件

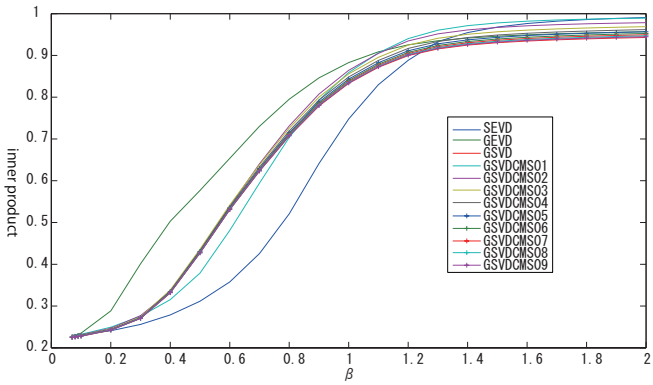


図 8: 方向誤差と ξ との関係

や実験的に求めた最適なパラメータ値の妥当性についてシミュレーション実験を通して議論を行った．結果，雑音の空間的な有色度と最適なパラメータ値に相関があったこと，以前に報告した最適パラメータがクアドロコプタの持つ４つの方向性雑音に対して妥当であったこと，雑音のパワーの推定誤差に対して CMS 法がロバストであったこと，雑音方向に対して iGSVD-MUSIC 法が既存法よりもより急峻な白色化が達成できることが示された．今後の課題として，クアドロコプタの実環境雑音データの有色度の検証，雑音の有色度を動的に推定して CMS 法のパラメータを適応的に変化する適応的 CMS 法の構築，方向性の点雑音・面雑音などの雑音の空間的広がりに合わせた iGSVD-MUSIC 法と iGEVD-MUSIC 法の使い分けなどが考えられる．

謝辞

本研究は科研費基盤 (S) No.24220006 の支援を受けた．

参考文献

[1] B. Kaushik, D. Nance, and K. K. Ahuj. A review of the role of acoustic sensors in the modern battlefield. In *11th AIAA/CEAS Aeroacoustics Conference (26th AIAA Aeroacoustics Conference)*, pp. 1–13, 2005.

[2] Takuma Ohata, Keisuke Nakamura, Takeshi Mizumoto, Taiki Tezuka, and Kazuhiro Nakadai. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array. In *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*. IEEE Press, 2014.

[3] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. on Antennas and Propagation*, Vol. 34, No. 3, pp. 276–280, 1986.

[4] 奥谷啓太, 吉田尚水, 中村圭佑, 中臺一博. クワドロコプタ搭載のマイクロホンアレイを用いた屋外音環境理解の逐次雑音推定による向上. ロボット学会誌, Vol. 31, No. 7, pp. 38–45, 2013.

[5] 中村圭佑, 中臺一博, インジュギョカン. ロボットを対象にした複数同時発話にロバストな音源定位の検討. 第 29 回日本ロボット学会学術講演会. 日本ロボット学会, 2011.

[6] 大畑琢磨, 手塚太貴, 中村圭佑, 水本武志, 中臺一博. クアドロコプタを用いた屋外音環境音源探索. 第 14 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. 0360–0363. 計測自動制学会, 2013.

[7] 大畑琢磨, 長峰諒英, 中村圭佑, 水本武志, 中臺一博. 相関行列スケーリングを用いた igsvd-music 法による屋外環境音源探索の向上. 日本ロボット学会第 32 回学術講演会, pp. 1I1–03, 2014.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-03

屋外音環境理解における音源検出の性能評価と可視化

Visualization of Sound Detection for Outdoor Scene Analysis

長峰 諒英 †，大畑 琢磨 ‡，上村 知史 ‡，小島 諒介 ‡，杉山 治 ‡，中村 圭佑 *，中臺 一博 ‡,*

Akihide Nagamine†, Takuma Ohata‡, Satoshi Uemura‡,
Ryosuke Kojima‡, Osamu Sugiyama‡, Keisuke Nakamura*, Kazuhiro Nakadai‡,*

† 東京工業大学 工学部 電気電子工学科, ‡ 東京工業大学 大学院 情報理工学研究科,
* (株) ホンダ・リサーチ・インスティチュート・ジャパン

†Department of Electric and Electrial Engineering, Tokyo Insistute of Technology,

‡Graduate School of Information Science and Engineering, Tokyo Insistute of Technology,

*** Honda Research Institute Japan Co., Ltd.**

Abstract

本稿では，屋外での音環境理解を目指して，クアドロコプターに搭載したマイクロホンを用いた音源定位を扱う．これまで，プロペラ音や風切り音が存在する環境下で音源を定位する手法を開発したが，1) 方位角のみを扱っていた，2) 音源検出結果を表示するビューアがなく直感的に状況がわかりにくかったという問題があった．本稿ではこれらの問題の解決を図るため，1 つ目の問題に対しては，仰角の定位を行うことができるように拡張するとともに，音源が地上付近にあることを仮定して，音源までの距離推定を行う．これによって，方位角，仰角，距離情報からなる 3 次元定位を可能にした．2 つ目の問題については，クアドロコプタのセンサから得られる 3 次元位置データ，および 3 次元音源定位結果を用いてマイクロホンアレイが 3 次元的に移動する場合でも，これらを 3 次元マップ上に表示するツールの開発を行った．これらを実装したプロトタイプシステムを構築し，3 種類の実機を用いて，実際に屋外で 21 種類の音源を用いた収録を行った．提案する 3 次元定位手法を，実機ベース，および音源ベースの指標で評価し，その有効性を示すとともに，ケーススタディベースで音源の直感的な可視化が実現できることを示した．

1 はじめに

屋外環境での音環境理解は，災害地での救助活動や異常音検出など様々な応用が期待できる有用な分野である．内閣府の革新的研究開発推進プログラム（ImPACT）では，極限災害環境でもタフに仕事ができる遠隔自律ロボットの実現を目指す「タフ・ロボティクス・チャレンジ」がプロジェクトとして採択され[1]，屋外ロボットの基盤技術への重要性が認知されてきている．屋外環境での音環境理解は，タフ・ロボティクス・チャレンジでも，極限音響という重要なテーマとして位置づけられている．

我々は，こうしたプロジェクトに先駆け，これまでに培ってきたロボット聴覚技術を用いて，屋外環境理解実現に向けた研究を行っている．ロボット聴覚は，主に屋内のロボットを対象にして，人とのインタラクションをロボットに備えた耳を用いて実現することを目的とした日本発の研究分野である[Nakadai 00]．ロボットの耳で音を聞く場合は，スマートホンの場合とは異なり，遠隔からの発話を認識する必要があるため，様々な雑音を扱う必要がある．そこで，マイクロホンアレイ処理を導入して，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[Nakamura 09, Nakajima 10, Yamamoto 07]．また，ロボット聴覚で培ってきた技術をロボット聴覚のオープンソースソフトウェア HARK (HRI-JP Audition for Robot with Kyoto University) として，一般公開を行っている．

1.1 屋内と屋外音環境理解の違い

屋内と屋外では，前述の雑音問題の性質が異なるため，同じ雑音抑圧技術でもそのフォーカスは異なる．屋内では，周囲の騒音と共に，残響が存在する（もしくは，残響を考慮する必要がある）ことが大きな特徴である．特に，音声認識が残響に対しての頑健性が低いという特徴を持っていることから，音声認識では残響が大きな問題である．一般的な屋内では壁，天井，床など音を反射するものに囲まれていることから残響を避けることは難しく，国際学会でも Reverb Challenge のような残響抑圧技術を競うコンペティションが行われている[2]．一方で，残響は，屋内の音響環境に関する情報が含まれている．例えば，方位角や仰角推定と比較すれば，音源距離推定の精度は低いものの，残響情報を積極的に利用することで音源距離推定が可能であることが報告されている[丹羽 14]．

屋外では，特殊な状況を除けば，一般に残響を考慮する必要はないといえる．これは，残響を扱う必要がない反面，屋外での音源距離推定が難しいことを示している．また，周囲の雑音が大きなダイナミックレンジで，動的に変化する．風，湿度，温度の変化があるため，音速自体が一様ではないばかりか，時間的にも変動する．点音源を仮定できない雑音源も多く存在し，そのモデル化も困難であるといった厄介な特徴を持っている．

[1] http://www.jst.go.jp/impact/program07.html

[2] http://reverb2014.dereverberation.com/

1.2 屋外音環境理解の関連研究

我々は，これまで，屋外音環境特有の問題を解決するため，音源定位にフォーカスして研究を行っている．例えば，奥谷らは，屋内の音源定位用に開発した一般化固有値展開に基づく GEVD-MUSIC (MUltiple SIgnal Classification based on Generalized EigenValue Decomposition) 法 [Nakamura 09] を時間的に変動する雑音に対応するように拡張した iGEVD-MUSIC (incremental GEVD-MUSIC) 法を報告した[奥谷 13]．ベースとなった GEVD-MUSIC 法は，クアドロコプタで事前収録した音響信号を用いて，雑音に関する知識である雑音相関行列の推定を行うため，上述のようにモデル化が難しい雑音源であっても精度よく推定できたが，動的に変化する雑音に対応することは難しかった．iGEVD-MUSIC 法における雑音相関行列の推定は，短時間での雑音は定常であるという仮定の下，対象区間より，時間的に少し前の時刻の音響信号を用いて雑音相関行列の推定を行うため，雑音相関行列を動的に推定することができ，屋外での音源検出性能を著しく向上できる．古川らは，この考え方をさらに発展させて，クアドロコプタ自身が作り出す雑音の変化に対応するため，クアドロコプタのステータス情報に対してガウス過程を用いることにより，雑音相関行列を動的に推定する手法を報告している [Furukawa 13]．大畑らは，GEVD の計算コストを削減するために，一般化特異値展開 GSVD (Generalized Singular Value Decomposition, GSVD) を導入した iGSVD-MUSIC 法を提案した[Ohata 14]．また，さらに雑音相関行列の推定誤差に対応するため，雑音相関行列の大きさをスケーリングできる CMS (Correlataion Matrix Scaling) 法を併せて用いることを提案した[大畑 14]．これらの手法を用いた結果，音声では 15 m 程度，ホイッスルなど検出しやすい音源では 20 m 程度遠方の音源でも精度良く検出をできることを示した．このように，要素技術としては，屋外環境に耐えうる音源定位技術が構築されつつある．

1.3 課題とアプローチ

しかし，こうした技術の実用化を考えた場合，以下のような課題を解決する必要がある．

1. 屋外は三次元環境であるにもかかわらず，一次元（方位角）のみの音源定位を扱っていた．
2. 音源検出結果を表示するビューアがなく直感的に状況がわかりにくかった．

本稿では，これらの問題の解決を図るため，1 つ目の問題に対しては，仰角の定位を行うことができるように拡張する．また，方位角と仰角平面上に音源探索を頑健に行うことができる音源探索法を提案する．さらに，音源が地上付近にあることを仮定して，音源までの距離推定を行う．これによって，方位角，仰角，距離情報からなる 3 次元定位を可能にした．2 つ目の問題については，クアドロコプタのセンサから得られる 3 次元位置データ，および 3 次元音源定位結果を用いてマイクロホンアレイが 3 次元的に移動する場合でも，これらを 3 次元マップ上に表示するツールの開発を行った．

2 音源定位手法

本稿では，オフラインでの評価を前提としていることから，MUSIC 法の中で性能がもっともよい iGEVD-MUSIC 法をベースに定位を行う．

2.1 iGEVD-MUSIC 法

iGEVD-MUSIC 法は，GEVD-MUSIC 法の雑音相関行列推定を逐次的に行うことができるように改良した手法である．これによって，動的な雑音が存在する環境でも頑健に音源定位を行うことが可能となる．以下に，そのアルゴリズムを説明する．

f フレーム目の入力音響信号をフーリエ変換して得られる $X(\omega, f)$ から，以下のように相関行列 $R(\omega, f)$ を定義する．

$$\boldsymbol{R}(\omega, f) = \frac{1}{T_R} \sum_{\tau=f}^{f+T_R-1} \boldsymbol{X}(\omega, \tau)\boldsymbol{X}^*(\omega, \tau) \qquad (1)$$

ただし，ω は周波数ビン番号，T_R は相関行列の計算に用いるフレーム数である．

次に，f 番目のフレームに対して，f_s 前のフレームから，T_N フレーム分の信号は雑音区間であると仮定して，雑音の相関行列 $\boldsymbol{K}(\omega, f)$ を求める．

$$\boldsymbol{K}(\omega, f) = \frac{1}{\mathrm{T}_N} \sum_{\tau=f-f_s-\mathrm{T}_N}^{f-f_s} \boldsymbol{X}(\omega, \tau)\boldsymbol{X}^*(\omega, \tau)\,, \qquad (2)$$

GEVD-MUSIC は，雑音の相関行列には，与えられた雑音区間から事前に計算したものを使用しており，動的な雑音の変化に対応できないという問題があった．iGEVD-MUSIC 法では，フレームごとに（逐次的に）雑音が推定できるため，動的な雑音変化に対応できることが期待できる．

$\boldsymbol{K}$ の逆行列を用いて，以下のように雑音成分を白色化することが出来る．こうして得られた $\boldsymbol{K}^{-\frac{1}{2}}(\omega, f)\boldsymbol{R}(\omega, f)\boldsymbol{K}^{-\frac{1}{2}}(\omega, f)$ を一般化固有値展開し，固有ベクトルを計算する．

$$\boldsymbol{K}^{-\frac{1}{2}}(\omega, f)\boldsymbol{R}(\omega, f)\boldsymbol{K}^{-\frac{1}{2}}(\omega, f) = \boldsymbol{E}(\omega, f)\boldsymbol{\Lambda}(\omega, f)\boldsymbol{E}^*(\omega, f) \qquad (3)$$

ただし，$\boldsymbol{\Lambda}(\omega, f)$ は降順に並んだ固有値を対角成分に持つ行列である．$\boldsymbol{E}(\omega, f)$ は，固有値ベクトルを並べた行列である．

これと音源方向 $\boldsymbol{\psi}$ に対応した伝達関数 $\boldsymbol{G}(\omega, \boldsymbol{\psi})$ を用いて MUSIC 空間スペクトル $P(\omega, \psi, f)$ を計算する．

$$P(\omega, \boldsymbol{\psi}, f) = \frac{|\boldsymbol{G}^*(\omega, \boldsymbol{\psi})\boldsymbol{G}(\omega, \boldsymbol{\psi})|}{\sum_{m=L+1}^{M} |\boldsymbol{G}^*(\omega, \boldsymbol{\psi})\boldsymbol{e}_m(\omega, \boldsymbol{\psi})|} \qquad (4)$$

ただし，L は目的音源数である．$\boldsymbol{e}_m$ は，$\boldsymbol{E}_l$ に含まれる m 番目の特異値ベクトルを表す．音源方向を推定するために $P(\omega, \psi, f)$ を以下のように ω 方向に平均する．

$$\bar{P}(\boldsymbol{\psi}, f) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} P(\omega, \boldsymbol{\psi}, f) \qquad (5)$$

なお ω_H，ω_L は使用する周波数ビンの上限と下限に対応したインデックスである．

最後に，$\bar{P}(\boldsymbol{\psi}, f)$ に対してピーク検出と閾値処理を行い，得られたピークに対する $\boldsymbol{\psi}$ を音源方向として検出する．

2.2 仰角推定と２次元音源探索手法

一般に，音源方向 ψ としては，方位角 θ のみを扱うことが多く，屋内では，このような 1 次元定位でも比較的問題になることが少ない．しかし，クアドロコプタなど屋外での音源定位を前提にする場合には，仰角に対する定位能力が求められる．そこで，本稿では，以下のように 2 次元に拡張して，定位を行う．

$$\boldsymbol{\psi} = (\theta, \phi) \tag{6}$$

このような定義を行っても，上述の MUSIC アルゴリズム自体は基本的に一般性を失わない．ピーク検出についても，θ 直線上ではなく，$\theta - \phi$ 平面上で行う必要がある．実際には，ピーク検出の問題はそれほど簡単な問題ではないが，本稿では，以下のように，音源数が，高々 1 であると仮定し，単純な最大値検出によって，ピークを検出し，定位を行った．

$$\boldsymbol{\Psi}(f) = \begin{cases} \mathrm{argmax}_{\boldsymbol{\psi}} \bar{P}(\boldsymbol{\psi}, f) & (\bar{P}(\boldsymbol{\Psi}, f) \geq P_{th}) \\ \emptyset & (otherwise) \end{cases} \tag{7}$$

P_{th} は音源かどうかを判断するための閾値であり，実験的に求めた．

2.3 音源距離推定

上述のように，音源方向は極座標表現 $\boldsymbol{\Psi}(f) = [\Theta(f), \Phi(f)]$ として得られる．これは，xyz 軸からなる直交座標系では 3 次元表現になるため，方位角と仰角からなる音源方向推定はしばしば 3 次元音源定位と呼ばれることがある．しかし，方位角と仰角の 2 次元の情報しか含まれていないため，実際には，3 次元音源定位とは言えない．真に 3 次元音源定位を実現するためには，音源までの距離情報を推定する必要があり，屋外音環境理解では，マップ上に音源表示を行うためにも距離情報まで推定できることが望ましい．しかし，本稿の冒頭に述べたように一般に音源距離推定問題は難しい．さらに，屋外では距離推定の重要なキューとなる残響情報の利用が困難であるため，音源距離推定問題は一層難しい問題となっている．

そこで，本稿では，音源は地上付近（主に人間の口元の高さ）にあるという仮定を置くことによって，この問題の解決を試みる．

まず，得られる音源方向は，クアドロコプター座標系での値になっているため，航法データを用いて，絶対座標への変換を行い，絶対座標系での方位角と仰角のペア $[A, E]$ を得る．

クアドロコプタの地表からの高度を h，音源の高度を h_{src} とすれば，音源距離は，以下のようにあらわすことができる．

$$D = \left| \frac{h - h_{src}}{\sin(E)} \right| \tag{8}$$

従って，クアドロコプターの中心を原点にとれば，3 次元音源位置は，以下のように表すことができる．

$$\begin{aligned} \boldsymbol{P_s} &= [A, E, D] && \text{(極座標系)} && (9) \\ &= [D\cos(E)\cos(A), -D\cos(E)\sin(A), D\sin(E)] && (10) \\ & && \text{(直交座標系)} \end{aligned}$$

3 音源可視化システム

得られた音源定位結果をクアドロコプタの航法データや地図データとともに可視化を行うシステムを構築した．Fig. 1 に構築した音源可視化システムの構成図を示す．我々が利用しているクアドロコプタである Asctec 社の Pelican は，ジャイロ，高度センサ，GPS，加速度センサ，磁気センサを搭載しており，位置，姿勢，速度，加速度が取得できる．これらに加えて，システムインフロンティア社の多チャンネル収録装置 RASP-24 と MEMS マイクロホンで構成される小型軽量の 16ch マイクロホンアレイを設置した（Fig. 2a) 参照)．クアドロコプタ搭載センサからの情報，およびマイクロホンアレイからの音響信号は WiFi (IEEE 802.11ac) 経由でデータ収録用の端末に送信される．この際，センサデータを同期収録する必要があるため，ROS [3] を用いて実現した．端末側では，受信した信号のうち，音響信号は，2 節で説明した音源定位手法を用いて，定位を行う．実装は HARK[4] を用いた．得られたクアドロコプタ極座標系での 2 次元の音源定位情報とクアドロコプタの情報を用いて絶対座標系での音源位置を算出し，KML (Google Earth(Keyhole) Markup Language) 形式に変換後，Google Map 上にこれらのデータを表示する．また，予め人の位置がわかっている場合には，その音源位置を登録し，その位置に人オブジェクトを表示しておくことができる．実際に，登録した人位置に音源があるとシステムが判断した場合には，これを人の発話と見なし，表示した人オブジェクトの色の変更を行う．

4 評価実験

構築したシステムの評価実験を行うため，実際に屋外で 21 種類の音源をスピーカから出力し，音源定位実験を行った．クアドロコプタには，Asctec 社の Pelican (Fig. 2a) 参照)，enRoute 社の Zion (Fig. 2b) 参照) を用いた．また，ヘリウムガスを入れたバルーンの周囲に 16 ch マイクロホンアレイを設置して，これを浮遊させ，クアドロコプタと同様の実験を行った (Fig. 2c) 参照)．

4.1 実験条件

実験の測定条件について，Tab. 1 にまとめる．「固定」は，屋外測定ではあるが，筐体をしっかり固定し，プロペラが回転しても動かない状態で収録を行った．ただし，バルーンは，固定しても風で流されてしまうため，完全な固定はできなかった．「移動」は，実際にクアドロコプタを浮遊させホバリングに近い動作を行った状態で収録を行った．固定条件と比べれば，風の影響が大きくなり，また，プロペラ音の動的な変化への対応が必要となる．音源の位置に関しては，大まかな方向は得られるものの正確なリファレンスを得ることは困難であった．

使用した 21 種類の音源，およびその音量を Fig. 3 にまとめた．音量は，wav ファイルの最大値を 0 dB として算出している．音量は一つの目安ではあるが，音源毎に周波数特性が異なるため，音源定位のしやすさと完全な相関はない．MUSIC に用いる伝達関数については，実測ではなく，幾何計算で算出した．MUSIC の処理で用いる音源数 L は 1 とした．

[3] http://www.ros.org/
[4] http://www.hark.jp/

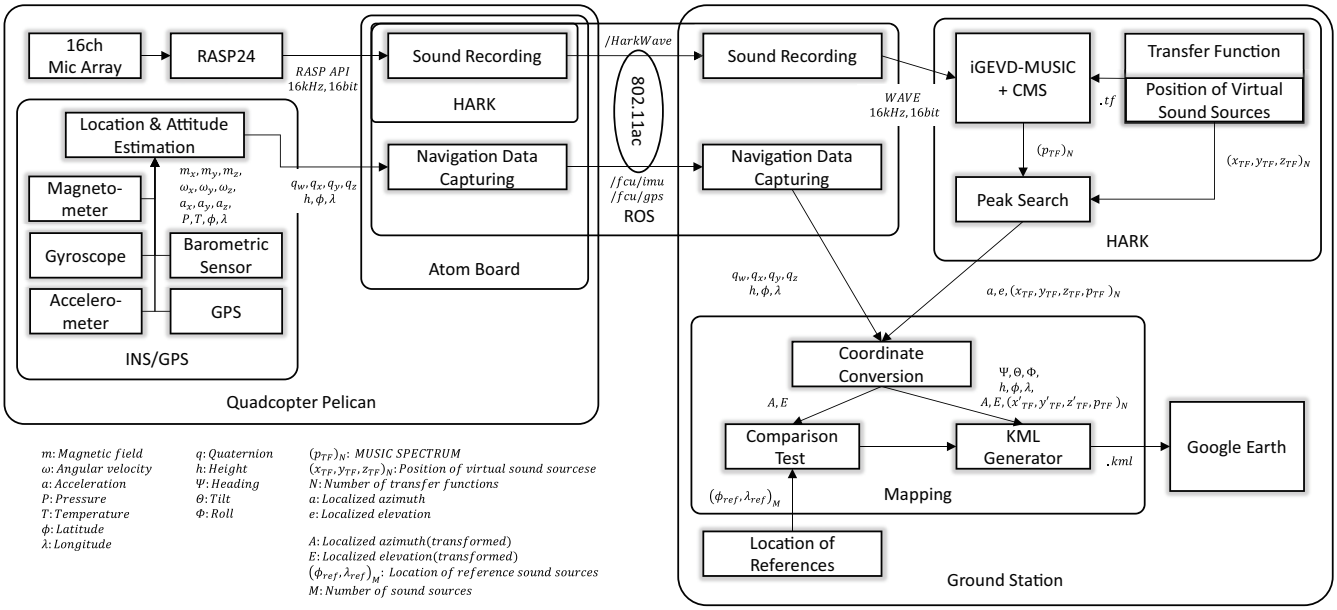
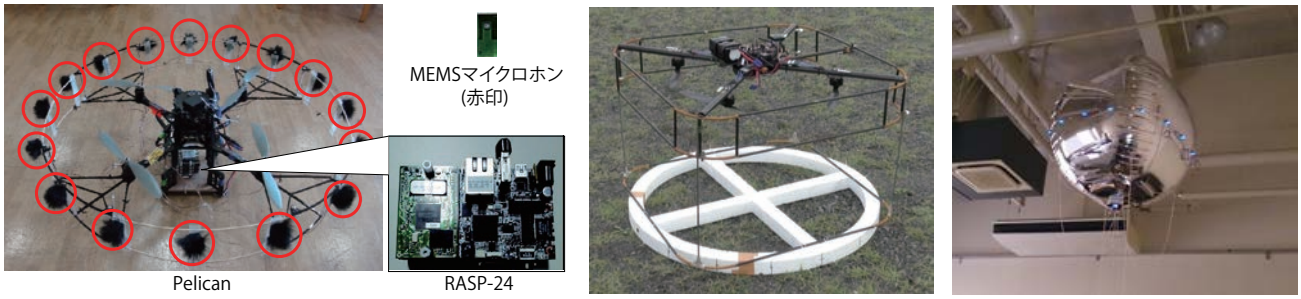


Figure 1: 可視化システム構成図



a) Pelican:丸印の位置にマイクロホン設置，風防付

b) Zion: 下部に円状発泡スチロールを取り付けマイクロホン設置

c) バルーン: 青く光っている点がマイクロホン

Figure 2: マイクロホンアレイ搭載クアドロコプタ

Table 1: 実験条件 (移動条件の高度，距離，角度は目安)

ラベル	クアドロコプタ		音源方向		使用音源	
	高度 [m]	水平距離 [m]	仰角 [deg]	方位角 [deg]	音源種類	測定回数 (音源毎)
バルーン固定	0	3	0	65	20	10
Pelican 固定	0	3	0	0	21	10
Zion 固定	0	3	0	0 – 360 (45 度毎)	1	10
Pelican 移動 A	5	3	60	0	7	3 – 10
Pelican 移動 B	5	5	45	0	7	3 – 10
Pelican 移動 C	5	10	27	0	7	3 – 10

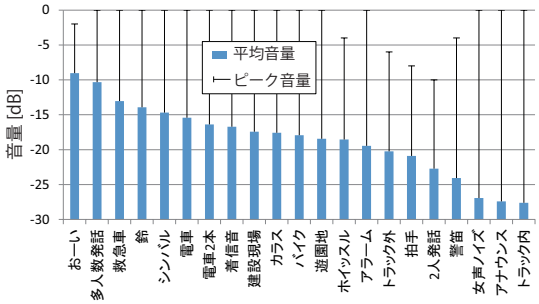


Figure 3: 使用音源の種類と音量

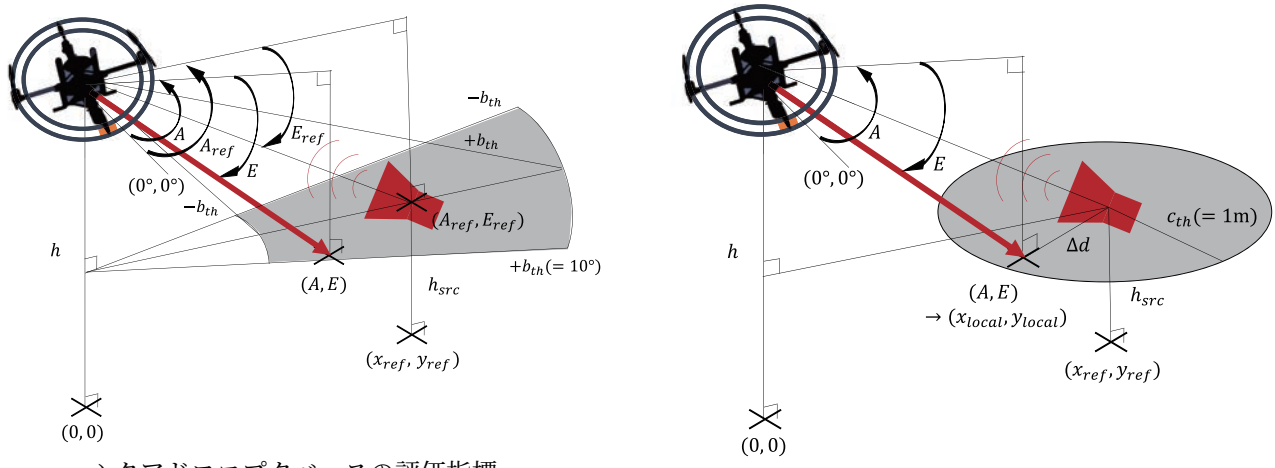
4.2 評価指標

音源定位の評価には，以下の 3 つの指標を用いた．

指標 1: 軸別定位正解精度
指標 2: クアドロコプタベース定位正解率
指標 3: 音源ベース定位正解率

指標 1 は，全音源数を N とした場合に，方位角，仰角別にクワドロコプタから見て，一定の角度 (a_{th}) 以内に定位した正解数 (C)，定位はしているものの角度が a_{th} 以内でない定位誤り数 (S)，音源の定位そのものがされなかった削除誤り数 (D), よけいに定位をしてしまう挿入誤り数 (I) をカウントし，$(N - S - D - I)/N$ を計算した値となる．$C = N - S - D$ であるので，I が多ければこの指標は負の値を持つ．奥谷らが用いた LAR (Localization Accuracy Rate) [奥谷 13] と同様の指標である．本稿では，a_{th} として，方位角に対しては，5°，仰角に対しては，10° を用いた．

指標 2 は，クワドロコプタから見て，実際の音源位置の一定角度 (b_{th}) 以内に定位しているかどうかを示す指標であり，**指標 1** と同様，クワドロコプタと音源の距離によ



a) クアドロコプタベースの評価指標　　b) 音源ベースの評価指標

Figure 4: 実験で用いた評価指標

らず精度が変わらない指標である．奥谷らが用いた LCR (Localization Correct Rate)[奥谷 13] に倣い，**指標 1** の C/N に相当する値とした．具体的な正解の判定条件は，Fig. 4a) における網掛け部分，つまり，以下の 2 つの条件を同時に満たす場合とした．

$$|A - A_{ref}| \leq b_{th} \quad (11)$$
$$|E - E_{ref}| \leq b_{th} \quad (12)$$

本稿では b_{th} として，10° を用いた．

指標 3 は，音源位置から一定範囲 (c_{th}) に定位したかどうかを示す指標である．**指標 1** と同様，C/N に相当する値とした．具体的な正解の判定条件は，Fig. 4b) における網掛け部分，つまり，以下の条件を満たす場合とした．

$$\Delta d = \sqrt{(x_{ref} - x_{local})^2 + (y_{ref} - y_{local})^2} \leq c_{th} \quad (13)$$

本稿では c_{th} として，1 m を用いた．この指標は，実際には，音源定位自体は極座標系で行われるので，たとえ音源定位結果が同じであっても，クワドロコプタからの距離が遠い音源ほど，精度が劣化する．

また，Tab. 1 記載のリファレンスデータは，正確性に欠けるため，この値をそのまま用いずに，定位結果のヒストグラムを作り，その中央値が Tab. 1 記載値から，±20° であれば，中央値をリファレンスの値とするようなキャリブレーションを行った．

4.3 実験結果

Tab. 2 に，Pelican 固定，バルーン固定条件に対する**指標 1** の結果を示す．また Tab. 3 に，Zion 固定条件に対する**指標 1** の結果を示す．Pelican では，プロペラ音が存在するにもかかわらず，ほぼ問題なく定位が実現できていることがわかる．着信音については，他の音源よりも性能の劣化が見られる．これは，着信音は特定の周波数のみにパワーが分布していることから，プロペラ音の周波数に埋もれやすいためではないかと考えられるが，より詳細な分析は今後の課題である．一方，バルーンはプロペラ音がないにもかかわらず，Pelican よりも定位性能が悪いという結果に終わった．特に，仰角については，音源定位結果は得られるものの，実際の方向とは大きく乖離した方向に定位が得られた．性性能が悪い原因としては，前述のようにバルーンが風で揺れてしまいしっかり固定できなかったこと，マイクロホンをバルーンの表面に貼りつけたため，風によるバルーンの変形に応じてマイクロホンアレイのレイアウトが変形してしまったこと，そもそもマイクロホンの位置を正確に計測することが難しかったことが挙げられる．また，仰角方向の定位に関しては，そもそものマイクロホンレイアウトが円状であったことから，上述の理由と合わさって性能が劣化したと考えられる．バルーンに対しては，プロペラ音が存在しないため，論理的には Pelican よりも良好な音源定位性能が得られるはずであるので，マイクロホン設置方法を工夫するなどして改良を行いたい．Zion については，音源の方向を変えながら定位性能の変化を調べた．概ね，Zion のプロペラ騒音下でも問題なく定位が可能であるといえる．方向毎の性能の変化が若干見られるが，少なくとも方位角に関しては，データ解析により，風の影響であることがわかった．

Tab. 4 に Pelican 移動条件に対する**指標 2** の結果，Tab. 5 に Pelican 移動条件に対する**指標 3** の結果を示す．移動条件は，固定条件と比較して，定位が難しいことがわかる．ホイッスルや警笛を見る限り，**指標 2** では距離とは無関係に性能が出ていること，**指標 3** では距離が大きくなる（A → C）につれて，性能が劣化していることがわかる．**指標 2** では，音源距離が 10 m 離れていても検出できている．これは，音源距離が 20 m 程度離れていてもホイッスルの定位が可能である（音声は 12 m 程度まで）とした報告[大畑 14]を裏付けるものである．しかし，**指標 3** では，ホイッスルでも定位性能が落ちており，距離まで含めた 3 次元的な音源定位性能で見るとこの距離では難しいことがわかる．これは，最初に 2 次元の音源方向だけ定位を行い，音源に近づくことで 3 次元音源定位を行う段階的なアクティブな音源定位が必要であることを示唆しているといえる．移動 A 条件では，音源までの距離が近いため，**指標 2** と**指標 3** で大きな差が見られない．つまり，指標によらず，定位しにくい音源は定位しにくいことがわかる．例えば，アナウンスは全く定位ができないという結果となった．これは，Fig. 3 にも示したように，アナウンスは音量が小さいために信号対雑音比が小さくなってしまっているためである．

Table 2: バルーン固定，Pelican 固定に対する軸別定位正解精度 [%]（すべて母数は 10)
無印は 500 Hz – 2800 Hz，$*$ 付は 2800 Hz–6000 Hz で評価を行った.「-」はデータ収録の失敗により算出できず.

		おーい	多人数発話	救急車	鈴	シンバル	電車	電車 2 本
バルーン	仰角	0	0	0	0	0	0	0
	方位角	100	80	100	90	60	100	80
Pelican	仰角	100	100*	100	100*	100	100*	100
	方位角	100	100*	100	100*	100	100*	100

		着信音	建設現場	カラス	バイク	遊園地	笛	アラーム
バルーン	仰角	0*	0	0	-	0	0	0
	方位角	60*	100	70	-	100	100	100
Pelican	仰角	60*	100*	100	100*	100	100*	100
	方位角	60*	100*	100	100*	100	100*	100

		トラック外	拍手	2 人発話	警笛	女声雑音	アナウンス	トラック内
バルーン	仰角	0	0	0	0	0	0	0
	方位角	70	60	80	50	70	80	70
Pelican	仰角	100	90*	100	100	100*	100	100*
	方位角	100	90*	100	100	100*	100	100*

Table 3: Zion 固定に対する軸別定位正解精度 [%]（母数は 10): 仰角範囲を 0–45° に制限

	Elevation	Azimuth
$-90°$	90	90
$-45°$	100	100
$0°$	100	100
$45°$	70	100
$90°$	90	100
$180°$	80	90

Table 4: Pelican 移動条件に対する指標 2 の結果 [%]

	おーい	救急車	鈴	カラス	ホイッスル	警笛	アナウンス
Pelican 移動 A	-	86	100	100	90	60	0
Pelican 移動 B	70	-	-	-	100	100	-
Pelican 移動 C	80	-	-	-	90	60	-

Table 5: Pelican 移動条件に対する指標 3 の結果 [%]

	おーい	救急車	鈴	カラス	ホイッスル	警笛	アナウンス
Pelican 移動 A	-	86	100	100	70	80	0
Pelican 移動 B	40	-	-	-	90	50	-
Pelican 移動 C	20	-	-	-	0	20	-

4.4 可視化の例

可視化については，ケーススタディとして，飛行実験の一例に処理を行ったケースを示す．Fig. 5 に実験の際の飛行データ，また Fig. 6 に収録した音響信号に対する MUSIC スペクトログラムを示す．Fig. 5 からは，クワドロコプタの 3 次元的な動きを把握することは難しい．また，クアドロコプタの向きが変化するため，Fig. 6 からだけでは，音源が検出できていそうであることは見て取ることができるものの，いつどこに音源があったのかまでを把握することは難しい．

Fig. 7 は，これらのデータをすべて統合し，Google マップ上に表示した結果のスナップショットとなっている．図は，上下 2 枚の画像のペアが時系列で並んでおり，各ペアの上の写真は，実際にカメラで現場を収録したデータからキャプチャしたものであり，下の画像は，カメラの画像と同じ視点に合わせて再合成した (Google マップ上に表示した) 結果である．再合成画像は，実際にカメラで収録したデータに近い結果が得られており，直感的に状況の把握が可能な可視化が実現できたといえる．また，さらにカメラ画像からでは音源がどこにあるかまではわからないが，再合成画像では音源の位置や発話時刻まで可視化することができ，より多くの情報をわかりやすく表示できていることがわかる．

5 おわりに

本稿では，2 種類のクワドロコプタ，およびバルーンに搭載したマイクロホンアレイを用いて，様々な音源を実際に屋外環境で収録し，2 次元，および 3 次元音源定位の性能評価を行った．また，定位結果の可視化について報告した．結果として，方位角，仰角からなる 2 次元の音源定位はプロペラ音や風が存在する環境下でもロバストに動作することが示された．また，距離の推定は音源までの距離が近ければ有効であることを示唆する結果を得た．一方で，実環境では正確なリファレンスデータの取得が難しく，評価を行う際には，リファレンスデータの誤差も考慮に入れる必要があることが分かった．今後は，システムのオンライン化，音源同定の導入を行う予定である．

謝辞

本研究は科研費基盤 (S) No.24220006 の支援を受けた．

参考文献

[Furukawa 13] Furukawa, K., Okutani, K., Nagira, K., Otsuka, T., itoyama, K., Nakadai, K., and Okuno, H. G.: Noise Correlation Matrix Estimation for Improving Sound Source Localization by Multirotor UAV, in *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, pp. 3943–3948, IEEE (2013)

[Nakadai 00] Nakadai, K., Lourens, T., Okuno, H. G., and Kitano, H.: Active Audition for Humanoid, in *Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000)*, pp. 832–839, AAAI (2000)

[Nakajima 10] Nakajima, H., Nakadai, K., Hasegawa, Y., and Tsujino, H.: Blind Source Separation with parameter-free adaptive step-size method for Robot Audition, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 6, pp. 1476–1485 (2010)

[Nakamura 09] Nakamura, K., Nakadai, K., Asano, F., Hasegawa, Y., and Tsujino, H.: Intelligent Sound Source Localization for Dynamic Environments, in *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, pp. 664–669, IEEE/RSJ (2009)

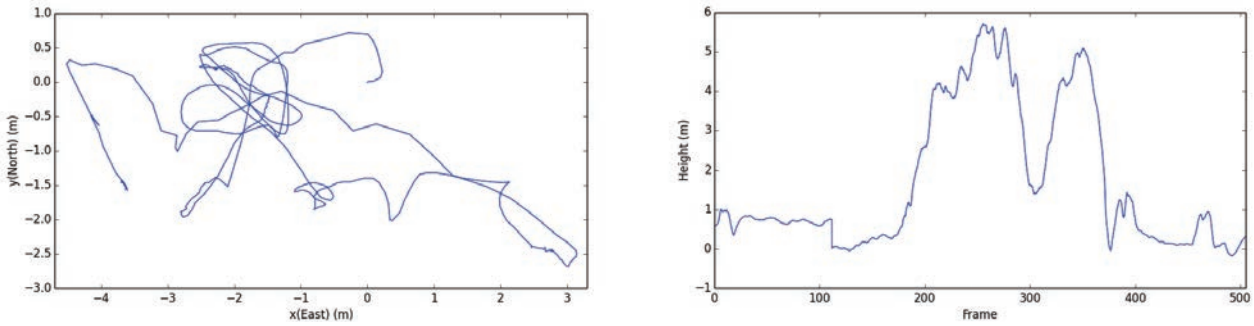


Figure 5: 航法データ (左：x-y 平面，右:高度)

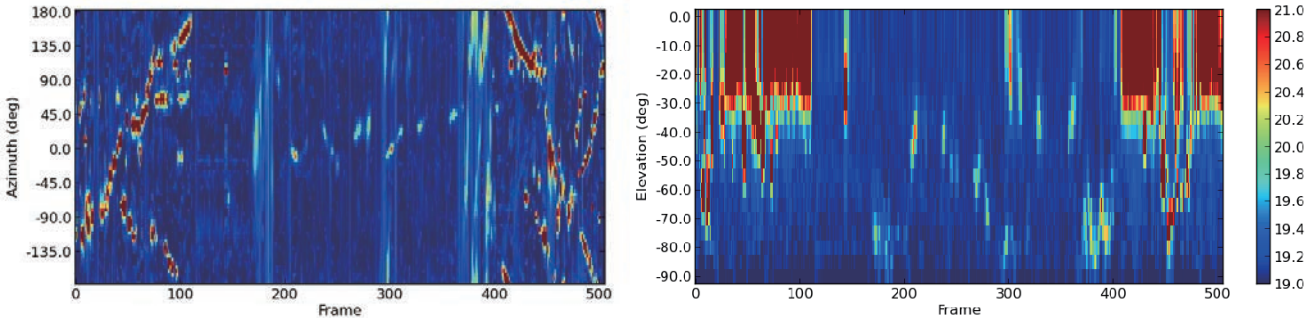


Figure 6: MUSIC スペクトル (左：方位角，右:仰角)

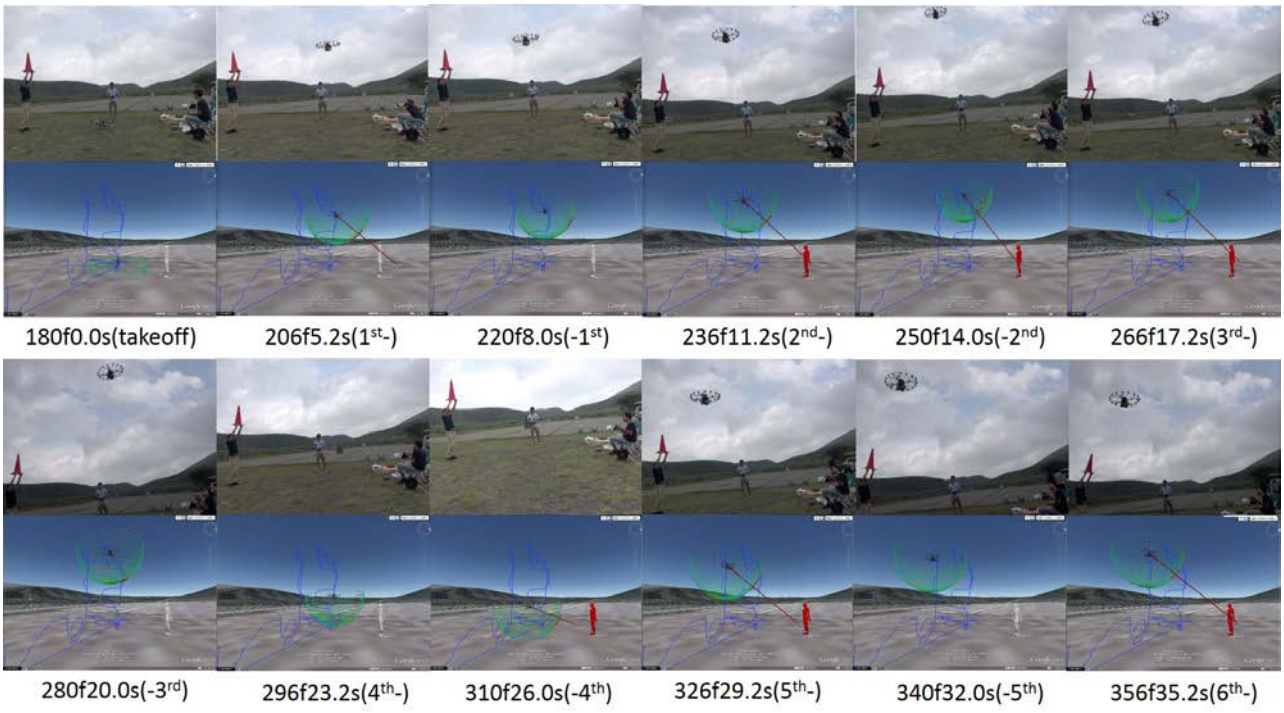


Figure 7: スナップショット (上：実際の画像，下:可視化システムでの表示，数字はフレーム数 f 時刻 s)

[Ohata 14] Ohata, T., Nakamura, K., Mizumoto, T., Tezuka, T., and Nakadai, K.: Improvement in Outdoor Sound Source Detection Using a Quadrotor-Embedded Microphone Array, in *Proc. of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS)*, IEEE Press (2014)

[Yamamoto 07] Yamamoto, S., Nakadai, K., Nakano, M., Tsujino, H., Valin, J.-M., Komatani, K., Ogata, T., and Okuno, H. G.: Design And Implementation Of A Robot Audition System For Automatic Speech Recognition Of Simultaneous Speech, in *Proc. of the 2007 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-2007)*, pp. 111–116, IEEE (2007)

[奥谷 13] 奥谷 啓太, 吉田 尚水, 中村 圭佑, 中臺 一博：クワドロコプタ搭載のマイクロホンアレイを用いた屋外音環境理解の逐次雑音推定による向上, ロボット学会誌, Vol. 31, No. 7, pp. 38–45 (2013)

[大畑 14] 大畑 琢磨, 長峰 諒英, 中村 圭佑, 水本 武志, 中臺 一博：相関行列スケーリングを用いた iGSVD-MUSIC 法による屋外環境音源探索の向上, 日本ロボット学会第 32 回学術講演会, pp. 1I1–03 (2014)

[丹羽 14] 丹羽 健太, 江崎 知, 日岡 裕輔, 西野 隆典, 武田 一哉：空間相関行列の固有値分布に着目した音源別距離推定, 電子情報通信学会論文誌 A, Vol. J97-A, No. 2, pp. 68–76 (2014)

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-04

深度センサとマイクロフォンアレイを用いた聴覚アウェアネスの提示

Proposal of auditory awareness using by depth sensor and microphone array

井山貴裕 [1]
Takahiro IYAMA

杉山治 [2]
Osamu SUGIYAMA

坂東宜昭 [1]
Yoshiaki BANDO

糸山克寿 [1]
Katsutoshi ITOYAMA

吉井和佳 [1]
Kazuyoshi YOSHII

奥乃博 [3]
Hiroshi G. OKUNO

[1] 京都大学大学院情報学研究科　　[2] 東京工業大学先進理工学研究科

[3] 早稲田大学実体情報学プログラム

Abstract

本稿では深度センサとマイクロフォンアレイを用いた音源位置推定に基づく聴覚アウェアネス可視化システムについて述べる．従来の音環境の可視化システムは MUSIC スペクトルをカメラ画像上に重畳するものであり，空間的· 時間的な聴覚アウェアネスが欠けている．空間的· 時間的な聴覚アウェアネスを提示するため，聴覚アウェアネス可視化のための三層モデルを設計し，本モデルに基づく可視化システムを開発する．本モデルでは，深度センサを用いることで空間的な聴覚アウェアネスを，音源を追跡し音源の時間変化を求めることで時間的な聴覚アウェアネスを提示する．また被験者実験により，本モデルに基づいて可視化された動画を視聴しながら，各レイヤごとに音源の発音したことを認識するまでの時間を比較し，本モデルの各レイヤごとの差異や有効性を確認した．

1　序論

環境の探索．監視システムの機能向上のためには，画像と音の情報統合に基づく充実した聴覚アウェアネスの提示が必要不可欠である．聴覚アウェアネスとは，音源の方向や位置，音量，種類，状態変化など，音源に対する総合的な気づきを意味する．単独マイクロフォンでは，当然ながら空間的な聴覚アウェアネスが提示できないのに加え，複数の音源が同時に発音した場合には音源種類の提示も困難になる．マイクロフォンアレイを用いると，音源方向など一部の聴覚アウェアネスは聴覚アウェアネスを提示できるが，奥行きも含めた音源の位置や音源状態変化などの提示は困難である．したがって，充実した聴覚アウェアネスの提示のためには，マルチチャネル信号処理技術で得られる音情報と RGB カメラや深度センサなどの画像情報の統合が不可欠である．

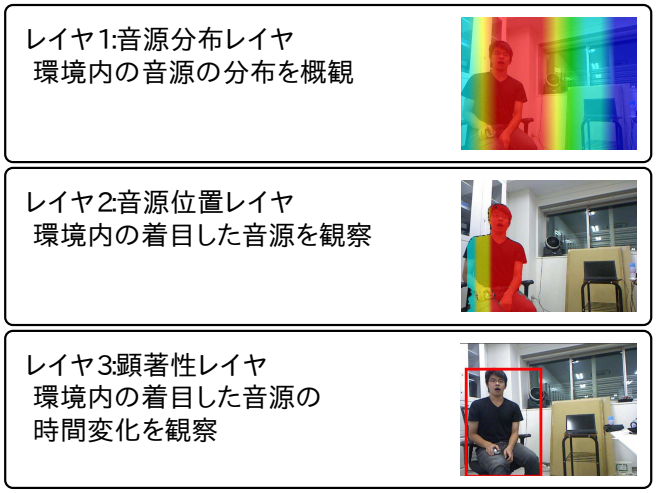


図 1: 聴覚アウェアネス可視化の三層モデルの構成

本研究では，聴覚アウェアネス可視化のための三層モデルを設計し，三層モデルに基づく聴覚アウェアネス可視化システムを設計する (図 1)．本モデルは音源分布レイヤ，音源位置レイヤ，顕著性レイヤの 3 つのレイヤから構成される．音源分布レイヤは環境中の音源の分布の様子を概観する機能を提供する．音源位置レイヤは着目した音源物体の音情報，すなわち空間的なアウェアネスを提示する．顕著性レイヤは着目した音源の時間変化の様子，すなわち時間的な聴覚アウェアネスを提示する．ユーザはこれらのレイヤを自由に選択しながら，着目した音源を観察することができる．

本稿の構成は以下の通りである．第 2 章では，従来の音情報可視化手法とその問題点について述べる．第 3 章では，聴覚アウェアネス可視化のための三層モデルについて述べ，第 4 章では，本モデルに基づく可視化システムについて述べる．第 5 章では，被験者実験により三層モデルの有効性を確認し，第 6 章でまとめを行う．

2　関連研究

聴覚アウェアネス可視化システムの開発のため，音環境の可視化· 深度センサを用いたマルチメディア統合に関する従来法を挙げ，本研究の位置づけを明確にする．

まず，音環境の可視化に関して神保ら[Jimbo et al., 2008]は，192 個のマイクロフォンアレイと CMOS カメラを使用し，RGB 画像上へ音高の帯域ごとの強さを重畳表示している．この可視化手法は，音源の分布を提示するが，空間的な聴覚アウェアネスである音源の位置や時間的な聴覚アウェアネスである音源の時間変化の提示は行っていない．

次に，深度センサを用いたマルチメディア統合に関して Even ら[Even et al., 2013]は，マイクロフォンアレイとレーザーレンジファインダを使用し，SLAM で作成した地図上に音源の位置を重畳表示している．この可視化手法は，音源の強さとレーザレンジファインダによって音源の位置を提示するが，音源の時間変化である時間的な聴覚アウェアネスの提示は行っていない．

これらの研究を受け井山ら[Iyama et al., 2014]は，マイクロフォンアレイと深度センサを使用し，聴覚アウェアネスを三層モデルで定義し，これを可視化するシステムを開発した．この三層モデルは，環境内の音の分布を概観する機能を提供する音源分布レイヤ，着目した音源の位置やパワーを抽出する音源位置レイヤ，新しい音源の出現や音源のパワーの大きな変化を抽出する顕著性レイヤから構成される．そのため，空間的· 時間的なアウェアネスの提示も行なっている．しかし，そのモデルの有効性が評価されていなかった．本稿では，聴覚アウェアネスの三層モデルを拡張し，可視化システムを開発し，その評価を行うことで三層モデルの有効性を確認する．

3　聴覚アウェアネス可視化のための三層モデル

聴覚アウェアネス可視化のための三層モデルは，音源分布レイヤ，音源位置レイヤ，顕著性レイヤの 3 つのレイヤから構成される．充実した聴覚アウェアネスの提示のため，音源位置レイヤは空間的な聴覚アウェアネスを，顕著性レイヤは時間的な聴覚アウェアネスを提示する．ユーザはこれらのレイヤを自由に切り替えながら，音環境の観察を行うことができる．各レイヤは，2 つの処理から構成される．はじめに，レイヤの入力データの可視化可能なデータへの変換や高次レイヤへのデータの受け渡しを行う．次に変換したデータから可視化画像の生成を行い，ユーザに提示する．次節以降で各レイヤの役割と処理について述べる．

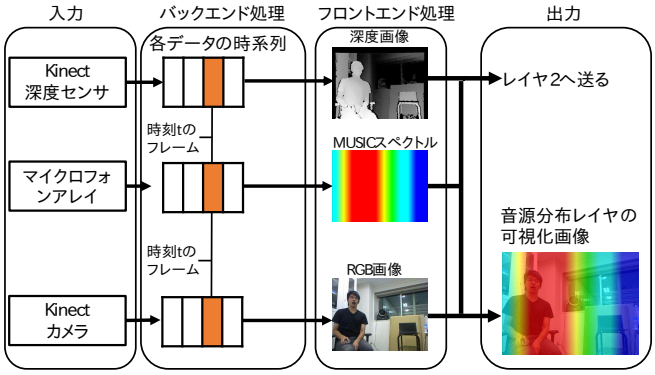


図 2: 音源分布レイヤの処理と可視化結果

3.1　音源分布レイヤ (レイヤ 1)

音源分布レイヤは環境内の音の分布を概観する機能を提供する．ユーザは環境内の音源の分布を MUSIC スペクトル[Asano et al., 2001]の色画像を RGB 画像に重畳した画像として提示される．ユーザが効率的に環境を概観するため，MUSIC スペクトルの色画像の濃淡や可視化する MUSIC スペクトルの範囲を変更することができる．ユーザが環境を観察したいとき，MUSIC スペクトルの色画像を RGB 画像に重畳した画像から環境全体を概観することができる．

音源分布レイヤのデータフローと処理は図 2 の通りである．入力データは RGB カメラから取得した RGB データ，深度センサから取得した深度データ，マイクロフォンアレイから取得した MUSIC スペクトルである．入力デバイスから取得したデータの時間同期を行い，各入力データの色画像への変換と，MUSIC スペクトルの色画像を RGB 画像に重畳した画像を生成する．MUSIC スペクトルの色画像は MUSIC スペクトルのパワーに対応した色を割り当てることで生成する．音源分布レイヤのこれらの処理によって，ユーザは環境内の音の分布を概観することができる．

3.2　音源位置レイヤ (レイヤ 2)

音源位置レイヤは環境内のユーザが着目した音源の音情報，すなわち空間的な聴覚アウェアネスをユーザに提示する．ユーザは着目した音源の RGB 画像上のみに MUSIC スペクトルの色画像を重畳した画像を提示される．ユーザは音源分布レイヤを用いて環境を概観した後，音源位置レイヤを用いて着目した音源のみの音情報を観察することができる．

音源位置レイヤのデータフローと処理は図 3 の通りである．まず，音源分布レイヤから送られる深度データからユーザが着目する音源物体の形状を推定する．着目する音源物体の形状は深度データに領域成長法[Ballard et al., 1982]を用いて算出する．そして，各入力データの色画像への変換と，MUSIC スペクトルの色画像を RGB 画像の着目した音源物体上に重畳した画像を生成する．音源位

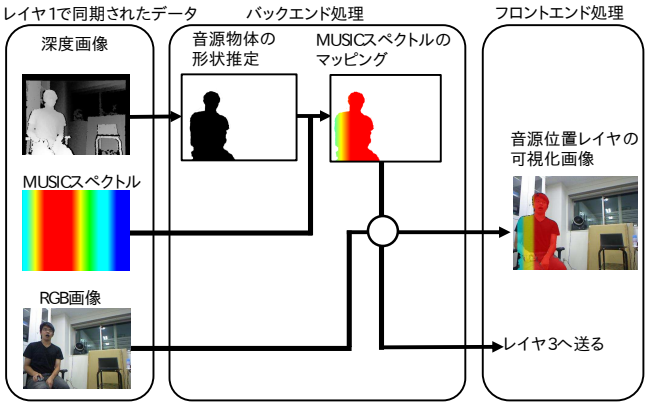


図 3: 音源位置レイヤの処理と可視化結果

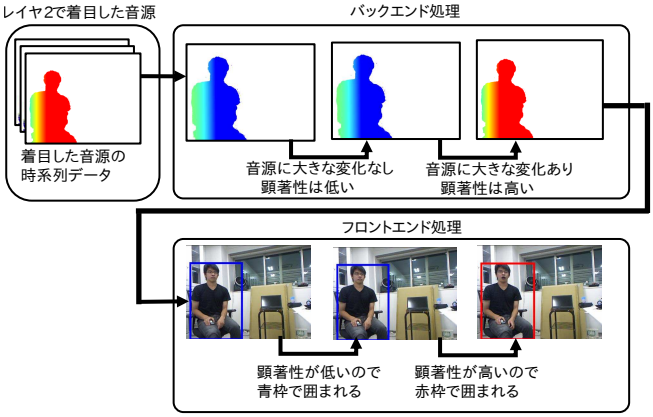


図 4: 顕著性レイヤの処理と可視化結果

置レイヤのこれらの処理によって，ユーザは環境内の着目した音源の音情報のみを観察することができる．

3.3 顕著性レイヤ (レイヤ 3)

顕著性レイヤは環境内のユーザが着目した音源の顕著性，すなわち時間的な聴覚アウェアネスをユーザに提示する．音源の顕著性とは，音響情報と音源の位置· 形状情報の時間変化によって定義される．音を発していなかった音源が音を発し始める場合や環境内に新しく音源が出現した場合は顕著性が大きくなる．一方，音源が音を発していない場合や，音源が発生している音に変化がない場合は顕著性が小さくなる．ユーザは RGB 画像に着目した音源の顕著性の大きさに対応した色枠を重畳した画像を提示される．ユーザは環境内の着目した音源の時間変化の様子を観察することができる．

顕著性レイヤのデータフローと処理は図 4 の通りである．まず，着目した音源物体の顕著性を算出する．顕著性 d_c はフレーム t とフレーム $t-1$ の音源物体の MUSIC スペクトルの変化量 l_m と位置の変化量 l_d の加重平均として求められる

$$d_c = \alpha \cdot l_d + (1.0 - \alpha) \cdot l_m. \tag{1}$$

l_m，l_d はカルバック· ライブラーダイバージェンスによって次のように求められる

$$\begin{cases} l_d = \frac{1}{2}\left[\log\frac{|\Sigma_{d_{t-1}}|}{|\Sigma_{d_t}|} + tr\{\Sigma_{d_{t-1}}^{-1}\Sigma_{d_t}\} \right. \\ \qquad \left. + (\mu_{d_t} - \mu_{d_{t-1}})^T\Sigma_{d_{t-1}}^{-1}(\mu_{d_t} - \mu_{d_{t-1}}) - 3\right] \\ l_m = \frac{1}{2}\left[\log\frac{\sigma_{m_{t-1}}^2}{\sigma_{m_t}^2} + \frac{\sigma_{m_t}^2}{\sigma_{m_{t-1}}^2} + \frac{(\mu_{m_t} - \mu_{m_{t-1}})^2}{\sigma_{m_{t-1}}^2} - 1\right] \end{cases}$$

ここで，Σ_{d_t}，$\Sigma_{d_{t-1}}$ はフレーム t，$t-1$ の深度データの共分散行列，μ_{d_t}，$\mu_{d_{t-1}}$ は深度データの平均，σ_{m_t}，$\sigma_{m_{t-1}}$ は MUSIC スペクトルの分散，μ_{m_t}，$\mu_{m_{t-1}}$ は MUSIC スペクトルの平均である．そして，顕著性の大きさに基づく色画像の生成と生成した色画像を RGB 画像に重畳した画像を生成する．顕著性レイヤのこれらの処理によって，ユーザは環境内の着目した音源の時間変化の様子を観察することができる．

表 1: インタフェースのパラメータ

レイヤ	パラメータ	概要	範囲
1	透明度 t	MUSIC スペクトルのパワー画像の透明度	$0 < t < 1$
	可視化の最小値 p	可視化する MUSIC スペクトルの最小値	$0 < p$
2	領域成長法の閾値 d	領域成長法で統合するかの閾値	$0 < d$
3	重み α	顕著性の重みパラメータ	$0 < \alpha < 1$

4 インタフェース設計

ユーザが要求する可視化結果を得るため，GUI の操作やパラメータの変更はスライダやボタンでなく，可視化画像に対するジェスチャを用いて行う．例えば，ユーザが着目したい音源対象を選択する操作は，画面内の画像を直接選択できるほうがより直感的である．本インタフェースはジェスチャの入力としてマウスの左クリック，右クリック，中クリック，マウスホイールを使用する．ユーザが要求する可視化結果を得るためのインタフェースの操作を簡易で直感的にできるよう設計する．

GUI は図 5 に示すように，画像表示部，ステータス部から成る．画像表示部に，三層モデルで生成した画像を組み合わせたものが表示される．ユーザは画像表示部上で三層モデルで使用するパラメータを変更することによって，表示する画像を変更できる．表 1 はユーザが変更できるパラメータの一覧である．レイヤごとにパラメータは存在し，ユーザは三層モデルにおける特徴量を柔軟に組み合わせることができ，自由に画像表示部に表示される描画結果を変更することができる．ステータス部は現在のレイヤやパラメータの変更内容などを表示する部分である．これによりユーザは現在のレイヤ状態やパラメータ状態を確認しながら操作することができる．以下では，各レイヤにおける操作について述べる．

図 5: インタフェースのデザイン

図 6: マウスホイールによる透明度の変更

4.1 音源分布レイヤのインタフェース設計

音源分布レイヤでは，ユーザは重畳される MUSIC スペクトルの色画像の濃淡と可視化する MUSIC スペクトルの帯域を変更できる (図 6)．これらの機能によって，ユーザは環境や要求に応じた可視化結果を得ることができる．例えば，音の分布を鮮明に観察するときは濃く，音の弱い部分が必要ないときは帯域の最小値を上昇させることができる．

これらのパラメータの変更は，いずれも増減であるため，マウスホイールの操作により変更する．マウスホイールを上に回転させると，色画像の透明度や可視化する MUSIC スペクトルの最小値が増大し，下に回転させると減少する．これら 2 つのパラメータのどちらを変更するかの切り替えは，右クリックで行えるようにする．

4.2 音源位置レイヤのインタフェース設計

音源位置レイヤでは，ユーザは着目する音源対象の選択と領域成長法の類似度の閾値パラメータの変更できる (図 7)．これらの機能によって，ユーザは着目したい音源を選択できる．

着目する音源の選択は，可視化領域内の着目する画像をマウスの左クリックにより行う．領域成長法の閾値パラメータが大きいと，より広い範囲に存在する複数物体を同一領域とみなし，小さいと，領域をより細かく分割する．このパラメータの変更は増減であるため，マウスホイールの操作により変更する．マウスホイールを上に回転させると，閾値パラメータは増大し，下に回転させると減少する．

図 7: クリックによる音源の選択

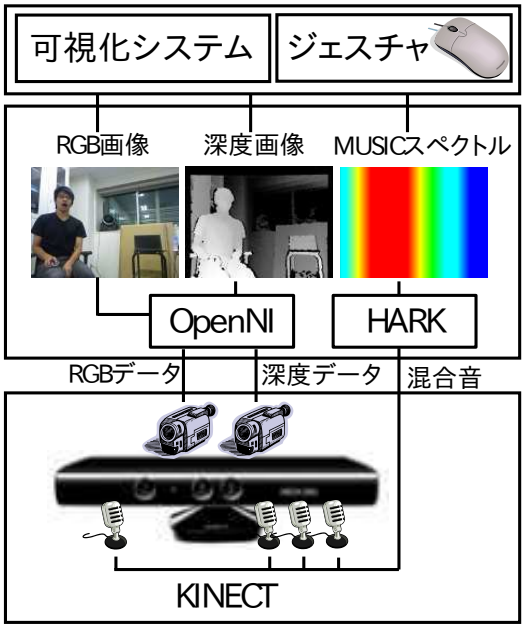


図 8: 聴覚アウェアネスの可視化システムの詳細

4.3 顕著性レイヤのインタフェース設計

顕著性レイヤでは，ユーザは顕著性を算出する際の音響情報と深度情報の重みの変更できる．音響情報の重みが大きくなるほど，着目した音源の音響情報の変化が顕著性に大きく影響し，深度情報の重みが大きくなるほど，着目した音源の移動量や形状の変化が顕著性に大きく影響するようになる．

これらのパラメータの変更は，いずれも増減であるため，マウスホイールの操作により変更する．また，音響情報と深度情報のどちらを変更したいかは，ユーザの要求に応じて変わるので，重みを変更する情報をマウスの右クリックで変更する．

各レイヤ間の移動については以下のように設計する．音源分布レイヤから音源位置レイヤへの移動は，音源位置レイヤで着目する音源を左クリックしたときに移動する．音源位置レイヤから顕著性レイヤへの移動は，音源位置レイヤで着目している音源の領域内を左クリックしたときに移動する．顕著性レイヤから音源位置レイヤ，音源位置レイヤから音源分布レイヤへの移動はマウスの中クリックを行うことで移動する．

5 実験

聴覚アウェアネス可視化のための三層モデルを用いて音環境の可視化を行い，三層モデルによって聴覚アウェアネ

(a) 音源が 2 個のときの音源の配置　(b) 音源が 4 個のときの音源の配置　(c) 音源が 8 個のときの音源の配置

図 9: 音源の配置と可視化動画の一例

スが提示されているかを被験者実験によって評価した．

5.1　システム構成

システム構成は図 8 の通りである．本システムへの入力データは Kinect を用いて取得した RGB 画像，深度画像，および多チャンネル音響信号である．RGB 画像と深度画像は OpenNI ライブラリ [sim, 2014]を，多チャンネル音は HARK(Honda Research Institute Japan Audition for Robots with Kyoto University)[Nakadai et al., 2010]を通じてそれぞれ取得され，システムに渡される．各レイヤのデータ処理や GUI への様々な画像の描画には Processing を用い，三層モデルの特徴量を柔軟に変化させることができるようシステムを設計する．

5.2　実験設定

実験では，空間的な聴覚アウェアネスの有効性を確認する．空間的な聴覚アウェアネスを考慮しない音源分布レイヤによる可視化結果と，音源位置レイヤによる可視化結果について，どちらが音の発生に即座に気づくかを比較した．音の発生から被験者の認識までの時間を比較するために，それぞれのレイヤで可視化された 30 秒程度の動画を被験者に視聴させ，あらかじめ指定した物体が音を発したと認識した時間を記録した．

視聴する動画は各レイヤについて，音源が 2 個，4 個，8 個の 3 種類，計 6 種類用意した (図 9)．各実験と音源の数，可視化するレイヤの対応は表 2 の通りである．音源の再生デバイスとしてはすべて同一のスピーカを使用した．使用した音源は，ATR の音素バランス文[Kurematsu et al., 1990]，RWC 音楽データベース[Goto et al., 2002]のクラシック曲，ホワイトノイズ，サイン波のテストトーンである．三層モデルに必要な各データは，Kinect から取得した 30fps の深度データ・RGB データと 4ch 同期，16bit 量子化，16kHz の音響信号を用いた．各動画の各音源の音の発生時刻は図 10 の通り．

実験手順は，音源が 2 個の動画の各レイヤによる実験，次は音源が 4 個の動画，最後に音源が 8 個の動画の実験という手順で行った．各実験でどちらのレイヤによる動画を視聴した順序による実験結果の偏りをなくすため，6

表 2: 各実験の音源数と可視化レイヤの対応

実験種類	音源数	可視化レイヤ
実験 1-1	2	音源分布レイヤ
実験 1-2	2	音源位置レイヤ
実験 2-1	4	音源分布レイヤ
実験 2-2	4	音源位置レイヤ
実験 3-1	8	音源分布レイヤ
実験 3-2	8	音源位置レイヤ

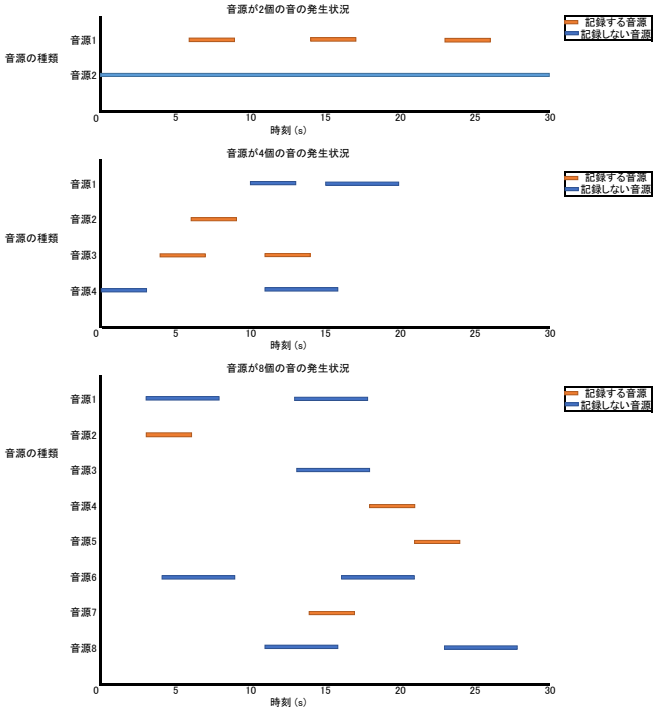
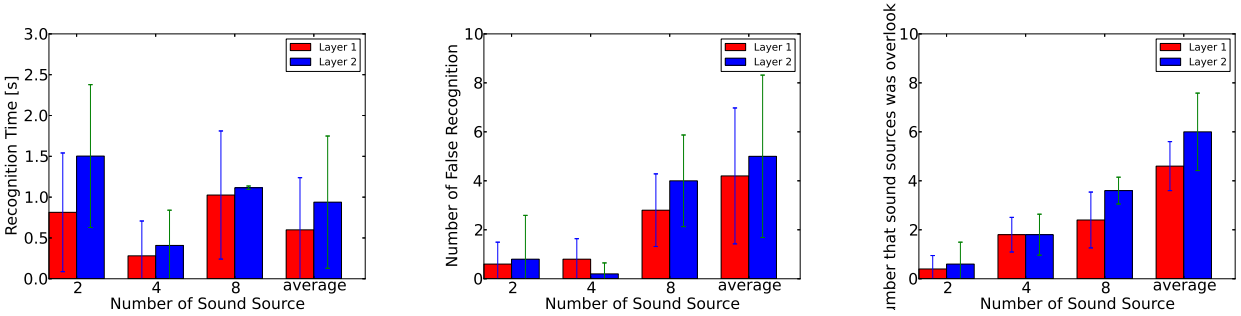


図 10: 各動画の各音源の音の発生時刻

人の被験者を 2 つのグループに分けて実験を行った．あるグループは，音源分布レイヤによる可視化動画による実験を行った後に音源位置レイヤによる可視化動画による実験を行い，一方のグループは音源位置レイヤによる可視化動画による実験を行った後に音源分布レイヤによる実験を行った．また，どの音源からどのような種類の音が発生するのかは事前に知らせておらず，常に未知の音を聞く状態にした．記録した時間が正しい範囲は，音源の再



(a) 各音源数の認識時間　(b) 各音源数の誤認識数　(c) 各音源数の見逃し数

図 11: 各音源数・レイヤの認識時間・誤認識数・見逃し数

生時間が 3 秒であるため正解時間から 3 秒間とした．

5.3 実験結果

図 11 に示すように，音源分布レイヤの場合の認識時間は音源位置レイヤの場合の認識時間よりも早い．平均認識時間はすべて正しく記録した時間の値のみから計算されており，見逃しや誤認識の時間は含まれていない．また，合計誤認識数，合計見逃し数ともに音源分布レイヤの場合のほうが少ない．合計誤認識数・合計見逃し数は音源数が 2 個，4 個，8 個の場合の合計である．

5.4 考察

平均認識時間・合計誤認識数・合計見逃し数の全ての値に関して，音源分布レイヤの結果が音源位置レイヤの結果より数値が小さく，どのような音が発生するか未知の場合では全体を概観する機能のほうが適していると考えられる．これは音源物体の大きさが画面に比べて小さく，音源定位の結果が少しでも誤った場合，正しく物体上に MUSIC スペクトルの色画像が重ねられないためであると考えられる．そのため，今後の実験においては実際の使用順序にしたがって，発生する音の種類を既知とした実験を行うべきだと考えている．また，正確なキャリブレーションや定位精度向上の手法に取り組む必要がある．

6 結論

本研究では，音源分布レイヤ，音源位置レイヤ，顕著性レイヤから構成される聴覚アウェアネス可視化の三層モデルを設計し，Kinect を用いた聴覚アウェアネス可視化システムを実装した．音源分布レイヤは環境内の音の分布を概観する機能を，音源位置レイヤは着目した音源情報を抽出する機能を，顕著性レイヤは音情報の時間変化，すなわち，新しい音源の出現や音源のパワーの大きな変化といった顕著性を抽出する機能を提供する．三層モデルに基づくデータ処理や可視化を行い，各レイヤのパラメータをジェスチャを用いて変更することで，音環境を分析するための直感的な操作が可能なインタフェースを開発した．被験者実験によって，三層モデルによって聴覚アウェアネスが提示されているかの実験を行った．その結果，発生する音の種類が未知の状況下では音源分布レイヤによる可視化が音源位置レイヤによる可視化より，高速な音源の認識や少ない誤認識を行うことができることを確認した．

今後，発生する音の種類を既知にするなど実験の情報を増やし，実際の使用に近い環境における実験を行い，再度モデルの有効性の評価や音源定位の精度向上などシステムの処理部分の改善を行う予定である．

謝辞 本研究の一部は科研費 No.24220006 と No.24700168 の支援を受けた．

参考文献

[Asano et al., 2001] F. Asano et al. Real-time sound source localization and separation system and its application to automatic speech recognition. In *INTERSPEECH*, pages 1013–1016, 2001.

[Ballard et al., 1982] D. H. Ballard et al. *Computer Vision*. Prentice Hall, 1982.

[Even et al., 2013] J. Even et al. Creation of radiated sound intensity maps using multi-modal measurements onboard an autonomous mobile platform. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3433–3438, Nov 2013.

[Goto et al., 2002] M. Goto et al. RWC Music Database: Popular, Classical and Jazz Music Databases. In *ISMIR*, volume 2, pages 287–288, 2002.

[Iyama et al., 2014] T. Iyama et al. Visualization of auditory awareness based on sound source positions estimated by depth sensor and microphone array. In *Intelligent Robots and Systems (IROS), 2014 IEEE/RSJ International Conference on*, pages 1908–1913, Sep 2014.

[Jimbo et al., 2008] N. Jimbo et al. Visualization of sound environment using multi channel acoustic measurement system. In *Acoustic Society Symposium,2008*, pages 1509–1510, Sep 2008.

[Kurematsu et al., 1990] A. Kurematsu et al. Atr japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, 9(4):357–363, 1990.

[Nakadai et al., 2010] K. Nakadai et al. Design and Implementation of Robot Audition System 'HARK' – Open Source Software for Listening to Three Simultaneous Speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[sim, 2014] simple-openni - openni library for processing. https://code.google.com/p/simple-openni/, 2014.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-05

臨場感の伝わる遠隔操作システムのデザイン：
マイクロフォンアレイ処理を用いた音環境の再構築
Design of tele-operation system for tele-presence: Recreating auditory scene using microphone arrays

劉超然[1]，カルロス石井[1]，石黒浩[2]，萩田紀博[1]
Chaoran LIU, Carlos ISHI, Hiroshi ISHIGURO, Norihiro HAGITA
国際電気通信基礎技術研究所
[1]知能ロボティックス研究所
[2]石黒特別研究所
[1]ATR/IRC
[2]ATR/HIL
chaoran.liu@irl.sys.es.osaka-u.ac.jp, carlos@atr.jp, ishiguro@sys.es.osaka-u.ac.jp, hagita@atr.jp

Abstract

コミュニケーションロボット遠隔操作システムにおいて，ロボット側の空間的音環境を操作者に再現することは，臨場感の伝達に大きな役割を担う。本稿では，ロボット周囲の音源位置情報に基づいて，３次元音環境を操作することのできる遠隔操作システムを提案した。ロボット側では、音源定位・分離において複数のマイクロフォンアレイとヒューマントラッキング技術を用いた。操作者側では，操作者の頭部回転をトラッキングし、操作者の動きを補正してロボット側の空間的音環境を再生する。提案システムを用いることによって，従来法よりも高い定位精度と強い臨場感・聞き取り安さが得られることを被験者実験により確認した。また，バーチャル音環境を操作するために，２種類のユーザインタフェースも提案し、検証した。

1 はじめに

近年，ロボット遠隔操作システムにおいて，操作者の存在感をロボット側に伝達する研究が広く行われている。しかし，操作者側へ遠隔地の臨場感を伝達することに注目した研究は少ない[Nishio 2007] [Ishi 2010] [Liu 2012] [Sumioka 2014]。対面コミュニケーションに比べて，遠隔地にいる人物がロボットを介して人とコミュニケーションする場合，空間情報などの欠落によって相手との共有情報が不足する。そのため，操作者側ではコミュニケーションが行われている現場の臨場感を感じることが困難である。

臨場感の伝達に大きな手助けとなるのは，バーチャルリアリティ技術である。現在では多くの遠隔医療・軍事・コミュニケーション目的のアプリケーションなどにおいてバーチャルリアリティ技術が利用されているが[Popescu 2000] [Piron 2009] [Billinghurst 2002] [Ogi 2001]，臨場感の伝達はこれらの一つの大きな目的となっている。しかし，これらバーチャルリアリティに関する研究の大部分は，視覚における臨場感伝達に着目している[Ogi 2001] [Bullinger 1997]。音環境の構築に関するバーチャルリアリティの研究は，ゲームなどのアプリケーションで用いられているものの，未だ少ないのが現状である。リッチな音環境の構築は，遠隔操作ロボットなどのソーシャルメディアにおいても，操作者に遠隔地での自身の存在感や現場の臨場感を伝えるために重要である。

以上の背景から，本研究は遠隔地にあるロボット周囲に分布している複数の音源から構成される音環境（3D 音場）を，操作者（オペレーター）側に再現・加工することで，音の臨場感を伝達する遠隔操作システムの開発を目的とする。提案システムはリアルタイム性を保ちながら，空間的に分布する複数の音源を定位・分離し，正確な位置に再生する能力を備えることが求められる。

3D 音場を再現するため従来広く使われた方法は，バイノーラル（両耳）レコーディングされた音声をステレオで再生することである。この方法は簡便であるという利点があるが，正確なステレオマイクロフォンのセッティングが必要で，尚且つダミーヘッドが動かないためダイナミックに音場を再現することができない。さらに，各音源に対して加工を加えることも不可能である。

サラウンドチャンネルスピーカーは空間的な音場の再現のために開発されており，DirAC（Directional Audio Coding）を用いた音場再現の研究は少なくない[Pulkki 2007] [Laitinen 2011]。だが，サラウンドスピーカーシステムには二つの問題点がある。一つ目は，音場を録音した環境とそれを再生する環境が異なる場合，部屋の大きさや形状などの環境的要素が音響の伝達に影響を与えてしまい，これらの影響を正確に補正することは困難であるという点である。二つ目は，サラウンドスピーカーシステムでは “sweet spot” の位置がシステムの中心付近に限られている[Rumsey 2001]，という点である。即ち，聴者の場所が制限される。

ヘッドフォンを用いた 3D 音場の再現も，これまで広く研究されてきた。日常，人は両耳に到達した音

波の違いによって音源定位を行っている[Meyer 1972]。この違いを再現することで，ステレオヘッドフォンで 3D 音場を合成することが可能になる。頭部伝達関数（HRTF: Head Relative Transfer Function）は空間内の音源から発した音波が人の両耳に到達する時点の違いを表現する関数であって，3D 音場のバイナル再現に多く使われている[Cheng 2001]。しかし，ヘッドフォンを使って空間上に存在する音源を再現する際，バーチャルな音源が聴者の頭部・体の動きと共に動いてしまうという問題点がある。人の日常経験を考えると，外部音源の位置は聴者の体の動きに関連せず，固定されている。ヘッドフォンによる 3D 音場の再現ではこの経験と異なるため，臨場感の伝達にマイナスに働き，不自然な印象の原因となる。さらに，頭部伝達関数を使った場合，前後の誤判断が起こるという問題がある。これは，前方にある音源が後方にあるように聞こえる，もしくはその逆の現象である。日常生活では音源を定位するために意識的・無意識的に頭部を回し，その効果を定位の補助に用いている。また，頭部を回転することで前後の誤判断率が有意に下がったことも報告されている[Iwaya 2003]。

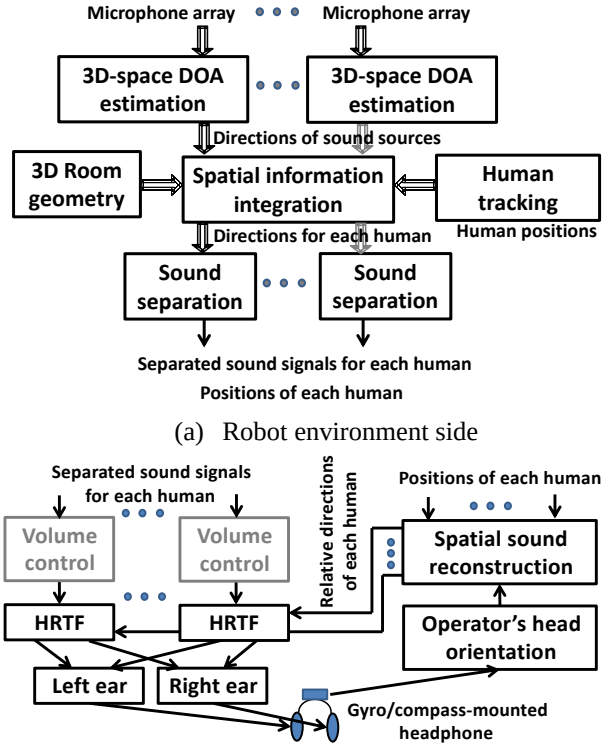
一方で，環境内の音源の空間的特性を保持するために多く使われているのは，マイクロフォンアレイ処理技術である。マイクロフォンアレイを用いた遠隔会議の研究では，音源定位や音源分離，雑音抑圧が応用されているが，多くの場合は分離音をモノラルで再生し、音場を再現している訳ではない。

これらを考慮し，提案システムではオペレーターの頭部回転をトラッキングすることで，頭部の向きに合わせた HRTF を用いてステレオ音声を合成した。正確な HRTF を選択するのに必要な連続的音源位置情報は，複数のマイクロフォンアレイの DOA（Direction Of Arrival）推定結果，および，人位置推定システムから取得する。さらに，合成したバーチャル音場の加工を制御するために 3 つのユーザインタフェースを提案し，被験者実験を通して検証した。

2 提案システム

提案システムは二つの部分から構成されている。一つはロボット側の音源位置推定・トラッキングと複数人の音源分離であり，もう一つはオペレーター側の頭部回転トラッキングとステレオ音声の合成である。Figure 1 に提案システムのブロック図を示す。

ロボット側の処理では，まず，各マイクロフォンアレイによって音の 3 次元到来方向（DOA）が推定される。環境とアレイの位置関係と各音源の DOA を統合することで，3 次元上での人位置情報が得られる。この人位置情報は，ヒューマントラッキングシステムにより，非発声時にも常時追跡されている。次に，推定した人位置情報に基づいて各人の音声を分離し，位置情報と合わせてオペレーター側のシステムに送信する。



(a) Robot environment side

(b) Tele-operator side

Figure 1. Block diagram of the proposed tele-presence system.

オペレーター側の処理では，まず，人位置情報とオペレーターの顔の向きによって，左右のチャンネルに対応した正確な HRTF をデータベースから選択する。次に，分離した音声に畳み込み演算を行い，ステレオヘッドフォンでオペレーターに再生する。オペレーターの頭部回転トラッキングには，ヘッドフォンの上部に取り付けたジャイロセンサーとコンパスを用いた。また，分離した各音源のボリュームは，ユーザインタフェースにて独立して調節することができる。

2.1 3 次元音源定位

音源定位に関して，まず，各マイクロフォンアレイで DOA 推定を行う。複数のアレイによる DOA 情報と人位置情報を統合することで，音源の 3 次元空間内の位置を推定する。

実環境での音の DOA 推定は広く研究されてきた。MUSIC 法は，複数のソースを高い分解能で定位できる最も有効な手法の一つである[Schmidt 1986]。この手法を使うには事前に音源数が必要であるため，本研究では[Ishi 2009]で提案した解決法を用いる。音源数を固定した数値に仮定し，閾値を超えた MUSIC スペクトルのピークを音源として認識する。この研究で使用した MUSIC 法の実装は 100 ms ごとに 1 度の分解能を有しており，2 GHz のシングルコア CPU でリアルタイムに探索することができる。

コミュニケーションロボットの遠隔操作システム

にとって，最も重要な音源は人の音声である。本研究では人の声を漏れ無く抽出するために，複数の2D-LRF（Laser Range Finder）で構成したヒューマントラッキングシステムを使用した[Glas 2007]。複数のマイクロフォンアレイからの DOA 推定出力と LRF のトラッキング結果が同じ位置で交差すれば，そこに音源がある可能性が高い[Ishi 2013] [石井 2014]。本システムでは 2D の LRF を用いているため、人位置情報は 2D に限られる。ここでは、検出された音源の位置が口元の高さの範囲内にあるかの制限もかけている（z = 1 ~ 1.6m）。 無音区間や音源方向推定が不十分な区間では、最後に推定された口元の高さと最新の 2D 位置情報を用いて，音源分離を行う。

2.2 音源分離

音源分離では，選択された複数の人物をパラレルに分離している。本研究では計算量が少なく且つロバストな Delay-Sum Beamformer を用いて，目的方向の人の声を分離した[Dudgeon 1977]。フレーム長は 20 ms で，シフト長は 10 ms である。

本研究で使用した 16 チャンネルのマイクロフォンアレイ（半球 30cm にマイクを配置した形状）の DS ビームフォーマのレスポンスの特徴として，低周波領域の分解能が低いことが挙げられる。そのため，無指向性雑音の低周波成分が分離音に多く混在してしまい、臨場感の伝達に悪影響を与える可能性がある。

空間に指向性音源 S と無指向性雑音源 N が存在すると仮定すると，DS ビームフォーマの出力は以下の形になる：

$$Y(f) = \boldsymbol{w}_{Sdir}(f) \cdot S(f) + \int_0^{2\pi} (\boldsymbol{w}_\theta(f) \cdot N(f)) d\theta$$

Y は周波数 f に対応したビームフォーマの出力で，Sdir は信号の方向，$\mathbf{w}_{\mathrm{Sdir}}$ は Sdir 方向のビームフォーマレスポンスを指す。式の二つ目の項目は，分離音声に混在する雑音を表している。この雑音成分を低減させるために，各周波数に以下のようなウェイトを掛けた。

$$w_{PF}(f) = \frac{1}{\int_0^{2\pi} \boldsymbol{w}_\theta(f)\, d\theta}$$

$$Y_{PF} = \sum_f w_{PF}(f) \cdot Y_{DS}(f)$$

Y_{PF} はウェイト掛けした後のビームフォーマ出力である。

さらに，各音源とアレイの間の距離による違いを補正するため，分離した各音声に対して距離によって以下のように正規化を行った。

$$g_i = \frac{\sum_{n=1}^{N} dist_n - dist_i}{(N-1) \cdot \sum_n^N dist_n}$$

$$Y_i = g_i \cdot Y_{PF,i}$$

このうち，N は音源の数で，$\mathrm{dist_n}$ は n 番目の音源とアレイの距離を表す。g_i は i 番目の音源に掛ける正規化ファクタで，Y_i は i 番目の音源の分離結果を示している。

2.3 HRTF による音場合成

一つの音声を特定の方向から聞こえるようにするため，その方向に対応した HRTF によってフィルタリングするステレオ化方法が一般的である。本研究では，一般公開されている KEMAR (Knowles Electronics Manikin for Acoustic Research) ダミーヘッドの HRTF データベースを利用した[Gardner 1995]。KEMAR は HRTF 研究のために一般的な頭部サイズを使って作られたダミーヘッドで，データベースには空間からのインパルス信号に対するダミーヘッドの左右耳のレスポンスとして，仰角-40 度から 90 度までの総計 710 方向のインパルス応答が含まれている。各インパルス応答の長さは 512 サンプルで，サンプリング周波数は 44.1 kHz である。

前述のように, HRTF を用いてダイナミックに音場を合成するには，頭部の向きのリアルタイム検出が必要である。このため，本研究ではヘッドフォンの上部にジャイロセンサーとコンパスを取り付け，頭部回転のトラッキングを行った。角度情報はシリアルおよびブルートゥース経由のいずれかでシステムに送られる。音場の合成に使う方向は音源方向から頭部角度を引いたもので，この方向に対応した左右チャンネルのインパルス応答がデータベースから選出され，分離結果と畳み込み演算を行った音声がオペレーターの両耳に再生される。

3 システム評価

提案システムを評価するため，被験者実験を行った。被験者はロボットを介してロボット側にいる人物と会話をし，ロボット側の視覚情報無しの状態で，その対話相手のいる方向を推定することが求められる。

比較対象として，ロボットの耳に位置するステレオマイクロフォンを用いた。この実験ではミニマルデザインされているヒューマノイドロボット Telenoid-R3（figure 3 左上）を使用した。このロボットは両耳位置にマイクの装着が可能で，且つ，首には 3 自由度があるため，人の頭部動作を線形的にマッピングすることができる。

以下に，比較対象の条件を述べる。この条件では，ロボットの耳にある二つのマイクロフォンから採った音を，そのままオペレーターのステレオヘッドフォンの左右チャンネルで再生する。トラッキングしたオペレーターの首の動きは，線形的にロボットにマッピングされる。

Figure 2. External appearance of the Telenoid R3 (top left), operator environment (bottom left) and the robot environment where interaction experiments were conducted (right).

Figure 2 の左下図にオペレーター側の環境を，右図にロボット側の環境の様子を示す。ロボット側の 3D 音源位置推定は, 3 つのマイクロフォンアレイによって行われた。Figure 2 右図に赤矢印で示してあるように，天井には直径 15 cm で 8 チャンネルのマイクが円形に配置されたマイクロフォンアレイが 2 つ設置してあり，卓上には直径 30 cm で 16 チャンネルのマイクが半球面上に配置されたマイクロフォンアレイが設置してある。

総計 20 名の被験者がこの実験に参加した。全て大学生で，ロボットや音響研究に関わりがない者である。被験者にはオペレーター役として，別室のロボット側にいる話者 1 名（研究補助者）とロボットを介して会話し，その相手のいる方向を判定するように指示した。実験補助者はランダムに方向を選び，その方向から会話を進める。被験者は方向の判定ができたら協力者に知らせ，協力者は次の方向に移動する。この手順を 4 回繰り返した。方向の判定は 8 方向に制限しており，被験者はそのうちのどの方向かを回答するという形式である。

実験の最後に，二つの条件について，臨場感と聞き取り易さに関する主観評価のアンケートを採った。1 から 7 までの七段階評価で，1 は「臨場感が低い/聞き取り難い」で，7 は「臨場感が高い/聞き取り易い」を示す。

Figure 3 上図に，提案システム条件と比較条件での方向定位の精度の平均値とその標準偏差を示す。T-test の結果，両者の精度差に有意差がみられた（$t = 0.59, p < 0.001$）。

主観評価アンケートでは，臨場感と聞き取り易さの評価で類似した結果が得られた。Figure 3 下図にその結果を示す。臨場感と聞き取り易さの両方において，提案システム条件での評価は，比較条件よりも有意に高い（$t = 6.68, p < 0.001$ と $t = 4.86, p < 0.001$）。

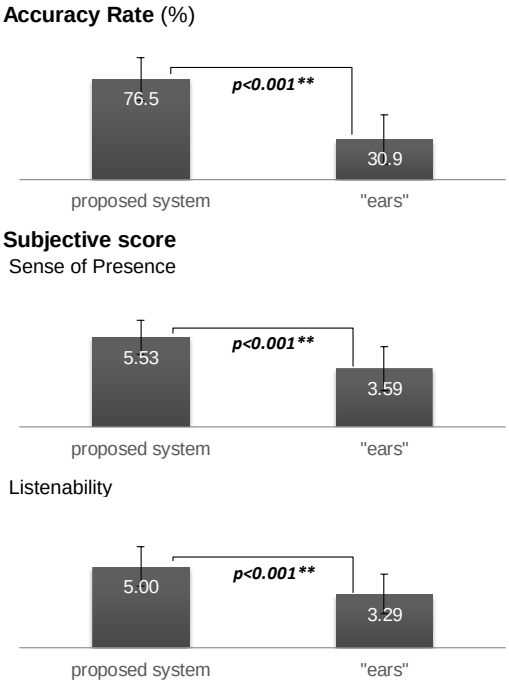


Figure 3. Accuracy rates for direction perception and subjective scores (1 to 7 scale) for sense of presence and listenability, in two conditions: “proposed system” and “robot’s ears”.

両条件で聞き取り易さに差が出た理由として，ロボットの両耳位置に埋め込まれたマイクロフォンの SNR が考えられる。このマイクロフォンはロボット内部のモーターに物理的に近いためモーターノイズの影響を受けやすく，これが SNR の低下に繋がったと考えられる。臨場感の評価にも両条件で有意差が見られたが，可能な理由としては，ロボットの首と人間の首の可動範囲が違うことが挙げられる。人間の首の可動範囲はロボットより広いため，オペレーターが首を回している途中でもロボットの首はすでに最大角度にヒットしている可能性がある。このオペレーターとロボットの頭部オリエンテーションのミスマッチが臨場感の評価に影響した可能性がある。

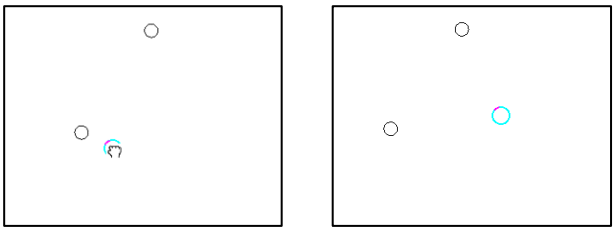
4 バーチャル音場における音源ボリュームの調整

提案システムでは，選択されたすべての音源に対して，位置情報を反映したステレオ音声を合成し，足し合わせて，バーチャル音場を表現する出力が再生される。しかし，これでは選択された各音源のボリュームが予測できない。もし，オペレーター側で各音源のボリュームを各々独立して操作することができれば，自分にとって最も快適な音環境を作ることができる。このことに注目して，オペレーターがバーチャル空間上にある音源や自分の位置を変えることができるように，二つのインタフェースを提案した。

4.1 提案のユーザインタフェース

このセクションでは，バーチャル音場をコントロールするための 2 つの異なる操作パターンのユーザインタフェースについて説明する。

Figure 4 に二つのインタフェースのスクリーンショットを示す。



(a) “drag-and-drop” (b) “face dir”

Figure.4. Screen shots of the displays for different user interfaces.

Figure 4 (a)に示す 1 つ目のインタフェースでは，オペレーターがスクリーン上の青い円（これはバーチャル空間上でのオペレーターの位置を表す）を任意の場所にマウスでドラッグ&ドロップすることによって，各音源のボリュームを調整する。希望の場所へ自身のバーチャルな位置を移動させることによって各音源との距離・角度が再計算され，音源のボリュームがその距離に従って変更される（特定の音源に接近させると，その音源のボリュームが大きくなる）。このインタフェースを“drag-and-drop”と表記する。実環境での会話シーンでは，会話参加者間の物理的距離は環境や相手との社会的関係に影響される。“drag-and-drop”は，この観点に注目したバーチャル音場コントロール法である。

Figure 4 (b)に示す 2 つ目のインタフェースでは，オペレーターの顔の向きによって各音源のボリュームが調整される。オペレーターの顔方向を利用して音源の音量を操作するため，両手が解放される。オペレーターの顔の前方にある音源は強調され，後方にある音源は減衰される。ボリュームを調節するファクタは角度と比例する。このインタフェースを“face dir”と表記する。顔の向きや視線方向は現時点における人の注意を示すだけでなく，次のターゲットやそのゴールをも示す[Langton 2000] [Yokoyama 2012]。“face dir”はこの観点に注目したバーチャル音場コントロール法である。

4.2 提案ユーザインタフェースの評価

提案のユーザインタフェースを評価するための被験者実験を行った。比較対象として，従来のモノラルマイクロフォンを使ったインタフェースを用いた。

前セクションで述べた実験被験者が，この実験にも参加した（大学生 16 名。前セクションの 20 名中最初の 4 名は従来法との比較を行っていないため除外）。実験のデザインは被験者内比較を採用した。被験者は提案インタフェース及び従来のインタフェースを使って，ロボット側の環境にいる対話者 2 名（研究補助者）と会話をする。会話トピックに制限はない。用いたインタフェースごとに会話のセッションを分けた。セッションの長さは 3 分間で，各セッション終了後にインタフェースの「使い易さ」「臨場感」「聞取り易さ」に関して前実験と同じく 1 から 7 まで 7 段階の主観評価アンケートを採った。

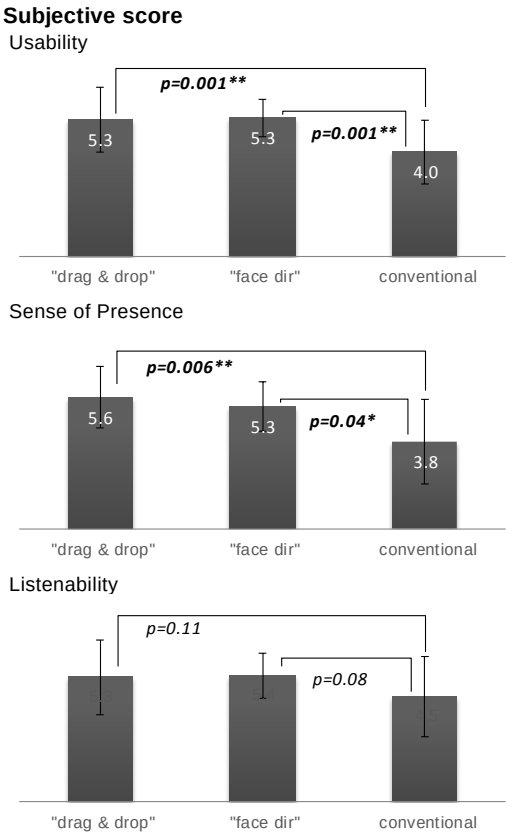


Figure 5. Subjective scores (1 to 7 scale) for three types of user interface: “drag and drop”, “face dir” and “conventional”.

Figure 5 に各インタフェースに対する主観評価の平均値と標準偏差を示す。実験結果に対して分散分析（ANOVA, with-in participants, Bonferroni’s posttest）を行った。

「使い易さ」（Figure 5 上図）と「臨場感」（Figure 5 中図）では，主観評価の平均値に有意差が見られた（*F(2,13)=16.03, p<0.001 and F(2,13)=6.74, p=0.009*）。多重比較（Bonferroni 法）の結果，提案法である“drag-and-drop”と“face dir”は従来法よりも使い易く（*“drag-and-drop” vs. “conventional”: p=0.001; “face dir” vs. “conventional”: p=0.001*），臨場感が高い（*“drag-and-drop” vs. “conventional”: p=0.006; “face dir” vs. “conventional”: p=0.04*）と評価された。しかし，「聞取り易さ」では有意差が見られなかった（*F(2,13)=3.67, p=0.052*）。

以上の結果は，提案インタフェースの有効性を示

している。

4.3 考察

ユーザインタフェースの評価実験は，興味深い結果を示している。通常、オペレーターとロボットは連動することで臨場感を感じるが、“drag-and-drop”インタフェース使用時では，被験者のみが自分（ロボット）の位置をバーチャル空間で変えるだけで、ロボットは実際に移動していないにも関わらず,「臨場感」の評価が高かった。

「聞取り易さ」の評価結果に関しては，提案インタフェースに対する評価スコアの平均値は従来法より高いものの，有意差が見られなかった。この可能性として，以下の理由が考えられる。今回の実験ではロボット側にいる対話者が２名のみであるため，多人数対話環境と比較して音の収録状況が良好である。そのため，従来法でも難なく音声を聴き取ることができたと考えられる。音源が増えるに連れて聞取り易さにも差が出る可能性があるが，これについての検証は今後行なう予定である。

また，今回の実験ではダミーヘッドの HRTF データベースを利用したが，被験者の頭部の形状に対応した HRTF を合成できれば，システムの効果の向上が期待できる。

5 おわりに

本稿では，操作者の頭部の動きに合わせて遠隔ロボットの環境の 3D 音場を合成する遠隔コミュニケーションロボット操作システムを提案し，被験者実験によってこれを評価した。

マイクロフォンアレイを用いて音源を収録し音場を合成する提案法は，ロボットの両耳にマイクを装着させて音源を収録した手法よりも，音源位置の同定実験では有意に高い精度を示し，臨場感と聞き取り易さの主観評価実験では、いずれも有意に高い評価が得られた。

また，バーチャル音場における音源のボリュームを操作するために 2 種類のユーザインタフェースを提案し，これを被験者実験によって評価した。

その結果，オペレーターがスクリーン上で音源に対する自身のバーチャルな位置を変更させてボリュームを調整する方法，及び，オペレーターの顔の向きに応じてボリュームを調整する方法は，従来法よりも「使い易さ」と「臨場感」の評価において有意に高く評価された。

謝辞

本研究は JST/CREST の委託研究により実施したものである。音源定位に関するシステムの一部は、総務省 SCOPE の委託研究により開発されたものを利用している。評価実験にご協力いただいた森田美香氏、波多野博顕氏に感謝する。

参考文献

[Nishio 2007] Nishio, S., Ishiguro, H., Hagita, N. Can a Teleoperated Android Represent Personal Presence? - A Case Study with Children. *Psychologia*, 50(4): 330-342. 2007.

[Ishi 2010] Ishi, C.T., Liu, C., Ishiguro, H., Hagita, N. 2010. Head motion during dialogue speech and nod timing control in humanoid robots. *In Proceedings of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010)*. OSAKA, JAPAN. 293-300.

[Liu 2012] Liu, C., Ishi, C. T., Ishiguro, H., Hagita, N. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. *In Proceeding of ACM/IEEE International Conference on Human Robot Interaction (HRI 2012)*. Boston, USA. 285-292, March, 2012.

[Sumioka 2014] Sumioka, H., Nishio, S., Minato, T., Yamazaki, R., Ishiguro, H. Minimal Human Design Approach for Sonzai-kan Media: Investigation of a Feeling of Human Presence. Cognitive Computation, 2014.

[Popescu 2000] Popescu, V. G., Burdea, G. C., Bouzit, M., Hentz, V. R. A virtual-reality-based telerehabilitation system with force feedback. *IEEE transactions on Information Technology in Biomedicine*. 4(1): 45-51. 2000.

[Piron 2009] Piron, L., Turolla, A., Agostini, M., Zucconi, C., Cortese, F., Zampolini, M., Zannini, M., Dam, M., Ventura, L., Battauz, M., Tonin, P. Exercises for paretic upper limb after stroke: a combined virtual-reality and telemedicine approach. *J. of Rehabilitation Medicine*. 41(12): 1016-1020(5). 2009.

[Billinghurst 2002] Billinghurst, M., Cheok, A., Prince, S., Kato, H. Real world teleconferencing. *IEEE Computer Graphics and Applications*. 22(6): 11-13. 2002.

[Ogi 2001] Ogi, T., Yamada, T., Tamagawa, K., Kano, M. Immersive telecommunication using stereo video avatar. *Proceedings of Ieee Virtual Reality*. Yokohama, Japan. 45-51. 2001

[Bullinger 1997] Bullinger, H., Riedel, O., Breining, R. Immersive Projection Technology- Benefits for the Industry, *International Immersive Projection Technology Workshop*, 13-25, 1997.

[Pulkki 2007] Pulkki, V. Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc*. 55(6): 503-516. 2007.

[Laitinen 2011] Laitinen, M., Kuech, F., Disch, S., Pulkki, V. Reproducing applause-type signals with directional audio coding. *J. Audio Eng. Soc*. 59(1/2): 29-43. 2011.

[Rumsey 2001] Rumsey, F. *Spatial Audio*. Focal Press, 2001.

[Meyer 1972] Meyer, E., Neumann, E. *Physical and Applied Acoustics: An Introduction*. Academic Press, New York, 1972. ISBN 0124931502.

[Cheng 2001] Cheng, C. I., Wakefield, G. H. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *J. Acoust. Soc. Am*, 49(4):231-249, April 2001.

[Iwaya 2003] Iwaya, Y., Suzuki, Y., Kimura, D. Effects

of head movement on front-back error in sound localization. *Acoustical Science and Technology*. 24(5): 322-324. 2003.

[Schmidt 1986] Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34, 276-280, 1986.

[Ishi 2009] Ishi, C. T., Chatot, O., Ishiguro, H., Hagita, N. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. Proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 09)*. 2027‑2032. 2009.

[Glas 2007] Glas, D.F. et al, 2007. Laser tracking of human body motion using adaptive shape modeling. In Proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, 602-608. 2007.

[Ishi 2013] Ishi, C., Even, J., Hagita, N. (2013). Using multiple microphone arrays and reflections for 3D localization of sound sources. In Proc. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, 3937-3942, Nov., 2013.

[石井 2014] 石井カルロス寿憲, Jani EVEN, 萩田紀博, (2014) "複数のマイクロホンアレイと人位置情報を組み合わせた音声アクティビティの記録システムの改善", 第３２回日本ロボット学会学術講演会, Sep. 2014.

[Dudgeon 1977] Dudgeon, D. E. Fundamentals of digital array processing. *Proceedings of the IEEE*. 65(6): 898-904. 1977.

[Gardner 1995] Gardner, W. G., Martin, K. D. HRTF measurements of a KEMAR. *J. Acoust. Soc. Am*. 97(6):3907-3908, Jun. 1995.

[Langton 2000] Langton, S. R., Watt, R. J., Bruce, I. I. Do the eyes have it? Cues to the direction of social attention. *Trends Cog. Sci*. 4, 50–59, 2000.

[Yokoyama 2012] Yokoyama, T., Noguchi, Y. Kita, S. Attentional shifts by gaze direction in voluntary orienting: evidence from a microsaccade study. *Exp. Brain Res.* 223, 291–300, 2012

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-06

非同期分散マイクロフォンアレーによる音源定位・音源分離

Source Localization and Separation with Asynchronous and Distributed Microphone Array

小野 順貴
Nobutaka ONO
国立情報学研究所 / 総合研究大学院大学
National Institute of Informatics / The Graduate University for Advanced Studies (SOKENDAI)
onono@nii.ac.jp

Abstract

本稿では，複数の録音機器を分散配置させ，これらをマイクロフォンアレーの素子として用いる，非同期分散マイクロフォンアレーという枠組みについて紹介する。従来のマイクロフォンアレー信号処理においては，チャンネル間の微小な時間差が空間情報の大きな手がかりであり，各チャンネルを正確に同期させるために，全てのマイクロフォンは多チャンネル A/D 変換器に接続されることが前提であった。これに対し，我々の身の周りには，ラップトップ PC，ボイスレコーダー，スマートフォンなどの録音機器が多数存在する。こうした機器によりマイクロフォンアレー信号処理が可能になれば，その利便性は大きく，適用範囲を格段に広げることが期待できる。本稿では，非同期録音機器を用いたマイクロフォンアレー信号処理の新しい展開について，関連研究を概観しつつ，著者らの取り組みを紹介する。

1 はじめに

マイクロフォンアレーは，複数のマイクロフォンにより音場の空間的な情報を取得し，単一マイクロフォンでは困難な，音源定位，音源強調，音源分離などを行う枠組みである。一般には，用いられるマイクロフォンの数が多いほど得られる空間情報が多くなるため，制御できる指向性の自由度が増加し，また，マイクロフォンを広範囲に配置することができるほどカバーできる範囲が広がり，定位や分離の性能向上が期待できる。

しかしながら，マイクロフォンアレー信号処理においては，厳密な同期録音が必要不可欠であることが大きな制約条件となっている。これは，マイクロフォンアレー信号処理では，各マイクロフォンで録音される信号間の微小な時間差（例えば，経路長 3.4 cm に対して 100 μs）が音源の空間情報の主要な手がかりとなっているためである。よって従来は，各チャンネルを正確に同期させるために，全てのチャンネルは多チャンネル A/D 変換器に接続され，同一クロックによりサンプリングされる必要があり，これがマイクロフォンアレーの多素子化や分散配置などに対して大きなコストを生じる主要因の一つとなっていた。

一方，我々の身の周りには，録音機能を持つ機器が多数存在している。音を録音することが目的であるボイスレコーダーや，通話を目的としたスマートフォンにとどまらず，ラップトップ PC やタブレット端末の多くも録音機能を有しているし，動画撮影機能を持つデジタルカメラやビデオカメラも録音機器として用いることができる。こうした別々の録音機器による多チャンネル録音に基づくアレー信号処理の枠組みは，近年，非同期分散マイクロフォンアレー，アドホックマイクロフォンアレーなどと呼ばれ，国内外で関心が高まりつつある [1]。

非同期分散録音機器でアレー信号処理が可能になれば，以下のような利点が期待できる。

1. 従来のように，同時サンプリング可能な多チャンネル A/D 変換器が不要なため，録音機器を増やすだけで，マイクロフォンアレーの素子数を容易に増やすことができる。
2. 広範囲に素子を分散配置することができる。
3. 録音機器間の有線接続の必要がなく，ワイヤレスのシステムを容易に構築することができる。

本稿では，非同期分散マイクロフォンアレーの新しい展開について，関連研究と共に著者らの取り組みを紹介する。なお本稿では特に，複数の機器を用いて録音した多チャンネル信号に対する信号処理を想定し，全チャンネルの信号が利用可能であることを仮定した音源定位と音源分離の問題を中心に紹介する。著者らの別の解説論文 [2, 3, 4] も併せて参照いただきたい。

また，本稿では扱わないが，マイクロフォン間の通信を仮定し，音のセンサネットワークを構築することを目指すワイヤレスアコースティックセンサネットワークという枠組みでは，各マイクロフォンは近傍のマイクロフォンとした通信できない，つまり全チャンネルの情報が利用できるわけではないとう制約の元で，どのように信号処理を行うか，といった問題も盛んに議論されている。興味のある読者は [1] や国際会議 EUSIPC2013 のチュートリアル資料 ([5] から参照可) などを参照していただきたい。

2 音源定位

2.1 マイクロフォン位置の自己校正の必要性

従来のアレー信号処理に用いられるマイクロフォンは，直線，円周のような規則的な配置に並べられ，各マイクロフォンの位置は既知であることが前提である。しかしながら非同期分散マイクロフォンアレーでは，個々の録音機器は有線接続すらされておらず，通常，その位置座標は未知であることが多い。そのため，音源位置を推定するためには，まずマイクロフォン位置を推定することが必要となる。スマートフォンなどを用いる場合には GPS (Global Positioning System) 情報が利用できる場合もあるが，室内環境での測位精度は一般に十分でなく，マイクロフォンが観測された音自体を用いて推定する自己校正 (self-calibration) が重要となる。

なお，位置推定は一般にアレー信号処理への応用に限らず重要であり，Indoor Positioning and Indoor Navigation (IPIN)[6] という室内での機器位置推定に関する国際会議が 2010 年以降毎年開催され，音に限らず，電波や光を含めた定位手法が議論されている。

2.2 ブラインドアラインメント

マイクロフォン位置や音源位置を推定するための観測量としては，到来時間 (Time of Arrival; TOA) や到来時間差 (Time Difference of Arrival; TDOA) がよく用いられる。ここでは録音機器が同期しておらず，音源も未知な場合，各音源に対して観測された TDOA のみから音源位置，マイクロフォン位置，各マイクロフォンの時間原点を推定するという問題を考えてみよう。我々はこれを，観測信号のみから (ブラインド)，音源，マイクロフォン，チャンネルを共通の時空間座標系に揃える (アラインメント) という意味で，ブラインドアラインメントと呼んでいる [7, 9]。

M 個のマイクロフォン，N 個の音源，の位置座標をそれぞれ $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_M$，$\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N$ とする。ここでは $\boldsymbol{r}_m$，$\boldsymbol{s}_n$ は p 次元のベクトルとし，$p = 2$ または $p = 3$ とする。また，各マイクロフォンはそれぞれ別の時間原点をもっており，それらを t_m で表わす。簡単のため，録音機器間のサンプリング周波数ミスマッチは，ここでは無視できると仮定する。

まず，音源 $\boldsymbol{s}_n$ に対する TDOA を求めることを考える。チャンネルが非同期である場合，正しい TDOA は直接には求まらず，マイクロフォン i，j の相互相関ピークから得られるチャンネル間の時間差 (みかけの TDOA) は，

$$\hat{\tau}_{nij} = \left(\frac{||\boldsymbol{s}_n - \boldsymbol{r}_i||_2}{c} - t_i \right) - \left(\frac{||\boldsymbol{s}_n - \boldsymbol{r}_j||_2}{c} - t_j \right) \quad (1)$$

のように，未知の録音開始時刻 t_i，t_j を含むことになる。よって，みかけの TDOA が得られたとしても，これが到来時間差によるものなのか，マイクロフォンの時間原点差によるものなのかはわからない。しかし，

$$(K - p - 1)(M - p - 1) \geq \frac{p(p+3)}{2} \quad (2)$$

の場合には，未知数の数より多くの観測量が得られる [7]。よって，TDOA に基づくチャンネル同期とマイクロフォン位置定位の問題は，式 (1) を観測方程式とし，多数得られた $\hat{\tau}_{nij}$ から $\boldsymbol{s}_n$，$\boldsymbol{r}_m$，t_m を推定する問題と考えることができる。我々は，補助関数法 [8] という最適化のアプローチを用いて式 (1) の最小二乗解を求める解法 [7, 9] を導出している。また，近年ロボットへの応用を想定したオンラインでのキャリブレーション [10] も検討されている。

2.3 距離行列のもつランク制約

チャンネルが同期していたとしても，マイクロフォンと音源の同時位置推定は難しい問題である。音速を既知とすれば TOA や TDOA は距離情報を与える。一般に距離情報から座標を推定する問題は，多次元尺度法 (Multi Dimensional Scaling; MDS) という手法によって解くことができるが，マイクロフォンと音源の位置推定問題の場合には，マイクロフォン間，音源間の距離情報は得ることができず，マイクロフォンと音源の間の距離情報しか得られないことが難しさの一因となっている。ここでは，こうした問題を解く上で大変有用な，距離行列の性質について触れておきたい。

いま，距離行列 $D = (D_{mn})_{M \times N}$，距離差行列 $\tilde{D} = (\tilde{D}_{mn})_{(M-1)\times(N-1)}$ を，マイクロフォンと音源の距離の 2 乗を用いて以下のように定義する。

$$\begin{aligned} D_{mn} &= \|\boldsymbol{r}_m - \boldsymbol{s}_n\|_2^2 = (\boldsymbol{r}_m - \boldsymbol{s}_n)^T(\boldsymbol{r}_m - \boldsymbol{s}_n) \\ &= ||\boldsymbol{r}_m||^2 + ||\boldsymbol{s}_n||^2 - 2\boldsymbol{r}_m \cdot \boldsymbol{s}_n \quad (3) \\ \tilde{D}_{mn} &= D_{m+1,n+1} - D_{m+1,1} - D_{1,n+1} + D_{1,1} \quad (4) \end{aligned}$$

このとき以下が成り立つ。

補題 1 距離行列 D のランクは高々$(p+2)$ である。

証明: 式 (3) より，

$$D = T^T I_N + I_M^T U - 2R^T S \quad (5)$$

とあらわせる。ただし，T は m 番目の要素が $||\boldsymbol{r}_m||^2$ である $1 \times M$ のベクトル，U は n 番目の要素が $||\boldsymbol{s}_n||^2$ である $1 \times N$

のベクトル，I_M，I_N は要素が全て 1 の $1\times M$，$1\times N$ のベクトル，$R=(\boldsymbol{r}_1\cdots\boldsymbol{r}_M)$ は $p\times M$ 行列，$S=(\boldsymbol{s}_1\cdots\boldsymbol{s}_N)$ は $p\times N$ 行列であり，また、^{T} は行列の転置を表す。各項のランクは高々，1，1，p であるから，D のランクは高々 $(p+2)$ である。 ■

補題 2 距離差行列 $\tilde{D}$ のランクは高々 p である。

証明: $\tilde{R}=(\boldsymbol{r}_2-\boldsymbol{r}_1\cdots\boldsymbol{r}_M-\boldsymbol{r}_1)$，$\tilde{S}=(\boldsymbol{s}_2-\boldsymbol{s}_1\cdots\boldsymbol{s}_N-\boldsymbol{s}_1)$ とおくと，$\tilde{D}=-2\tilde{R}\tilde{S}^T$ とあらわされる。$\tilde{R}$ は $p\times(M-1)$ 行列，$\tilde{S}$ は $p\times(N-1)$ 行列であるので，$\tilde{D}$ のランクは高々 p である。 ■

任意の配置に対して，距離行列がこのようなランク制約をもつことは興味深い。近年，このランク制約を用いた TOA，TDOA ベースの位置推定法が提案されている [11, 12, 13]。我々も現在，いくつかのアルゴリズムを研究中 [14, 15] であり，今後はブラインドアラインメントへ応用していきたいと考えている。

2.4 音の発信を利用した機器位置推定

スマートフォンなど，音を発することができる録音機器を利用できる場合には，音の発信を積極的に利用するアプローチが考えられる。例えば，1 台のスマートフォンに装備されているスピーカーとマイクの位置は厳密には異なっているが，これを近似的に等しい $(\boldsymbol{r}_m\simeq\boldsymbol{s}_m)$ とし，2 台のスマートフォンから互いに音を発信して TDOA を求めたとすると，

$$\hat{\tau}_{mn}=-\frac{||\boldsymbol{r}_m-\boldsymbol{r}_n||_2}{c}-t_m+t_n \tag{6}$$

$$\hat{\tau}_{nm}=\frac{||\boldsymbol{r}_n-\boldsymbol{r}_m||_2}{c}-t_m+t_n \tag{7}$$

と表せるので，

$$||\boldsymbol{r}_m-\boldsymbol{r}_n||_2=\frac{c}{2}(\hat{\tau}_{nm}-\hat{\tau}_{mn}) \tag{8}$$

$$t_m-t_n=-\frac{1}{2}(\hat{\tau}_{mn}+\hat{\tau}_{nm}) \tag{9}$$

のように，互いの TOA から距離と時間原点の差が直接的に求まる [16, 17, 18]。各機器間の距離が求まれば，あとは前述の多次元尺度法により相対位置を決めることができる。

我々はさらに，音の発信を利用し，位置と時間原点のキャリブレーションだけでなく，サンプリング周波数ミスマッチもあわせて補償する手法を提案している [19]。図 1 に，4 台の iPod touch による移動音源定位の実験結果例を示す。それぞれから発信された TSP (Time-Stretched Pulse) 信号 [20] を用いて，各機器の位置，録音開始時刻の推定，サンプリング周波数ミスマッチを推定してキャリブレーションを行った後にスピーカーを音源定位したものであり，移動音源であるスピーカーの定位がうまく行われていることがわかる。

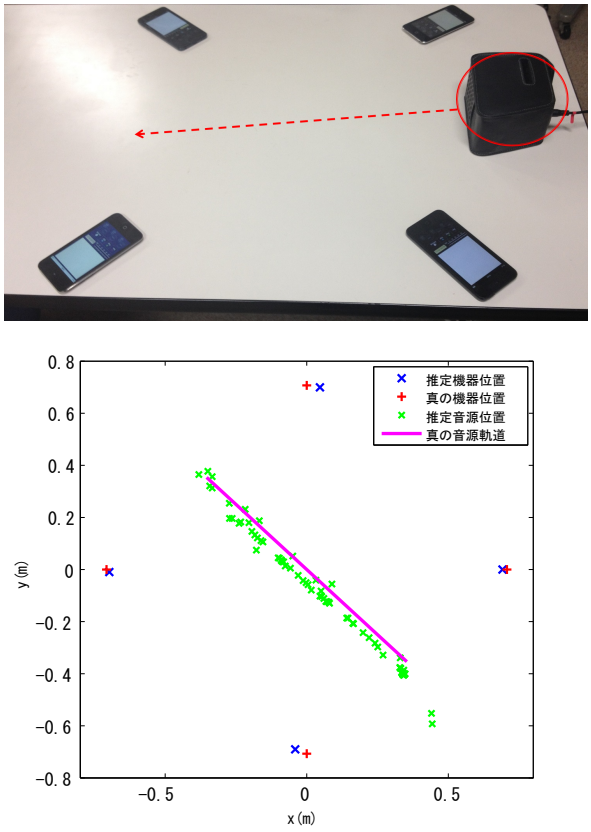


図 1: 非同期分散マイクロフォンアレーによる移動音源定位実験の写真 (上) と実験結果 (下) [19]

3 音源分離

3.1 チャンネル非同期がアレー信号処理に与える影響

一般に非同期分散マイクロフォンアレーにおいては，通常マイクロフォン位置は未知であり，各録音チャンネルは同期していない。いわゆるビームフォーミングや音源定位を行うためには，マイクロフォン位置の推定も必要となるが，SN 比最大化ビームフォーマ [21] や独立成分分析 [22] など，音源強調，音源分離手法の中には，マイクロフォンの位置情報を必要としない手法もある。そこでここでは主に，チャンネル非同期の影響について考える。

非同期の主な要因には，1) 録音を開始する時刻が同一でないこと，2) サンプリング周波数が同一でないこと，の 2 つがある。前者は定常的な時間軸シフトを，後者は時間軸の伸縮をもたらす (図 2 参照)。定常的な時間シフトについては，信号間の相互相関が最大となるように信号をシフトすることで十分な場合も多い。音源からマイクまでの伝達関数は，SN 比最大化ビームフォーマにおいては学習区間の信号から，独立成分分析においてはブラインドに推定されるため，たとえ信号間に小さな一定のシフト誤差が残っていても，あまり問題にならないためである。

一方，録音機器 1, 録音機器 2 のサンプリング周波数を f_1，f_2 とすると，

$$\varepsilon=\frac{f_2}{f_1}-1 \tag{10}$$

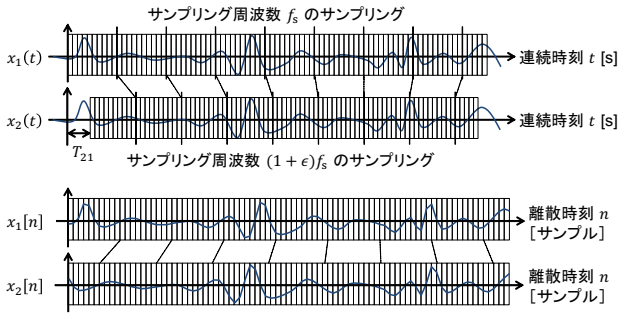


図 2: サンプリング周波数ミスマッチの影響：連続波形 (上)，離散波形 (下)

が，録音機器間の相対的なサンプリング周波数ミスマッチを表わす無次元量となる。ε は，10 ppm (ppm は parts per million で 10^{-6} を表わす) の数倍程度に収まることが多い。しかし，こうしたわずかなミスマッチがアレー信号処理に与える影響は大きい。

例えば，録音機器 1，2 のサンプリング周波数を 16000 Hz，16000.5 Hz とし (サンプリング周波数ミスマッチは 31.25 ppm)，これらを 0.2 m だけ離した配置で，正面方向の音源信号を 10 秒間録音を行ったとしよう。簡単のため，録音は同時に開始したと考える。このとき 10 秒間の音響信号は，録音機器 1 では 160000 サンプル，録音機器 2 では 160005 サンプルに相当し，サンプリング周波数ミスマッチにより，10 秒間の録音信号の最後では 5 サンプルの時間差が生じる。一方，30 度方向から到来する音波に対する到来時間差は，$0.2\sin 30°/340.0\times 16000 \simeq 4.7$ で，約 4.7 サンプルに相当する。つまり，31.25 ppm という微小なサンプリング周波数のずれが引き起こす時間差は，10 秒間に音源が正面から 30 度方向に移動したのと区別がつかないことになる。多くの線形アレー信号処理においては，音源からマイクロフォンまでの伝達関数は線形時不変であることが仮定されているため，こうしたチャンネル間の時間差のドリフトは，音源分離に深刻な破綻を引き起こすことになる [23, 24]。

3.2 サンプリング周波数ミスマッチのブラインド補償に基づくブラインド音源分離

サンプリング周波数ミスマッチの問題に対しては，ミスマッチを推定し，補償した後に，従来のアレー信号処理を適用する方法が考えられる。前述のように，サンプリング周波数ミスマッチは離散波形の伸縮を引き起こすため，同一音源から到来する信号を 2 回観測できれば，それらの間隔の比率からサンプリング周波数ミスマッチが推定できる [25] し，観測信号のみからのブラインド推定と補償も検討されつつある [24, 26]。

前述のようにサンプリング周波数ミスマッチは，音源が移動しているかのように，時間とともに拡大するチャンネル間時間差を生じる。そこで我々は，音源が動かないことを仮定し，STFT (Short-Time Fourier Transform) 領域で定常な確率モデルに基づく最尤法によりサンプリング周波数ミスマッチをブラインドに推定し，補償する手法を提案している [27, 28, 29, 30, 31, 32]。

図 3 は，複数話者の混合音声を録音した，複数の長さのデータに対するサンプリング周波数の補正精度を平均二乗誤差 (RMSE) で評価した実験結果の例 [28, 29] である。最も短い 3 秒の観測信号でも RMSE は元のサンプリング周波数ミスマッチ ε の 10 分の 1 以下に収束し，データ長が増えるにつれて急速に小さくなることがわかる。次に，音源分離への適用例を図 4 に示す。分離フィルタの学習は，補助関数型独立ベクトル分析 [33] により行った。分離性能評価には BSS Toolbox [34] を用い，評価尺度としては SDR (Signal to Distortion Ratio) を用いた。まず未処理の分離性能が非常に低い値を示しており，サンプリング周波数ミスマッチの補償をしなければ音源分離ができない厳しい条件であるということが分かる。ε の正しい値を与えた位相補償 (図中位相補償) は，サンプリング周波数ミスマッチがない場合よりも SDR が 2 dB 程度低いだけであり，STFT 領域におけるミスマッチ補償法の有効性が確認できる。また，サンプリング周波数ミスマッチをブラインドに推定して補償する提案手法は，正しい ε を与えた場合とほとんど性能差がなく，提案手法の最尤推定は十分高い精度で ε を推定していることを示している。

以上より，非同期録音であっても，音源が動かないと仮定できる場合には，こうした手法によりブラインドにチャンネルを同期化し，ブラインド音源分離が適用できることがわかる。また我々は，符号化録音に対しても，本手法が破綻せず動作することを実験により確認している [35]。

3.3 伝達関数ゲイン基底 NMF を用いた非同期録音に対する音源分離

非同期分散マイクロフォンアレーでは，アレー素子として別々の録音機器を用いるため，関心のある個々の音源の近くにマイクロフォンを配置できる場合も多い。こうした場合にはチャンネル間の時間差だけでなく，チャンネル間の振幅比も音源を区別する重要な手がかりとなる。そこで我々は，サンプリング周波数ミスマッチを正確に推定・補償する代わりに，ミスマッチに頑健な振幅スペクトル情報のみを用いた音源分離手法についても検討している。

いま，サンプリング周波数ミスマッチの影響により位相情報は信頼できないと考え，通常の線形混合モデルの代わりに，振幅スペクトルに対する近似的な線形混合モデルを仮定する。すなわち，

$$\bar{\boldsymbol{X}}(\tau,\omega) \simeq \bar{A}(\omega)\bar{\boldsymbol{S}}(\tau,\omega) \tag{11}$$

ただし，$\bar{\boldsymbol{X}}(\tau,\omega)$，$\bar{\boldsymbol{S}}(\tau,\omega)$ は，$\boldsymbol{X}(\tau,\omega)$，$\boldsymbol{S}(\tau,\omega)$ の要素ごとに絶対値をとったベクトル (振幅スペクトルベクトル)，

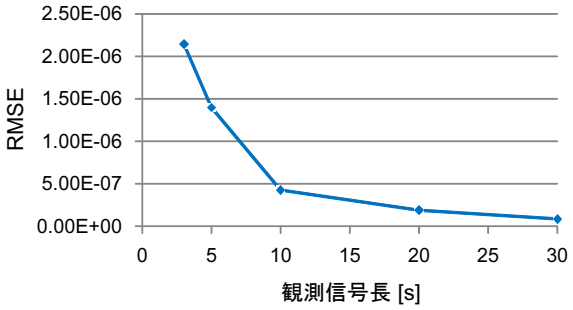


図 3: 推定されたサンプリング周波数ミスマッチ ε の平均二乗誤差 [28]

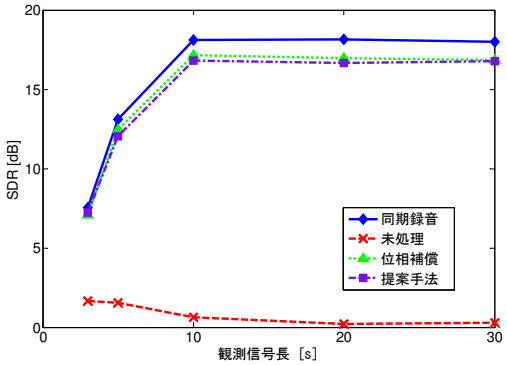


図 4: サンプリング周波数ミスマッチ補償のブラインド音源分離による性能評価 [28, 29]

$\bar{A}(\omega)$ は，各音源から各マイクロフォンまでの伝達関数のゲインを要素とする混合行列である。

このような振幅領域での混合モデルを用いたビームフォーマ [36] も提案されているが，行列が全て非負であることから，非負値行列分解 (Non-negative Matrix Factorization; NMF) の適用も考えられる [37]。ただし，時間周波数領域での NMF と異なり，時間チャンネル領域での NMF の場合には，優決定であったとしても，チャンネル数（観測行列の行数）と音源数（基底数）がそれほど大きくは変わらないことが多い。そのような場合には，NMF の低ランク近似としての拘束力が弱く，通常の NMF で観測行列をブラインドに分解することは難しい。

よって我々はまず，各音源に対する単一音源観測が教師信号として得られることを仮定し，あらかじめ $\bar{A}(\omega)$ を学習した後，観測信号 $\bar{\boldsymbol{X}}(\tau,\omega)$ に対して，$\bar{\boldsymbol{S}}(\tau,\omega)$ のみを更新していく，教師あり NMF の適用を非同期録音に対して検討した。以下では，実際の会議録音を想定した非同期録音データにより，教師あり NMF による音源分離の性能を評価した例 [38] を示す。

分離性能を定量評価するために，まず話者ごとに同期録音を行い，その後，話者ごとの録音を足し合わせて混合した後，マイクごとに，サンプリング周波数が 16000, 16001, 16002, 16003Hz になるようリサンプリングを行って，人工的に非同期録音データを生成したものを対象とした。分離性能評価には前述の BSS Toolbox [34] を用い，評価尺度としては SDR を用いた。また評価は，未処理 (No processing) のほか，教師あり伝達関数ゲイン NMF を板倉斎藤ダイバージェンス規準 (SNMF-ISdiv)，I ダイバージェンス規準 (SNMF-Idiv)，フロベニウスノルム規準 (SNMF-Fnorm) で適用したものに対して行った。

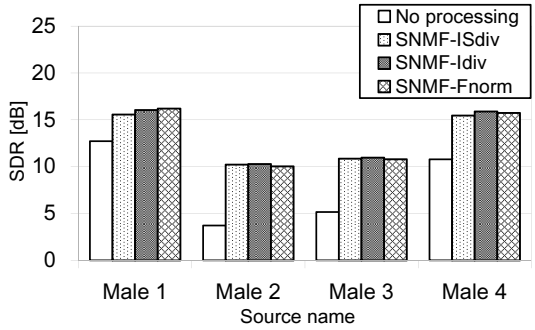


図 5: 非同期録音に対する伝達関数ゲイン基底 NMF の音源分離性能 [38]

結果を 図 5 に示す。未処理の場合と比較して，教師ありの伝達関数ゲイン基底 NMF では各観測信号において SDR が大きく向上している。また，ここでは結果を省略しているが，同期録音と非同期録音の結果にほとんど差が見られなかった。よってこの手法が非同期録音に頑健であり，非同期分散マイクロフォンアレーにおける目的音強調に適した手法であることが確認できる。我々はその他，マイク数・マイク配置との関係 [39] や，スパース性を導入したブラインド化 [40] についても検討を進めている。

4 おわりに

本稿では，非同期分散マイクロフォンアレーという新しいアレー信号処理の枠組みについて紹介した。十分紹介しきれなかったトピックもあるが，引用した文献をあわせて参照していただければ幸いである。

謝辞

本稿で紹介した研究は，筑波大学牧野研究室，東京大学亀岡研究室との共同研究によるものである。ここに共同研究者各位へ謝意を表する。また，本研究の一部は，科学研究費補助金基盤研究 (B) (課題番号 25280069) の助成を受けて行われたものである。

参考文献

[1] A. Bertrand, “Applications and trends in wireless acoustic sensor networks: a signal processing perspective,” Proc. Symposium on Communications and Vehicular Technology (SCVT), 2011.

[2] 小野 順貴, Trung-Kien Le, 宮部 滋樹, 牧野 昭二, “アドホックマイクロフォンアレー –複数のモバイル録音機器で行う音響信号処理–,” 電子情報通信学会 Fundamental Review, vol. 7, no. 4, pp. 336–347, 2014.

[3] https://www.jstage.jst.go.jp/article/essfr/7/4/7_336/_pdf

[4] 小野 順貴, 宮部 滋樹, 牧野 昭二, “非同期分散マイクロフォンアレーに基づく音響信号処理,” 日本音響学会誌, vol. 70, no. 7, pp. 391–396, 2014.

[5] http://www.eurasip.org/Proceedings/Eusipco/Eusipco2013/

[6] http://ipin-conference.org/

[7] N. Ono, H. Kohno, N. Ito and S. Sagayama, “Blind alignment of asynchronously recorded signals for distributed microphone array,” Proc. WASPAA, pp.161–164, Oct. 2009.

[8] 小野 順貴, “補助関数法による最適化アルゴリズムとその音響信号処理への応用,” 日本音響学会誌, vol. 68, no. 11, pp. 566–571, 2012.

[9] K. Hasegawa, N. Ono, S. Miyabe and S. Sagayama, “Blind estimation of locations and time offsets for distributed recording devices,” Proc. LVA/ICA, pp. 57–64, Sept. 2010.

[10] H. Miura, T. Yoshida, K. Nakamura and K. Nakadai, “SLAM-based online calibration of asynchronous microphone array for robot audition,” Proc. IROS, pp. 524–529, 2011.

[11] M. Crocco, A. Del Bue and V. Murino, “A bilinear approach to the position self-calibration of multiple sensors,” IEEE Trans. Signal Process., vol. 60, no. 2, pp. 660–673, Feb. 2012.

[12] F. Jiang, Y. Kuang and Kalle Åström, “Time delay estimation for TDOA self-calibration using truncated nuclear norm regularization,” Proc. ICASSP, pp. 3885–3889, May, 2013.

[13] Y. Kuang and Kalle Åström, “Stratified sensor network self-calibration from tdoa measurements,” Proc. EUSIPCO, 2013.

[14] T. -K. Le and N. Ono, “Numerical Formulae for TOA-based Microphone and Source Localization” Proc. IWAENC, pp. 179–183, Sept. 2014.

[15] T. -K. Le and N. Ono, “Reference-distance Estimation Approach for TDOA-based Source and Sensor Localization,” Proc. ICASSP, 2015. (submitted)

[16] V. Raykar, I. Kozintsev and R. Lienhart, “Position calibration of microphones and loudspeakers in distributed computing platforms,” IEEE Trans. Speech Audio Process., vol. 13, no. 1, pp. 70–83, 2005.

[17] M. Hennecke, T. Pl´otz G. A. Fink, J. Schmalenstr´oer and R. H´ab-Umbach, “A hierarchical approach to unsupervised shape calibration of microphone array networks,” Proc. SSP, pp. 257–260, 2009.

[18] 柴田 一暁, 小野 順貴, 亀岡 弘和, “音の発信を利用したスマートフォンアレイの機器位置推定,” 音講論 (秋), pp. 591–592, 9 月, 2013 年.

[19] 柴田 一暁, 小野 順貴, 亀岡 弘和, “音の発信を利用したキャリブレーションに基づくアドホックマイクロホンアレイによる音源定位,” 音講論 (春), pp. 707-710, 3 月, 2014 年.

[20] Y. Suzuki, F. Asano, H.-Y. Kim, and Toshio Sone, “An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses,” J. Acoust. Soc. Am., vol. 97, no. 2, pp. 1119–1123, 1995.

[21] H. L. Van Trees, Ed., Optimum Array Processing, John Wiley & Sons, 2002.

[22] A. Hyvärinen, J. Karhunen and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.

[23] E. Robledo-Arnuncio, T. S. Wada and B.-H. Juang, “On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation,” Proc. WASPAA, pp. 34–37, Oct. 2007.

[24] Z. Liu, “Sound source separation with distributed microphone arrays in the presence of clock synchronization errors,” Proc. IWAENC, 2008.

[25] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada and Shoji Makino, “Speech enhancement with ad-hoc microphone array using single source activity,” Proc. APSIPA, Oct. 2013.

[26] S. Markovich-Golan, S. Gannot and I. Cohen, “Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming,” Proc. IWAENC, 2012.

[27] 宮部 滋樹, 小野 順貴, 牧野 昭二, “非同期録音信号の線形位相補償によるブラインド同期と音源分離への応用,” 音講論 (秋), pp. 689–690, 9 月, 2012 年.

[28] 宮部 滋樹, 小野 順貴, 牧野 昭二, “非同期録音ブラインド同期のための線形位相補償の効率的最尤解探索,” 音講論 (春), pp. 733–734, 3 月, 2013 年.

[29] S. Miyabe, N. Ono and S. Makino, “Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain,” Proc. ICASSP, pp.674–678, May, 2013.

[30] 宮部 滋樹, 小野 順貴, 牧野 昭二, “非整数サンプルシフトのフレーム分析を用いた非同期録音の同期化,” 音講論 (秋), pp. 593–596, 9 月, 2013 年.

[31] S. Miyabe, N. Ono and S. Makino, “Optimizing frame analysis with non-integer shift for sampling mismatch compensation of long recording,” Proc. WASPAA, Oct. 2013.

[32] S. Miyabe, N. Ono and S. Makino, “Blind Compensation of Interchannel Sampling Frequency Mismatch for Ad hoc Microphone Array Based on Maximum Likelihood Estimation,” Elsevier Signal Processing (to appear)

[33] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” Proc. WASPAA, pp. 189–192, Oct. 2011.

[34] E. Vincent, C. Fevotte and R. Gribonval, “Performance measurement in blind audio source separation,” IEEE Trans. Acoust. Speech Signal Process., vol. 14, no. 4, pp. 1462–1469, 2006.

[35] 宮部 滋樹, 小野 順貴, 牧野 昭二, 高橋 祐, “非同期マイクロホンアレーの符号化録音におけるビットレートと同期性能の関係,” 音講論 (春), pp. 725–726, 3 月, 2014 年.

[36] 加古 達也, 小林 和則, 大室 仲, “非同期分散マイクアレーのための振幅スペクトルビームフォーマの提案,” 音講論 (春), pp. 829–830, 3 月, 2013 年.

[37] 戸上 真人, 川口 洋平, 小窪 浩明, 大淵 康成, “音源のチャンネル間振幅差を基底ベクトルとする音源分離,” 音講論 (春), pp. 803–804, 3 月, 2010 年.

[38] 千葉 大将, 小野 順貴, 宮部 滋樹, 山田 武志, 牧野 昭二, 高橋 祐, “伝達関数ゲイン基底 NMF による分散配置非同期録音における目的音強調の検討,” 音講論 (春), pp. 757–760, 3 月, 2014 年.

[39] 村瀬 慶和, 千葉 大将, 小野 順貴, 宮部 滋樹, 山田 武志, 牧野 昭二, “伝達関数ゲイン基底 NMF におけるマイク数・マイク配置と目的音強調性能の関係,” 音講論 (秋), pp. 523–526, 9 月, 2014 年.

[40] 千葉 大将, 小野 順貴, 宮部 滋樹, 山田 武志, 牧野 昭二, 高橋 祐, “教師なし伝達関数ゲイン基底 NMF による目的音強調における罰則項の特性評価,” 音講論 (秋), pp. 527–530, 9 月, 2014 年.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-07

マイクアレイ伝達関数のオンライン校正とそのロボットへの適用

Online Calibration of Microphone Array Transfer Functions for Robots

中村圭佑, 中臺一博

Keisuke NAKAMURA, Kazuhiro NAKADAI

（株）ホンダ・リサーチ・インスティチュート・ジャパン

Honda Research Institute Japan Co., Ltd.

keisuke@jp.honda-ri.com, nakadai@jp.honda-ri.com

Abstract

本稿ではマイクアレイベースのロボット聴覚システムで事前情報として用いられるマイクアレイ伝達関数の校正について述べる．伝達関数は主に数値計算と計測の二つの方法によって得られるが，数値計算はロボットや部屋の形状に起因する反射や回折などを誤差なく模擬することは難しく，計測は正確であるものの時間がかかり，専門の機器を必要とする．そこで，本稿ではロボットや部屋の音響特性を含めたオンザスポット（ユーザーがその場で手軽にできる）オンライン伝達関数校正法を提案する．

1 序論

ロボットと人が自然なインタラクションを実現するには，ロボット聴覚 [1] を用いた周囲の音の聞き分けが不可欠である．実環境における人・ロボットインタラクションでは，ロボットに埋め込まれたマイクを用いることから，音源からの距離が遠く，信号対雑音比は接話マイクを使う場合よりも低い．それゆえ，音源定位や音源分離などのマイクアレイ処理はロボット聴覚において重要な役割を果たし，様々な応用がなされている [2; 3; 4; 5].

一般的にマイクアレイ処理は音源とマイク間の伝達関数を事前情報として使用するが，ロボット聴覚における応用では主に二つの方法で得られたものを使用している．一つは，マイクの位置から伝搬波モデルに基づいて幾何計算して求めた伝達関数である (幾何計算法) [2; 3]．幾何計算法はマイクが自由空間上に存在することが仮定されているが，ロボットに搭載されたマイクアレイを用いる場合は，ロボット形状に起因する反射や回折が誤差を生じるため，処理の性能を劣化してしまう．ロボットの形状を考慮した計算手法 [6; 7] も存在するが，計算コストが大きく，環境要因の音響特性 (壁からの反射等) を含めた計算ができない制約がある．もう一つは，*Time Stretched Pulse (TSP)* 信号や，M 系列信号を各方向に対して計測する方法である [8; 9]．この手法は上述の音響特性まで計測できるため，性能を確保できる [4; 5] ものの，計測に時間と手間がかかってしまう上，ロボットが移動すると音環境が計測した時と変化するため，その変化が誤差を生じ，性能が劣化してしまう．それゆえ，音環境が変化した時にユーザーが時間と手間をかけることなく伝達関数校正ができることが望ましい．

そこで，本稿ではマイクアレイ処理のロボット応用を考慮した，マイクアレイ伝達関数のオンライン校正について述べる．本稿は，特別な機器やマイク位置などの事前情報を必要とせず，十分に短いデータのみによって，マイクアレイをその場で高精度に校正する (オンザスポット校正) ことを目的とする．これまでにも，マイクアレイ周辺で移動する人の拍手音などの短い録音データを用いて，マイクアレイのオンザスポット校正が提案されてきた[10; 11; 12; 13; 14]．しかし，これらの手法は伝達関数ではなくマイク位置の校正となっており，伝達関数は伝搬波モデルを用いて幾何計算していた．この場合，上述のようにロボットや部屋の音響特性を考慮することができないため，その誤差がマイクアレイ処理性能を劣化してしまう．

そこで，本稿では，ロボットや部屋の音響特性を含めた伝達関数のオンザスポット校正を提案する．提案手法は *Frequency-domain Ordinary Least Squares (FOLS 法)* と *Frequency- and Time-Domain Linear Interpolation (FTDLI 法)*[15] から構成される．FOLS 法はマイクアレイ周囲を移動する人の拍手音から直接，伝達関数を推定する．既存手法とは違い，マイクアレイ位置を事前情報とせず，ロボットや部屋の音響特性を直接推定するため，ロボット応用に向いている．FOLS 法で得られる伝達関数は拍手音が観測された方向のみ得られるため，音源定位で使用されるような等間隔 (5◦ 毎など) に並んでいない．そこで，FTDLI 法を用いて伝達関数を補間することで，伝達関数を所望の解像度で整列させる．伝達関数を補間するため

Table 1: SLAM 問題とマイクアレイ校正問題の対応関係

SLAM(ロボット)	マイクアレイ校正
自己位置	音源位置
地図 (ランドマーク位置)	マイク位置
予測誤差最小	同期時刻ずれ推定付予測誤差最小

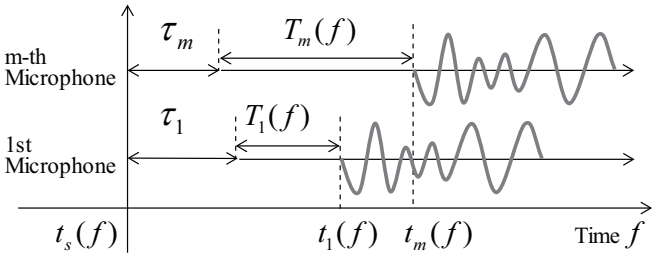


Figure 1: f 回目の拍手音到達時間

には，FOLS 法で得られた伝達関数の方向情報が必要となるため，*Simultaneous Localization And Mapping (SLAM)* に基づくオンラインマイク位置推定 (2 章参照 [13]) を導入し，幾何的に計算された伝達関数で拍手音を定位することで，方向情報を得る．最後に，提案法で推定された伝達関数を音源定位・音源分離に適用し有効性を確認する．

2 SLAM に基づくマイク位置推定

SLAM は，オンラインでロボットの自己位置と周囲の地図推定を同時に行う問題であり，ロボット分野で盛んに研究が行われている [15]．SLAM を解くため，様々なアルゴリズムが提案されているが，本稿では，*Extended Kalman Filter (EKF)* ベースの SLAM (EKF-SLAM) をマイクアレイ校正処理に適用する手法を導入する．この手法では，SLAM における地図推定を各マイクの位置推定，ロボットの自己位置推定を音源の位置推定に当てはめ，同期時刻のズレを含む推定誤差を最小になるよう更新を行うことによって校正処理を実現する (表 1 参照)．これによって，例えば，人が拍手をしながら，マイクアレイを周回することによって，マイクアレイのオンライン逐次校正を行うことができる．具体的には，M 個のマイクからなる非同期分散マイクアレイを考え，f 回目の拍手に対して，m 番目のマイクの状態ベクトル $\xi_m(f)$ (ただし，$1 \le m \le M$) と音源の状態ベクトル $\xi_s(f)$ を定義する．$\xi_m(f)$ は，2 次元のマイク位置情報と同期時刻ズレを含み，$\xi_s(f)$ は，2 次元の音源位置情報と進行方向を含む 3 次元ベクトルとして，$\xi_m(f) = [x_m(f), y_m(f), \tau_m(f)]^T$，$\xi_s(f) = [x_s(f), y_s(f), \theta_s(f)]^T$ と定義する．

2.1 観測モデル

m 番目のマイクで f 番目の拍手の到達時刻 $t_m(f)$ を観測する．m 番目のマイクの位置，および時刻ズレを (x_m, y_m)，τ_m，音速を c とすると，$t_m(f)$ は，図 1 に示されているように，音源が音を発した時刻 $t_s(f)$ を用いて，以下のように求めることができる．

$$t_m(f) = t_s(f) + T_m(f) + \tau_m \tag{1}$$

$$T_m(f) = \frac{\sqrt{(x_s(f) - x_m)^2 + (y_s(f) - y_m)^2}}{c} \tag{2}$$

音を発した時刻 $t_s(f)$ は未知であるため，基準マイク (マイク 1) での観測時刻との差をとると，観測モデルは，以下のように相対時刻で表すことができる．

$$\eta(f) = \begin{bmatrix} T_2(f) - T_1(f) + \tau_2 - \tau_1 \\ \vdots \\ T_M(f) - T_1(f) + \tau_M - \tau_1 \end{bmatrix} + \delta(f)\,, \tag{3}$$

観測誤差 $\delta(f)$ は白色雑音に従うものとする．

2.2 状態遷移モデル

音源，つまり人は，$\theta_s(l)$ 方向に等速 $v_s(l)$ で移動するとする．ここで，l は，歩数を表すインデックスであり，拍手の回数を表すインデックス f とは異なるものとする．状態は l，つまり 1 歩進むごとに更新されるため，音源の状態遷移モデルは以下のように表すことができる．

$$\xi_s(l+1) = \xi_s(l) + \begin{bmatrix} \sin(\theta_s(l)) & 0 \\ \cos(\theta_s(l)) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_s(l) \\ \dot{\theta}_s(l) \end{bmatrix} \Delta t + \mathscr{W}_s(l)\,, \tag{4}$$

$\mathscr{W}_s(l)$ はモデル誤差を表す白色雑音である．一方，マイクの位置は固定であるため，モデル誤差を $\mathscr{W}_m(l)$ とすれば，状態遷移モデルは以下のように表される．

$$\xi_m(l+1) = \xi_m(l) + \mathscr{W}_m(l)\,. \tag{5}$$

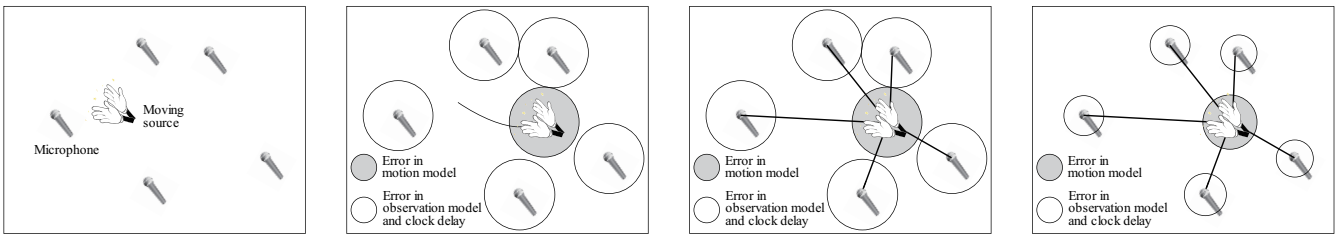
2.3 マイク位置推定

EKF-SLAM では，図 2(a) のように状態予測，観測予測，観測更新の 3 ステップを繰り返すことで校正を行う．まず，状態予測ステップでは，l 歩目での状態ベクトル $\xi_s(l)$，$\xi_m(l)$ を式 (4), (5) を用いて更新する (図 2(b))．観測予測ステップでは，状態予測ステップで更新された状態ベクトルと式 (3) を用いて f 回目の拍手の観測予測値 $\eta(f)$ を算出する (図 2(c))．観測更新ステップでは，f 回目の拍手で得られる実際の観測値と，観測予測値 $\eta(f)$ との誤差を最小にするようにカルマンゲインを導出し，状態ベクトルを更新する (図 2(d))．

3 伝達関数推定

伝達関数は室内における音源からマイクまでの音伝搬のモデルである．$S(\omega, f)$ と $X_m(\omega, f)$ をそれぞれ，短時間フーリエ変換後の f フレーム目の音源信号と m 番目のマイクでの観測信号とする．また，$A_m(\omega, \psi)$ を ψ 方向にある音源と m 番目のマイク間の周波数 ω での伝達関数とすると，ψ 方向にある音源 $S(\omega, f)$ は以下で表される．

$$X_m(\omega, f) = A_m(\omega, \psi) S(\omega, f) \tag{6}$$



(a) 初期状態: 基準マイクを原点とする．マイク位置は未知.

(b) 状態予測: 状態遷移モデルを用いて状態を更新する．

(c) 観測予測: 観測モデルを用いて観測を予測する．

(d) 観測更新: 予測誤差が最小となるようにカルマンゲインを更新する．

Figure 2: EKF-SLAM に基づく状態推定

ただし，フレーム長は十分長いとする．

伝達関数推定では $A_m(\omega,\psi)$ を推定することを目的とする．1 章で述べたように，伝達関数は数値計算もしくは計測によって求めるのが主流である．既存の伝達関数計測手法 [8; 9] では，$S(\omega,f)$ は既知である (TSP など) ことが前提であり，$A_m(\omega,\psi)$ は以下で得られた，

$$A_m(\omega,\psi) = X_m(\omega,f)/S(\omega,f)\,. \tag{7}$$

しかし，本稿では，拍手音から伝達関数を推定したいので，音源信号 $S(\omega)$，音源方向 ψ ともに未知として，伝達関数推定問題を解く必要がある．

3.1 FOLS 法による伝達関数推定

まず，拍手が行われた位置での音源とマイク間の伝達関数を推定する．拍手音の元信号 $S(\omega)$ は未知であり，事前情報として用いることができない．音源定位・分離では，あるチャンネルを基準にした相対的な伝達関数がわかれば処理上問題がないので，マイク 1 の観測信号 $X_1(\omega,f)$ を基準として，相対伝達関数を求める．すなわち，伝達関数は式 (7) の代わりに以下となる．

$$A_m(\omega,\psi) = X_m(\omega,f)/X_1(\omega,f) \tag{8}$$

しかし，$A_m(\omega,\psi)$ は，m 番目のマイクが基準マイクより拍手音を早く観測する場合，非因果成分を持ってしまう．そこで，$A_m(\omega,\psi)$ が非因果成分を持たないように．基準マイクの信号を T_o サンプルずらしたものを使う ($\tilde{X}_1(\omega,f)$ を表すこととする)．FOLS 法は，F フレームの $\tilde{X}_1(\omega,f)$ と $X_m(\omega,f)$ を用いて，回帰モデルを用いて雑音ロバストに伝達関数を推定する手法であり，以下で表される．

$$\underbrace{\begin{bmatrix} X_1(f+1) & \dots & X_M(f+1) \\ \vdots & & \vdots \\ X_1(f+F) & \dots & X_M(f+F) \end{bmatrix}}_{X_{[1:F]}} = \underbrace{\begin{bmatrix} \tilde{X}_1(f+1) \\ \vdots \\ \tilde{X}_1(f+F) \end{bmatrix}}_{\Omega_{[1:F]}} \underbrace{\begin{bmatrix} A_1(\psi) \\ \vdots \\ A_M(\psi) \end{bmatrix}^T}_{A^T(\omega,\psi)}$$

ここで，$\Omega_{[1:F]}$ はリグレッサである．最後に，$A(\omega,\psi)$ は，以下で求められる．

$$A^T(\omega,\psi) = (\Omega^T_{[1:F]}\Omega_{[1:F]})^{-1}\Omega^T_{[1:F]}X_{[1:F]} \tag{9}$$

フレーム数 F を長く取れば雑音ロバスト性能を高くすることが可能である．FOLS 法によって推定された伝達関数はマイク位置から得られる直接音成分だけでなく，ロボットや部屋の音響特性も含めたものであるため，実環境下のロボット聴覚応用に向いていると言える．

3.2 補間による伝達関数の整列

次に，音源方向 ψ について考える．拍手を行った際の音源の方向は，2 章の EKF-SLAM によって得ることができる．しかし，実際には，人の移動は状態遷移モデルに完全には従っていないため，EKF-SLAM から得られる音源方向の誤差は大きい．一方，マイク位置は，そもそも移動しないため，モデル誤差が小さく，精度のよい結果が得られる．そこで，各拍手の ψ を，より精度よく推定するため，推定したマイク位置を用いて，伝搬波モデルを用いて伝達関数を計算し，各拍手音 $S(\omega,f)$ のビームフォーミングを用いた定位を行い，精度のよい拍手方向 ψ を得る．

得られる音源方向 ψ のセットは拍手の位置であるため，伝達関数が等間隔に並んでおらず，音源定位や分離で使い勝手が悪い．そこで，伝達関数の補間を行い，所望の間隔 (5° 毎など) に配置された伝達関数を得る．

具体的には，K を総拍手数，ψ_k を k 回目の拍手の音源方向とする．また，得たい伝達関数を水平各方向に一周を N 等分した ψ_n ($1 \le n \le N$) とする．各 ψ_n に対して，ψ^- と ψ^+ を ψ_k の中で ψ_n に最も近い近傍の 2 点とする ($\psi^- \le \psi_n < \psi^+$)．FTDLI 法 [15] を用いて，$\psi^-$ と ψ^+ における伝達関数 $A(\omega,\psi^-)$ および $A(\omega,\psi^+)$ から $\hat{\psi}_n$ における伝達関数 $A(\omega,\psi)$ を補間する。

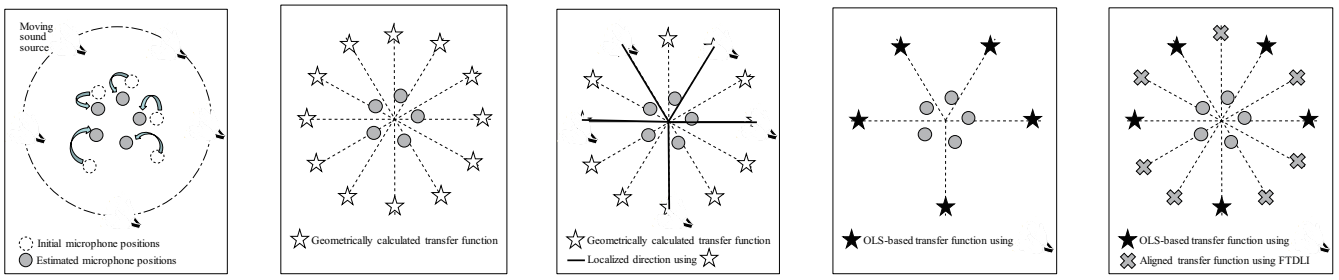
1) 周波数領域上と時間領域上で ψ_n が $[\psi^-\psi^+]$ の内分点となる α を算出し，線形補間を以下のように行う:

$$A_{m[F]}(\omega,\psi) = \alpha A_m(\omega,\psi^-) + (1-\alpha)A_m(\omega,\psi^+)$$
$$A_{m[T]}(\omega,\psi) = A^{\alpha}_m(\omega,\psi^-)A^{1-\alpha}_m(\omega,\psi^+),$$

ここで，$\psi^- \le \psi \le \psi^+$, $\alpha = \frac{\psi^+-\psi}{\psi^+-\psi^-}$ である．

2) 得られた伝達関数を振幅情報と位相情報に分離する:

$$A_{m[F]}(\omega,\psi) = \lambda_{m[F]}\exp(-j\omega t_{m[F]})$$
$$A_{m[T]}(\omega,\psi) = \lambda_{m[T]}\exp(-j\omega t_{m[T]})$$



(a) EKF-SLAM によるマイク位置推定　(b) 幾何計算法による伝達関数計算　(c) ビームフォーミングによる拍手音の定位　(d) 拍手音を用いた FOLS 法による伝達関数推定　(e) FTDLI 法による伝達関数の整列

Figure 3: 伝達関数のオンザスポット校正

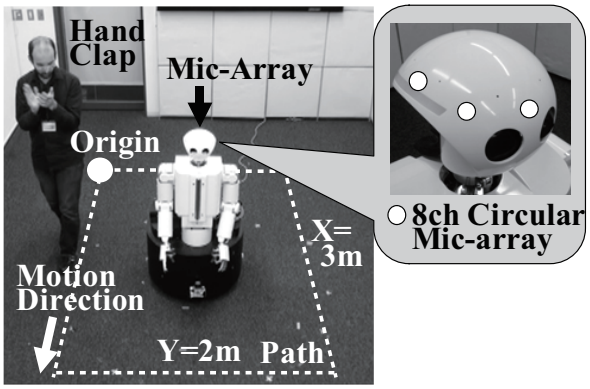


Figure 4: 実験環境

3) 振幅情報は時間領域，位相情報は周波数領域の補間を採用して最終的な伝達関数補間を行う:

$$A_m(\omega,\psi) = \lambda_{m[T]} \exp(-j\omega t_{m[F]})$$

この手法は，周波数領域の線形補間[16] と時間領域の線形補間[17] のハイブリッド法となっており，振幅と位相の両方が正しく補間できる．

3.3 システム構成

伝達関数推定の一連の流れを，図 3 に示す．EKF-SLAM によって得られるマイク位置 (図 3(a)) を用いて幾何計算法により，伝達関数を計算する (図 3(b))．得られた伝達関数を用いて，ビームフォーミングを行い，EKF-SLAM の際に観測した拍手の位置を推定する (図 3(c))．FOLS 法を用いて，拍手位置に対する伝達関数を推定する (図 3(d))．最後に FTDLI 法を用いて，伝達関数の補間を行い，等間隔に並んだ所望の方向の伝達関数を算出する (図 3(e))．

4 評価

本章では，まず，2 章のマイク位置推定の結果を示し，どのくらい短い計測でマイク位置が校正できるかを議論する．次に，3 章で推定された伝達関数をロボットを用いた音源定位・音源分離に適用し，既存手法で得られた伝達関数を用いた場合と比較を行う．

評価では提案手法をオープンソースのロボット聴覚ソフトウェア HARK [18] 上に実装し，2.0GHz の Intel Core i7 の CPU を持つ計算機で実時間動作することを確認した．本稿では，マイクアレイを搭載したロボットを残響時間が 0.2 秒 (RT20) の 7.0 m × 4.0 m の部屋の中央に設置した．マイクアレイは図 4 のように 8 チャネルの円状アレイを用いた．入力音響信号は 16kHz, 16 ビットでサンプリングした．音響信号処理のフレーム長とシフト長はそれぞれ，512, 160 サンプルとした．

4.1 マイク位置推定の評価

上述のように，ロボットは 7.0 m × 4.0 m の部屋の中央に設置されている．ロボットに搭載された円上アレイは半径約 0.1 m であった．人は，図 4 で示された原点からスタートし，3.0 m × 2.0 m の長方形の点線に沿って反時計回りに一定速度 $v_s(l)$ で移動した．拍手は 1 回/秒とした．

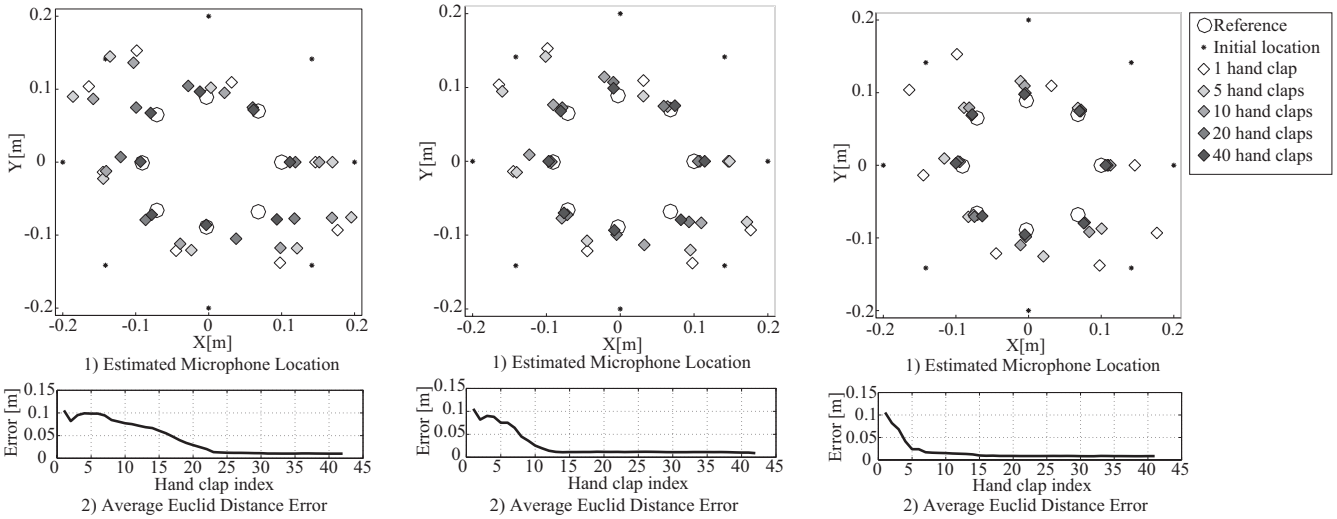
EKF-SLAM のため，人の初期位置と $v_s(l)$ は正解データを与えた．式 (3) の白色雑音の標準偏差 $\delta(f)$ は 0.0005 とした．式 (4) の白色雑音の標準偏差 $\mathscr{W}_s(l)$ は，$x_s(f), y_s(f)$ に対して 0.25，$\theta_s(f)$ に対して 1.0 とした．式 (5) の白色雑音の標準偏差 $\mathscr{W}_m(l)$ は，$x_m(f), y_m(f)$ に対して 0.25，$\tau_m(f)$ に対して 1.0 とした．収束速度を評価するため，マイクの初期位置はランダムに与えず，半径 0.2 m の円上アレイとなるように設定した．

図 5 に，各拍手回数での推定マイク位置，および全マイクの推定位置のユークリッド距離平均誤差の変化を示す．$v_s(l)$ を 0.1 m/s, 0.2 m/s, 0.6 m/s と変化させて評価した．

まず，図 5(a)-2), 5(b)-2), 5(c)-2) を比較すると，$v_s(l)$ が速いほど，収束速度が速いことがわかる．人はまず x 方向に 3 m 移動し，一辺を終えるのに $\frac{3}{v_s(l)}$ 回の拍手を必要とする (例えば，$v_s(l) = 0.1$ m/s の場合は 30 回)．収束速度はその一辺を終えるまでにかかる時間と相関があることから，観測時間差に大きな分散があるように移動すれば収束が速いことがわかる．

また，図 5(a)-1), 5(b)-1), 5(c)-1) に共通して，マイクの x 方向の位置から校正されている傾向が見られる．これは人が x 方向に最初に移動するためだと考えられる．このことからマイクをより早く校正するための最適な移動方法があると考えられる．

いずれの場合もマイク位置は高精度に校正されている



(a) $v_s(l) = 0.1$ m/s (b) $v_s(l) = 0.2$ m/s (c) $v_s(l) = 0.6$ m/s

Figure 5: マイク位置推定の結果

ことから提案法の有効性を確認することができた．また，$v_s(l) \geq 0.2$ m/s の場合は 20 回の拍手でマイク位置が十分に校正できていることから，20 秒ほどの録音でマイクを校正することができることがわかり，オンザスポット校正が十分可能であることが示された．4.2 章の評価では 20 回の拍手で推定されたマイク位置を用いることとした．

4.2 伝達関数推定の評価

伝達関数推定の有効性を音源定位・音源分離を通じて評価する．提案する伝達関数推定法に加え，2 種類の手法を比較した。TSP 法 (**TSP**) は，TSP 信号を用いて実際に測定した伝達関数を用いる手法であり，最も精度が良いことが期待できる [8]．幾何計算法 (**Calc**) は，マイク位置は EKF-SLAM で得られた位置を用い，自由音場を仮定した幾何計算によって算出した伝達関数を用いた手法である．

4.2.1 音源定位の性能比較

音源定位には，適応ビームフォーマの一種である，*MUltiple SIgnal Classification (MUSIC)*[15]を用いた．MUSIC では，M チャネルの入力音響信号の空間相関行列を計算し，その固有値展開を行う．$E(\omega, f) = [e_1(\omega, f), \ldots, e_M(\omega, f)]$ を f フレーム目で得られた固有ベクトルとする．定位では以下で表される MUSIC スペクトルを算出し，ここで提案法で推定された伝達関数 $A(\omega, \psi)$ を用いた．

$$P(\omega, \psi, f) = \frac{|A^*(\omega, \psi)A(\omega, \psi)|}{\sum_{m=L+1}^{M} |A^*(\omega, \psi)e_m(\omega, f)|}, \qquad (10)$$

ここで，$()^*$ は共役転置作用素を，L は音源数を表す．

音源定位では，図 6a) に示すように，ロボットの周囲で 10 cm ごとに 100 箇所でスピーカから 10 秒間ずつ白色雑音を出力し，水平方向の定位を行った．なお，TSP 法は，100 箇所すべてにおいて測定した伝達関数を用いた．提案法，幾何計算法では，20 回の拍手を用いてマイク位置を推定し，伝達関数は，5° ごとに推定・算出した．得られた伝達関数を式 (10) に適用し，評価を行った．

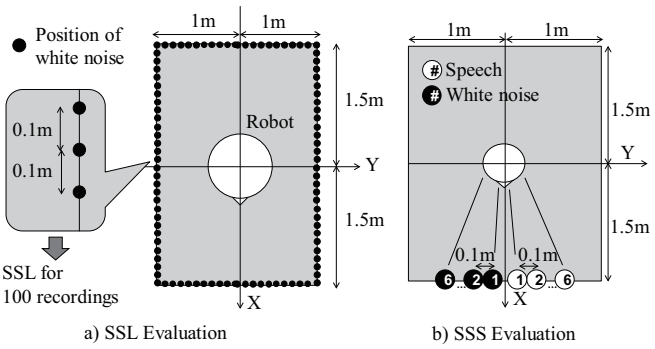


Figure 6: a) 音源定位評価のための白色雑音録音位置, b) 音源分離評価のための白色雑音・音声録音位置

Table 2: 音源定位評価結果

	TSP	Calc	Proposed
平均誤差 [deg]	5.11±2.01	6.93±2.04	6.82±2.13

表 2 に，音源定位の水平角度推定誤差の平均と標準偏差を示す．TSP 法が最も良い結果を示した．提案法 (Proposed) が，Calc に対して定位誤差が改善していることが確認でき，伝達関数にロボットや部屋の音響特性を考慮できたことの有効性が示された．

4.2.2 音源分離の性能比較

音源分離には，幾何拘束とブラインド分離のハイブリッドアルゴリズムである *Geometric High-order Decorrelation-based Source Separation (GHDSS)* [19] を用いた．GHDSS では，パーミュテーションとスケーリング問題を解決するために幾何拘束を用いており，伝達関数が使われている．音源分離は以下で表される．

$$Y(\omega, \psi) = W(\omega, \psi)X(\omega, \psi), \qquad (11)$$

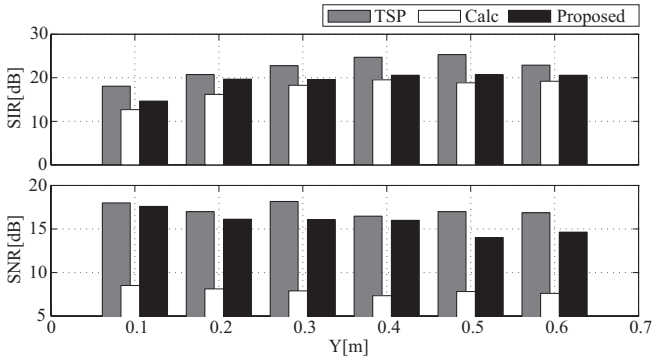


Figure 7: 音源分離評価結果

ここで，$Y(\omega,\psi)$ は分離音，$W(\omega,\psi)$ は分離行列，$X(\omega,\psi)$ は M チャネルの入力音響信号を表す．コスト関数 $J(W(\omega,f))$ を最小化するように $W(\omega,\psi)$ を更新することで音源分離が行われるが，GHDSS ではコスト関数を以下のように設定している．

$$J(W(\omega,f)) = \alpha J_1(W(\omega,f)) + \beta J_2(W(\omega,f)), \qquad (12)$$

ここで，$J_1(\cdot)$ はブラインド音源分離のためのコスト関数を，$J_2(\cdot)$ は幾何拘束のためのコスト関数を表す．α と β は，$\alpha+\beta=1$ を満たす重みを表す．提案法で推定された伝達関数 $A(\omega,\psi)$ は，以下のように $J_2(\cdot)$ で用いられる．

$$J_2(W(\omega,f)) = \|diag[W(\omega,f)A-I]\|^2. \qquad (13)$$

評価では，*Signal-to-Interference Ratio (SIR)* と *Signal-to-Noise Ratio (SNR)* の 2 つの指標を用いた．SIR の評価には，BSS EVAL Toolbox [20] を用いた．Toolbox では，分離音は $y_i(t) = s_t(t) + s_i(t) + s_n(t)$ としてモデル化されている．ここで，$s_t(t)$ は目的音のみがある場合の分離信号を，$s_i(t)$ は非目的音のみがある場合の分離信号を，$s_n(t)$ 背景雑音のみがある場合の分離信号を表す．SIR は $SIR = 10\log_{10}\frac{|s_t|^2}{|s_i|^2}$ として，SNR は $SNR = 10\log_{10}\frac{|s_t+s_i|^2}{|s_n|^2}$ として計算される．

図 6(b) に示すように，ターゲット音源の位置を白丸の 1〜6 から選択，雑音源 (白色雑音) は，ターゲット音源と対称になる位置の黒丸を選択した。各手法によって得られた伝達関数を GHDSS の式 (13) の D に用い，２音源からの混合音に対し音源分離を行った．

図 7 に結果を示す. 音源分離でも TSP 法が一番良い結果を示している。提案法 (Proposed) は，SIR, SNR 共に TSP 法に近い性能を示しており，Calc と比較すると良好な結果が得られていることがわかる。いずれも伝達関数にロボットや部屋の音響特性を考慮したことが効果を示した．

TSP 法は音源定位・音源分離の両者でもっとも良い性能を示したが，100 箇所の伝達関数の計測には特別な機器を用いても 60 分を要している．一方，提案法は 20 回の拍手 (20 秒) のみで伝達関数を校正できており，結果として TSP 法と同様な性能を得られた．このことから，提案法は初心者にも簡易に実現可能な実用的なマイクアレイ校正手法であると考えている．

5 結論

本稿ではロボットや部屋の音響特性を考慮したマイクアレイ伝達関数のオンザスポット (その場で簡易に可能な) 校正について述べた．EKF-SLAM に基づくマイク位置推定を導入し，FOLS 法と FTDLI 法による伝達関数推定を提案した．評価では，20 回の拍手 (20 秒の録音) でマイクアレイが精度良く校正でき，音源定位・音源分離を通して提案法で得られた伝達関数がマイクアレイ処理性能を向上していることが確認できた．今後の課題は音声などの一般音を用いた校正や三次元モデルへの拡張である．

参考文献

[1] K. Nakadai *et al.*, “Active Audition for Humanoid”, in *Proc. of 17th AAAI*, pp. 832–839, 2000.

[2] Y. Sasaki *et al.*, “Nested igmm recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition,” in *IROS*, pp. 3930–3936, 2013.

[3] J.-M. Valin *et al.*, “Enhanced robot audition based on microphone array source separation with post-filter,” in *IROS*, pp. 2123–2128, 2004.

[4] K. Nakamura *et al.*, “Intell. sound source localization for dynamic environments,” in *IROS*, pp. 664–669, 2009.

[5] F. Asano *et al.*, “Sound source localization and signal separation for office robot Jijo-2,” in *Proc. of IEEE Int ’l Conf. Multisensor Fusion and Integ. for Intell. Sys. (MFI)*, pp. 243–248, 1999.

[6] K. Yamamoto *et al.*, “An acoustic simulation for speech interface of humanoid robot,” in *Proc. of Acoustical Society of Japan Autumn Meeting*, pp. 815–818, 2009.

[7] K. Nakadai *et al.*, “Applying scattering theory to robot audition system: robust sound source localization and extraction,” in *IROS*, pp. 1147–1152, 2003.

[8] Y. Suzuki *et al.*, “An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses”, *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119-1123, 1995.

[9] G. B. Stan, J. J. Embrechts and D. Archambeau, “Comparison of different impulse response measurement technique,” *Journal of the Audio Engineering Society*, vol. 50, pp. 249–262, 2002.

[10] S. Thrun, “Affine structure from sound,” *Advances in Neural Information Processing Systems*, vol. 18, pp. 1355–1362, 2005.

[11] Y. Kuang and K. Astrom, “Stratified Sensor Network Self-Calibration From TDOA Measurements,” in *EUSIPCO*, 2013.

[12] N. Ono *et al.*, “Blind Alignment of Asynchronous Recorded Signals for Distributed Microphone Array,” in *WASPAA*, pp. 161–164, 2009.

[13] H. Miura *et al.*, “SLAM-based Online Calibration of Asynchronous Microphone Array for Robot Audition,” in *IROS*, pp. 524–529, 2011.

[14] Y. Bando *et al.*, “Posture estimation of hose-shaped robot using microphone array localization,” in *IROS*, pp. 3446–3451, 2013.

[15] K. Nakamura, K. Nakadai and G. Ince, “Real-time Super-resolution Sound Source Localization for Robots,” in *IROS*, pp. 694–699, 2012.

[16] T. Nishino *et al.*, “Interpolating head related transfer functions in the median plane,” in *WASPPA*, pp. 167–170, 1999.

[17] M. Matsumoto *et al.*, “A method of interpolating binaural impulse responses for moving sound images,” *Acoust. Sci. Tech.*, Vol 24, pp. 284–292, 2003.

[18] K. Nakadai *et al.*, “Design and Implementation of Robot Audition System HARK”, *Advanced Robotics*, vol. 24, pp. 739–761, 2009.

[19] H. Nakajima *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.

[20] E. Vincent, R. Gribonval and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

マイクロホンアレイとスピーカをもつ柔軟索状ロボットのための動的スピーカ選択による姿勢推定の高速化

坂東宜昭 [1] Yoshiaki Bando
糸山克寿 [1] Katsutoshi Itoyama
昆陽雅司 [2] Masashi Konyo
田所諭 [2] Satoshi Tadokoro
中臺一博 [3] Kazuhiro Nakadai
吉井和佳 [1] Kazuyoshi Yoshii
奥乃博 [4] Hiroshi G. Okuno

1 京都大学 大学院情報学研究科 2 東北大学 大学院情報科学研究科
3 東京工業大学 大学院情報理工学研究科, (株) ホンダ・リサーチ・インスティチュート・ジャパン
4 早稲田大学 理工学術院

Abstract

レスキューロボットの一つである柔軟索状ロボットは細長い形状を生かし，災害現場で人の進入が難しい狭い空間へ進入し探索できるが，柔軟な本体の制御，姿勢推定が難しいという課題がある．本稿では柔軟索状ロボットにマイクロホンアレイと小型スピーカを装着し，音の到達時間差を利用した姿勢推定を行う．従来，ロボット上のスピーカの再生順は端から順番に鳴らしていたが，推定姿勢の収束に時間を要することや精度が低下する問題があった．そこで，次に鳴らすべきスピーカを強化学習によりオンラインで決定する問題に取り組み，エントロピー最小化基準による動的スピーカ選択法を Unscented 変換による近似を用いて開発した．ロボットの姿勢から幾何的に観測を生成する数値実験による評価を行い，従来の順番にスピーカを鳴らす場合に比べて最大 67%の収束高速化と 50%の精度向上化がなされることを確認した．

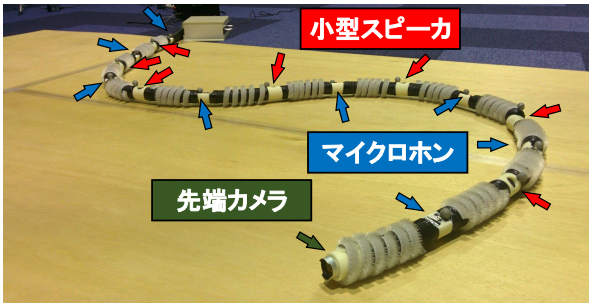


Figure 1: マイクロホンアレイとスピーカをもつプロトタイプ・柔軟索状ロボット

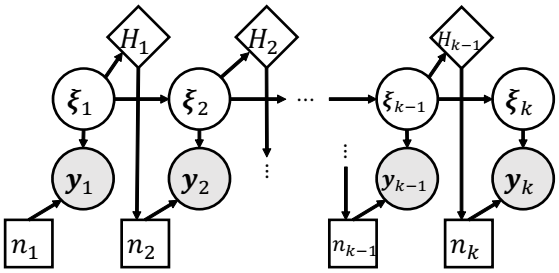


Figure 2: 動的スピーカ選択のためのダイナミック決定ネットワーク

1 はじめに

柔軟索状ロボット[Kitagawa et al., 2003; Hatazaki et al., 2007; Namari et al., 2012]はレスキューロボットの一種で，細く長い筐体を持ち，瓦礫内の探索といった人や動物が侵入できない環境の探索[Ohno et al., 2011; Nagatani et al., 2011; Voyles et al., 2012; Baiocchi et al., 2013]のために開発されている．リモートオペレータはロボットに搭載された駆動機構を用いて柔軟索状ロボットを探索対象へ移動させることができる．例えば，Active Hose-II [Kitagawa et al., 2003]は小型の車輪を用いて，Active Scope Camera (ASC) [Namari et al., 2012]は表面に接着された繊毛を振動させることで瓦礫内を進む．また，ASC はアメリカでの実際の災害現場での適用例も報告されている [Tadokoro et al., 2009]．

柔軟に形状変化する本ロボットの姿勢制御では，姿勢推定が不可欠である．既存の内界センサによる姿勢推定法 [Ishikura et al., 2012]は加速度センサとジャイロセンサの値を積分して姿勢を推定する積分型計測法である．このような姿勢の変化率から現在の姿勢を推定する手法は，長時間の運用では誤差が蓄積するという問題がある．また，GPS や曲げセンサといった，過去の姿勢に依存しない従来の非積分型計測法では，屋内や長い筐体で精度が低下するという問題があった．

これまで我々は非積分型計測法として，音を用いた柔軟索状ロボットの姿勢推定法を開発してきた [Bando et al., 2013]．柔軟索状ロボットにマイクロホンと小型スピーカを装着し (Figure 1)，小型スピーカから発する試験音の各マイクロホンへの到達時間差を用いてマイクロホンと小型スピーカの位置関係を推定する．本手法で使用する到達時間差は，現在のマイクロホンと小型スピーカの位置関係にのみ依存するため，累積誤差の問題を回避できる．また，ロボット上のマイクロホンアレイは，音源定位や分離といった声による被災者発見への応用が期待できる．遠隔地の音源方向提示による聴覚アウェアネスの有効性は，ロボット聴覚ソフトウェア HARK [Nakadai et al., 2010]を用いたテレプレゼンスロボットの開発 [Mizumoto et al., 2011]でも指摘されている。

従来はロボット上のスピーカは端から順番に繰り返して鳴らして姿勢推定を行っていた．しかし，ロボットの姿勢やマイクロホンの個数によって，推定姿勢の収束に時間を要することや精度が低下する問題があった．そこで本研究では，次に鳴らすべきスピーカを強化学習によりオンラインで決定する手法を開発し，姿勢推定の収束高速化と安定化を行う．

オンライン強化学習による行動最適化手法として，アクティブ･ビジョンによる位置推定法が提案されている [Czarnetzki et al., 2011]．この手法は，ロボカップでのロボットの位置推定のための手法で，ロボット位置の事後分布のエントロピーが最も小さくなる行動を選択する．本手法はエントロピーと事後分布の計算に Particle Filter [Arulampalam et al., 2002]によるモンテカルロ積分を使用しており，柔軟索状ロボットの姿勢のような高次元状態空間への適応は，計算量の観点から困難である[Ishikura et al., 2012]．提案法では，事後分布とエントロピーの計算を Uncsented 変換 [Julier, 2002]により計算量を削減してスピーカ選択へ応用する．

本稿の構成は以下のとおりである．第 2 章では音を用いた柔軟索状ロボットの姿勢推定法について述べる．第 3 章では，ロボットの姿勢から幾何的に計算した到達時間差を用いる数値実験による評価を行い，提案法により収束速度と精度が向上することを確認する．第 4 章でまとめる．

2 柔軟索状ロボットのマイクロホンアレイを用いた姿勢推定

Figure 3 に示すように，本稿では柔軟索状ロボットに，マイクロホンと小型スピーカを交互に間隔 l だけ離して装着する．各マイクロホンと小型スピーカはそれぞれ手元から順に mic_1, src_1, $\cdots$, src_N, mic_M とする．ここで，M, N はそれぞれマイクロホンと小型スピーカの個数を表し，$N = M - 1$ である．$\mathrm{mic}_m, \mathrm{src}_n$ の各座標を，それぞれ $\boldsymbol{u}_{m,k}, \boldsymbol{v}_{n,k} \in \mathbb{R}^2$ とする．k は観測のインデックスである．本稿で用いる表記を Table 1 にまとめた．

Table 1: 各記号の意味

記号	意味
M	マイクロホンの個数
N	小型スピーカの個数 ($N = M - 1$)
C	音速
k	観測のインデックス
mic_m	m 番目のマイクロホン ($1 \le m \le M$)
src_n	n 番目のスピーカ ($1 \le n \le N$)
$\boldsymbol{u}_m$	mic_m の座標 $\in \mathbb{R}^2$
$\boldsymbol{v}_n$	src_n の座標 $\in \mathbb{R}^2$
$\boldsymbol{\xi}_k$	ロボットの姿勢 $\in \mathbb{R}^{M+N-2}$
$\boldsymbol{y}_k$	k 番目の観測 $\in \mathbb{R}^{M-1}$
$\tau^n_{m_1,m_2}n$	スピーカ src_n 再生時のマイクロホン mic_{m_1}, mic_{m_2} 間の到達時間差 $\in \mathbb{R}$

本稿で扱う姿勢推定は次の手続きを繰り返し逐次実行する (Algorithm 1)．まず 1) 小型スピーカ src_{n_k} から試験音を再生し，2) 試験音の録音から試験音の各マイクロホンへの到達時間差 $\tau^{n_k}_{m_1 \to m_2,k}$ を推定する．3) 得られた到達時間差から姿勢を表すマイクロホンと小型スピーカの座標 $\boldsymbol{u}_{m,k}, \boldsymbol{v}_{n,k}$ を推定する．そして 4) 次に再生すべきスピーカ $\mathrm{src}_{n_{k+1}}$ をエントロピー最小基準で選択する．

Algorithm 1 マイク位置の逐次推定

1: 最初に再生するスピーカ src_{n_1} を src_1 と設定
2: **for** $k \in 1, 2, 3, \cdots$ **do**
3: 　スピーカ src_{n_k} から試験音を再生
4: 　M チャネルマイクロホンアレイで録音
5: 　試験音の到達時間差から姿勢の事後分布を推定
6: 　エントロピー最小基準でスピーカ $\mathrm{src}_{n_{k+1}}$ を決定
7: **end for**

以下に本稿で扱う問題設定を示す．

入力: src_{n_k} で再生した試験音の各マイクロホン間での到達時間差 $\tau^{n_k}_{m_1 \to m_2,k}$
出力: ロボット上のマイクロホンと小型スピーカの座標 $\boldsymbol{u}_{m,k}, \boldsymbol{v}_{n,k}$，および次に鳴らすスピーカ番号 n_{k+1}

ただし，試験音とは到達時間差推定のために小型スピーカで再生する元信号である．到達時間差はマイクロホンと小型スピーカの相対的な位置関係を表すので，出力はロボット上のマイクロホンと小型スピーカの相対位置 $\boldsymbol{u}_{m,k}, \boldsymbol{v}_{n,k}$ である．また，一般性を失わずに $\boldsymbol{u}_{1,k}, \boldsymbol{v}_{1,k}$ を既知とする．

マイクロホンと音源の位置を同時推定する関連研究に，補助関数法による手法[Ono et al., 2009]と EKF-SLAM による手法[Miura et al., 2011]がある．前者はオフラインで

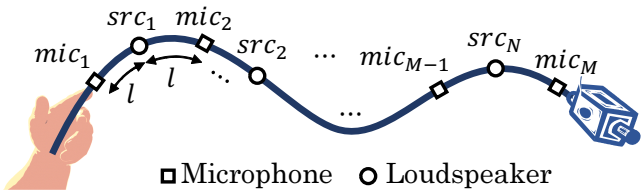


Figure 3: ロボット上のマイクロホンと小型スピーカの配置

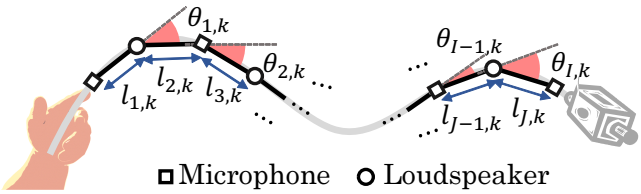


Figure 4: 姿勢のモデル

動作することを前提としておりロボットの姿勢推定には適さない．後者はオンライン手法だが運動モデルが既知の単一音源を仮定している．瓦礫中では音源の移動は困難であり，静止した 1 音源ではマイク位置推定できないため，本問題には適さない．ロボット上に配置された複数個の音源を仮定し，提案法は EKF-SLAM 法を改良したロボットの姿勢を表す状態空間モデルを用いたオンライン推定を実現する．

2.1 姿勢と観測のモデル

Figure 4 に示すように，柔軟索状ロボットの姿勢を表す状態変数 $\boldsymbol{\xi}_k$ は，隣り合うマイクロホンと小型スピーカ間を線分で結んだリンクモデルで近似する．よって姿勢 $\boldsymbol{\xi}_k$ は，各ノード間の角度 $\theta_{i,k}(1 \leq i \leq N+M-2)$ からなる $N+M-2$ 次元ベクトルで表現する．

$$\boldsymbol{\xi}_k = [\theta_{1,k}, \cdots, \theta_{N+M-2,k}]^{\mathrm{T}} \tag{1}$$

各マイクロホンと小型スピーカの座標 $\boldsymbol{u}_{m,k}, \boldsymbol{v}_{n,k}$ は $\boldsymbol{x}_{i,k}$ を $[\boldsymbol{u}_{1,k}, \boldsymbol{v}_{1,k}, \cdots, \boldsymbol{u}_{M-1,k}, \boldsymbol{v}_{N,k}, \boldsymbol{u}_{M,k}]$ の i 番目の要素として次のように計算される．

$$\boldsymbol{x}_{i,k} = \boldsymbol{x}_{i-1,k} + l_{i,k} \times [\cos(\sum_{i'=1}^{i-1} \theta_{i',k}), \sin(\sum_{i'=1}^{i-1} \theta_{i',k})] \tag{2}$$

観測モデル　観測モデル $p(\boldsymbol{y}_k|\boldsymbol{\xi}_k)$ は，小型スピーカ src_n が再生時のマイクロホン $\mathrm{mic}_{m_1}, \mathrm{mic}_{m_2}$ 間の到達時間差 $\tau^n_{m_1 \to m_2}(\boldsymbol{\xi}_k)$ を用いて次のように表現する．

$$p(\boldsymbol{y}_k|\boldsymbol{\xi}_k, n_k) = \mathcal{N}(\boldsymbol{y}_k|\boldsymbol{T}_{n_k}(\boldsymbol{\xi}_k), \boldsymbol{R}_k) \tag{3}$$

$$\boldsymbol{T}_{n_k}(\boldsymbol{\xi}_k) = [\tau^n_{n_k \to 1}(\boldsymbol{\xi}_k), \cdots, \tau^n_{n_k \to n-1}(\boldsymbol{\xi}_k), \tau^n_{n_k \to n+1}(\boldsymbol{\xi}_k), \cdots, \tau^n_{n_k \to M}(\boldsymbol{\xi}_k)]^{\mathrm{T}} \tag{4}$$

ただし $\boldsymbol{y}_k \in \mathbb{R}^{N-1}, \boldsymbol{R}_k \in \mathbb{R}^{N-1 \times N-1}$ はそれぞれ，観測ベクトルと観測誤差を表す共分散行列である．

到達時間差 $\tau^n_{m_1 \to m_2}(\boldsymbol{\xi}_k)$ はロボット上のマイクロホンと小型スピーカの座標から次のように定義する．

$$\tau^n_{m_1 \to m_2}(\boldsymbol{\xi}_k) = \frac{D^{n,m_2}(\boldsymbol{\xi}_k) - D^{n,m_1}(\boldsymbol{\xi}_k)}{C} \tag{5}$$

ここで $D^{n,m}(\boldsymbol{\xi}_k)$ は src_n と mic_m 間の距離を表し，C は音速を表す．

状態更新モデル　状態更新モデル $p(\boldsymbol{\xi}_k|\boldsymbol{\xi}_{k-1})$ は，ランダムウォークで表現する．

$$q(\boldsymbol{\xi}_k|\boldsymbol{\xi}_{k-1}) = \mathcal{N}(\boldsymbol{\xi}_k|\boldsymbol{\xi}_{k-1}, \boldsymbol{Q}_k) \tag{6}$$

ただし $\boldsymbol{Q}_k \in \mathbb{R}^{L \times L}$ はモデル誤差を表す共分散行列である．

2.2 推論アルゴリズム

本手法では，姿勢推定と次のスピーカの選択を Figure 2 に基づき，以下のステップで行う (Algorithm 2)．まず，Unscented Kalman Filter [Wan et al., 2000]を用いて現在の観測 $\boldsymbol{y}_k$ と事後分布 $p(\boldsymbol{\xi}_k|\boldsymbol{y}_{1:k}, n_{1:k})$ を推定する．次に，時刻 $k+1$ の姿勢の予測分布 $p(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k}, n_{1:k})$ を計算する．最後に，すべてのスピーカ n_{k+1} について，エントロピー $E[H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k}, n_{1:k+1})]$ を計算し，これが最小となるスピーカを次に鳴らすスピーカとして選択する．

Algorithm 2 マイク位置の逐次推定

1: 現在の観測から姿勢の事後分布 $p(\boldsymbol{\xi}_k|\boldsymbol{y}_{1:k}, n_{1:k})$ を計算
2: 次の時刻の姿勢の予測分布 $p(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k}, n_{1:k})$ を計算
3: **for** $n_{k+1} \in 1, \cdots, N$ **do**
4: 　　エントロピー $E[H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k}, n_{1:k+1})]$ を計算
5: **end for**
6: $E[H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k}, n_{1:k+1})]$ が最小となる n_{k+1} を選択

Unscented 変換　提案法ではすべての分布をガウス分布で近似し，Unscented 変換 [Julier, 2002]による計算量削減を行う．従来法では事後分布とエントロピーの計算に Particle Filter によるモンテカルロ積分を行っているが，柔軟索状ロボットの姿勢のような高次元状態空間への適用は，計算量の観点から困難である[Ishikura et al., 2012]．Unscented 変換は単峰性の確率分布に従う確率変数に任意の非線形変換を施した後の確率分布の平均と分散を求める手法で，変換前分布の平均と分散が既知と仮定することにより，少ないサンプルで効率的な分布推定を行う．

変換前分布の確率変数を $\boldsymbol{x} \in \mathbb{R}^L$，その平均と分散を $\boldsymbol{\mu}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}$ としたときの，非線形変換 f によって変換された $f(\boldsymbol{x})$ の分布を Unscented 変換により推定する方法を述べる．まず，$2L+1$ 個のシグマ点 $\boldsymbol{\chi}_0, \cdots, \boldsymbol{\chi}_{2L}$ と呼ばれる

サンプル点を生成する

$$\boldsymbol{\chi}_0 = \boldsymbol{\mu}_{\boldsymbol{x}} \tag{7}$$

$$\boldsymbol{\chi}_i = \boldsymbol{\mu}_{\boldsymbol{x}} + (\sqrt{(\alpha^2 L)\Sigma_{\boldsymbol{x}}})_i \quad \text{for } i = 1, \cdots, L \tag{8}$$

$$\boldsymbol{\chi}_{i+L} = \boldsymbol{\mu}_{\boldsymbol{x}} - (\sqrt{(\alpha^2 L)\Sigma_{\boldsymbol{x}}})_i \quad \text{for } i = 1, \cdots, L \tag{9}$$

ただし，$\sqrt{\cdot}$ は行列の平方根，$(\cdot)$ は行列の i 番目の列，α はスケーリングパラメータを表す．次にシグマ点に非線形変換 f を施し，変換後のシグマ点 $\boldsymbol{Z}_0, \cdots, \boldsymbol{Z}_{2L+1}$ を得る．

$$\boldsymbol{Z}_i = f(\boldsymbol{\chi}_i) \quad \text{for } i = 0, \cdots, L \tag{10}$$

得られたサンプル点を用いて，変換後分布の平均 $\boldsymbol{\mu}_{\boldsymbol{z}}$ と分散 $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ は以下のように推定される．

$$\boldsymbol{\mu}_{\boldsymbol{z}} = \sum_{i=0}^{2L} w_i^m \boldsymbol{Z}_i \tag{11}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{z}} = \sum_{i=0}^{2L} w_i^c (\boldsymbol{Z}_i - \boldsymbol{\mu}_{\boldsymbol{z}})(\boldsymbol{Z}_i - \boldsymbol{\mu}_{\boldsymbol{z}})^T \tag{12}$$

$$w_i^m = \begin{cases} (\alpha - 1)^2 L/(\alpha^2 L) & \text{if } i = 0 \\ 1/\{2(\alpha^2 L)\} & \text{otherwise} \end{cases} \tag{13}$$

$$w_i^c = \begin{cases} (\alpha - 1)^2 L/(\alpha^2 L) + 1 - \alpha^2 + \beta & \text{if } i = 0 \\ 1/\{2(\alpha^2 L)\} & \text{otherwise} \end{cases} \tag{14}$$

ただし，β はスケーリングパラメータである．

エントロピーの計算　スピーカ選択のコスト関数であるエントロピー $E[H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k}, n_{1:k+1})]$ の定義とその計算法について述べる．$E[H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k}, n_{1:k+1})]$ は以下のように定義される．

$$E[H(X|\boldsymbol{y}_{1:k}, n_{1:k+1})] = \int p(\boldsymbol{y}_{k+1}|\boldsymbol{y}_{1:k}) H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k+1}, n_{k+1}) d\boldsymbol{y}_{k+1} \tag{15}$$

ただし $H(X)$ は $H(X) = -\int p(X)\log(p(X))$ で定義されるエントロピー関数である．

以上より $E[H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k}, n_{1:k+1})]$ は，$p(\boldsymbol{y}_{k+1}|\boldsymbol{y}_{1:k})$ の平均と分散に，f を以下のように定義して Unscented 変換を行うことで得る．

$$f(\boldsymbol{y}_{k+1}) = H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k+1}, n_{1:k+1}) \tag{16}$$

計算に必要な $p(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k+1}, n_{1:k+1})$ は，Unscented Kalman Filter を用いて計算する．また，$H(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k+1}, n_{1:k+1})$ は $p(\boldsymbol{\xi}_{k+1}|\boldsymbol{y}_{1:k+1}, n_{1:k+1})$ がガウス分布 (平均 $\boldsymbol{\mu}$，分散 $\boldsymbol{\Sigma}$) と仮定し，以下として計算する．

$$H(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2}(1 + ln(2\pi)) + \frac{1}{2} ln|\boldsymbol{\Sigma}'| \tag{17}$$

ただし，$\boldsymbol{\Sigma}'$ は Unscented 変換による誤差を軽減するための $\boldsymbol{\Sigma}$ の対角成分以外を 0 とした共分散行列である．

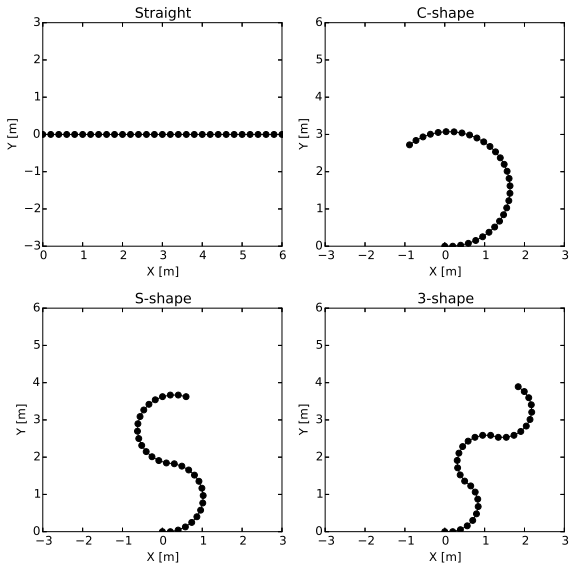


Figure 5: 実験で用いた姿勢 (マイク数:16)．それぞれ直線，C 字，S 字，3 の字を表す．

3　実験

実験ではロボットの姿勢から幾何的に計算した到達時間差を用いる数値実験による評価を行う．本実験では，スピーカの再生順について 1) 順番 (従来法)，2) ランダム，3) 提案法の 3 種を比較する．

3.1　実験設定

マイク・スピーカ間隔が 20 cm の柔軟索状ロボットを想定し，以下の条件において，姿勢推定の収束速度と姿勢推定精度を評価した．

- スピーカ選択法: 順番 (従来)，ランダム，提案法
- 姿勢: 直線，C 字，S 字，3 の字 (Figure 5)
- マイク数: 8, 16, 24 個 (それぞれ 2.8, 6, 9.2 m)

また，状態変数の初期値 $\boldsymbol{\xi}_0$ は，正解データを平均として標準偏差 $2\pi/(N + M - 2)$ rad の正規分布からサンプリングした．その他パラメータは実験的に与え，すべての試行において同じパラメータを使用した．

10 種の異なる初期値を用いて姿勢推定し，その先端位置誤差を評価した．先端位置誤差は，正解データと推定姿勢の手元側のマイクロホンと小型スピーカの座標 $\boldsymbol{u}_{1,k}, \boldsymbol{v}_{1,k}$ を一致させたときの，先端のマイクロホン mic_M の位置誤差である．

3.2　結果

Figure 6 に各観測ごとの推定結果の先端位置誤差を示す．まず従来の順番にスピーカを鳴らした場合では，マイク数が多くなると，マイク数回の周期で先端位置誤差が振動している．一方で，ランダムや提案法の鳴らし方では，振動が抑圧されている．Figure 7 に各条件での先端位置

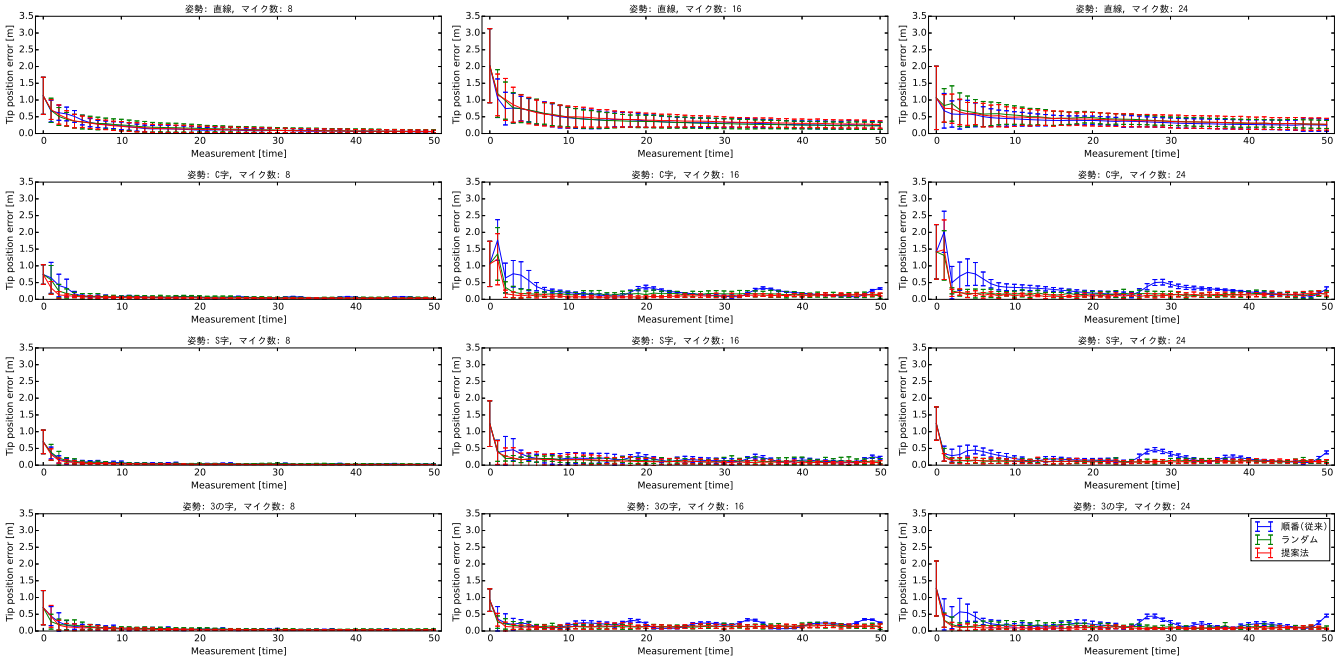


Figure 6: 姿勢推定結果の先端位置誤差．青が従来の順番にスピーカを選択する場合，緑がランダム，赤が提案法を表す．先端位置誤差の平均と標準偏差をそれぞれ折れ線とエラーバーを示す．

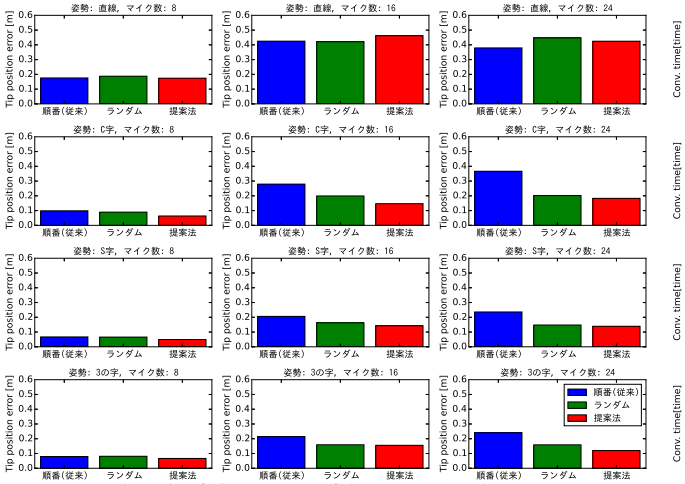


Figure 7: 姿勢推定結果の先端位置誤差の各条件での平均値．青が従来の順番にスピーカを選択する場合，緑がランダム，赤が提案法を表す．

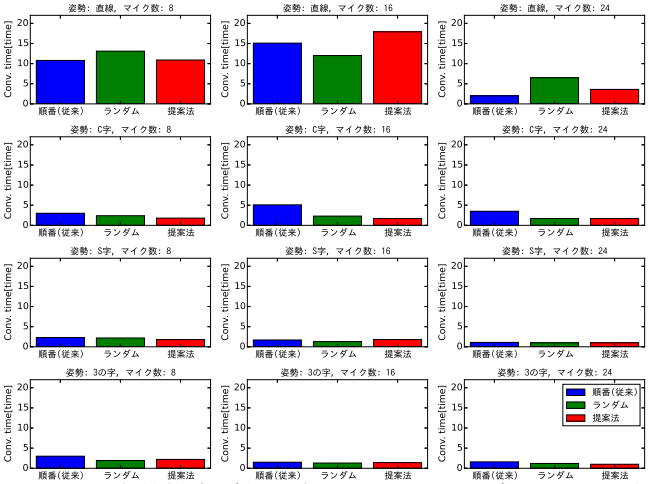


Figure 8: 姿勢推定が収束するまでに必要な観測回数の各条件での平均値．青が従来の順番にスピーカを選択する場合，緑がランダム，赤が提案法を表す．

誤差の平均を示す．提案法はマイク数が 8/16 で直線の姿勢を除いたすべての場合で最も先端位置誤差が小さい．

Figure 8 に各条件での推定姿勢が収束するまでの観測の平均回数を示す．収束判定は，先端位置誤差がロボットの全長の 5%になった時点とした．全 12 条件 8 条件で，提案法が従来の順番に鳴らす場合より収束が早い．一方で，12 条件中 4 条件で，ランダムにマイクを選択した方が提案法より早く収束しており，提案法には改善の余地がある．現在，提案法では 1 観測先の状態のみ予測している．オンライン強化学習による行動最適化手法として知られる partially observable Markov decision process [Thrun et al., 2005]では，複数観測先までの状態を予測し，行動決定を行う．提案法も同様に，予測する観測を増やしスピーカ選択の効率化が期待できる．

4　おわりに

本稿では，柔軟索状ロボットの音を用いた姿勢推定のために，スピーカの再生順序を強化学習により最適化する手法を開発し，収束速度と精度の向上を確認した．ロボットの姿勢から幾何的に計算した到達時間差を用いる数値実験を行い，従来の順番にスピーカを選択する場合より収束速度が最大 67%，先端位置精度が最大 50%向上することを確認した．一方で，直線型やマイク数が少ないときなどで，ランダムにスピーカを選択する場合の方が収束速度が早いことから，提案法には改善の余地があることが分かった．

今後は予測する観測のステップ数を増やし，より効率的なスピーカ選択法を開発する．また，モンテカルロ積

分や解析的な方法でよるスピーカ選択法と Unscented 変換を用いた提案法を比較し，提案法の精度と妥当性を確認する．さらに，これまで我々が開発した障害物存在下での信頼できないマイクロホンを棄却する技術[坂東 et al., 2014]と統合し，瓦礫内でも頑健で効率的な姿勢推定を実現する．

謝辞 本研究は科研費基盤 (S) No.24220006 の支援を受けた．

参考文献

[Arulampalam et al., 2002] M. Arulampalam et al. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[Baiocchi et al., 2013] V. Baiocchi et al. Development of a Software to Plan UAVs Stereoscopic Flight: An Application on Post Earthquake Scenario in L'Aquila City. In *ICCSA*, pages 150–165. Springer, 2013.

[Bando et al., 2013] Y. Bando et al. Posture estimation of hose-shaped robot using microphone array localization. In *IEEE/RSJ IROS*, pages 3446–3451, 2013.

[Czarnetzki et al., 2011] S. Czarnetzki et al. Real-time active vision by entropy minimization applied to localization. In *RoboCup 2010: Robot Soccer World Cup XIV*, pages 266–277. Springer, 2011.

[Hatazaki et al., 2007] K. Hatazaki et al. Active scope camera for urban search and rescue. In *IEEE/RSJ IROS*, pages 2596–2602, 2007.

[Ishikura et al., 2012] M. Ishikura et al. Shape estimation of flexible cable. In *IEEE/RSJ IROS*, pages 2539–2546, 2012.

[Julier, 2002] S. J. Julier. The scaled unscented transformation. In *American Control Conference*, volume 6, pages 4555–4559. IEEE, 2002.

[Kitagawa et al., 2003] A. Kitagawa et al. Development of small diameter Active Hose-II for search and life-prolongation of victims under debris. *Journal of Robotics and Mech.*, 15(5):474–481, 2003.

[Miura et al., 2011] H. Miura et al. SLAM-based online calibration of asynchronous microphone array for robot audition. In *IEEE/RSJ IROS*, pages 524–529, 2011.

[Mizumoto et al., 2011] T. Mizumoto et al. Design and implementation of selectable sound separation on the texai telepresence system using hark. *IEEE ICRA*, pages 2130–2137, 2011.

[Nagatani et al., 2011] K. Nagatani et al. Redesign of rescue mobile robot Quince. In *IEEE SSRR*, pages 13–18, 2011.

[Nakadai et al., 2010] K. Nakadai et al. Design and implementation of robot audition system HARK – open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[Namari et al., 2012] H. Namari et al. Tube-type active scope camera with high mobility and practical functionality. In *IEEE/RSJ IROS*, pages 3679–3686, 2012.

[Ohno et al., 2011] K. Ohno et al. Robotic control vehicle for measuring radiation in Fukushima Daiichi Nuclear Power Plant. In *IEEE SSRR*, pages 38–43, 2011.

[Ono et al., 2009] N. Ono et al. Blind alignment of asynchronously recorded signals for distributed microphone array. In *WASPAA*, pages 161–164, 2009.

[Tadokoro et al., 2009] Satoshi Tadokoro et al. Application of active scope camera to forensic investigation of construction accident. In *IEEE ARSO*, pages 47–50, 2009.

[Thrun et al., 2005] S. Thrun et al. *Probabilistic robotics*. MIT Press, 2005.

[Voyles et al., 2012] R. Voyles et al. Hexrotor UAV platform enabling dextrous interaction with structures – preliminary work. In *IEEE SSRR*, pages 1–7, 2012.

[Wan et al., 2000] Eric A Wan et al. The unscented Kalman filter for nonlinear estimation. In *IEEE AS-SPCC*, pages 153–158, 2000.

[坂東 et al., 2014] 宜昭 坂東 et al. マイクロホンアレイの位置推定によるホース型ロボットの姿勢推定. In **情報処理学会第 *76* 回全国大会**, *5R-7*, 2014.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-09

Robust Hands-free Human-Robot Communication in Reverberant Environments

Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto and Kazuhiro Nakadai

***Abstract*— Speech-based human-robot interaction is often plagued with issues such as reverberation and changes in speaker position that impacts overall performance. In this paper, we show a method in compensating the joint effects of reverberation and the change in speaker position. The acoustic perturbation caused by these two takes its toll on the Automatic Speech Recognition (ASR) and then the Spoken Language Understanding (SLU). Consequently, these will lead to a failure in the human-robot interaction experience. The proposed method is specifically designed to address the challenging environment condition in which robots are deployed. First, we analyze the impact of reverberation in the form of temporal smearing per change in speaker position. Then, we extract the smearing coefficients that capture the joint dynamics between the speech signal at current position and the room acoustics as observed by the robot. These coefficients are utilized to update the room transfer function (RTF) and the suppression parameters are stored offline. Moreover, all of these processes are optimized in the context of the ASR system for robot application. In the online mode, the reverberant data at an arbitrary position is processed using the parameters pre-computed offline. This effectively compensates the joint effects of reverberation at the arbitrary speaker position. Experimental results using real data gathered in a human-robot communication setting show that the proposed method outperforms existing methods.**

***Index Terms*— Speech Enhancement, Dereverberation, Robustness, Automatic Speech Recognition**

I. Introduction

Reverberation is a phenomenon caused by the reflections of the speech source in an enclosed environment. It is characterized by the smearing effect to the original speech due to the different time delays of arrival of the reflected speech source. The smearing degrades the Automatic Speech Recognition (ASR) system due to mismatch in the Hidden Markov Model (HMM) and impacts the Spoken Language Understanding (SLU) system as well. The overall impact may lead to the failure in human-robot communication experience. To mitigate this, the observed reverberant speech is enhanced through dereverberation. There exists different types of dereverberation methods [1][2] and most of these are originally formulated using human perception criterion and later applied to the ASR system in robot applications. Although this approach works well, the dereverberation method is not optimized for robot environment which is a very challenging task.

We note that in real robot environment, it is very difficult to control the position of the speaker when interacting with the robot as depicted in Fig. 1. For an immersive interaction, the speaker cannot be restricted as to where he initiates the conversation. Moreover, to effectively suppress the effects of the smearing caused by reverberation for robot application, it is important to incorporate the dynamic speaker position in the reverberation model. This means that the changes in speaker positions and dereverberation should be analyzed altogether when addressing the reverberation problem, to be effective in robot application. In this paper, we improve our previous work [4][13] by compensating the changes in speaker position via ASR optimization.

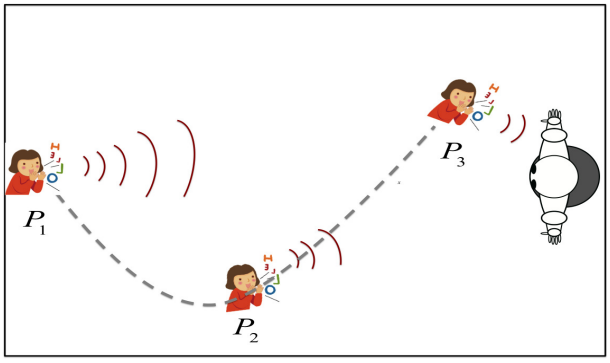


Fig. 1. Problem: Changes in speaker position in an enclosed reverberant room.

Our previous work [4][13] does not take into consideration the joint dynamics of the room characteristics and the change in speaker position. These two were treated independently in our previous work [4][13] but in reality there is a very strong link between the two. In addition, our previous method is more focused on the temporal side of the speech (i.e., waveform) and just stops right there. When dealing with ASR, the temporal representation of speech has its dual in the form of connected symbols which represents the sound units. Each of the sound unit are modelled by the Hidden Markov Models (HMMs). Therefore, it is very important to treat the latter equally likely with the waveform which is not addressed in our previous work [4][13]. In short, there is no mechanism in the previous method to operate in the HMM level. For example, the concept of frame-wise energy may exist but its analysis does not go deeper as to relate it with energy transfer across HMM states. This renders a very coarse treatment of the effect of reverberation when applied to HMM-based ASR, especially in challenging robot environment. In the proposed method, we have expanded the traditional reverberation model to treat jointly the effects of both the waveform and the HMMs. In particular, the effect of acoustic perturbation due to the changes in speaker position is tightly integrated in the dereverberation mechanism

Honda Research Institute Japan Ltd. Co., Honcho Wako-shi, Japan

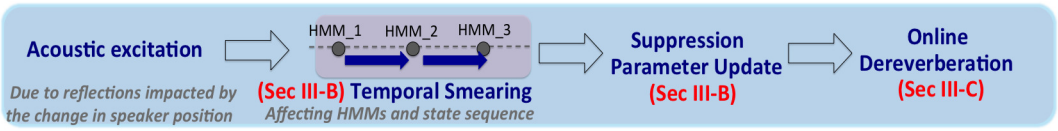


Fig. 2. Overall diagram of the proposed concept.

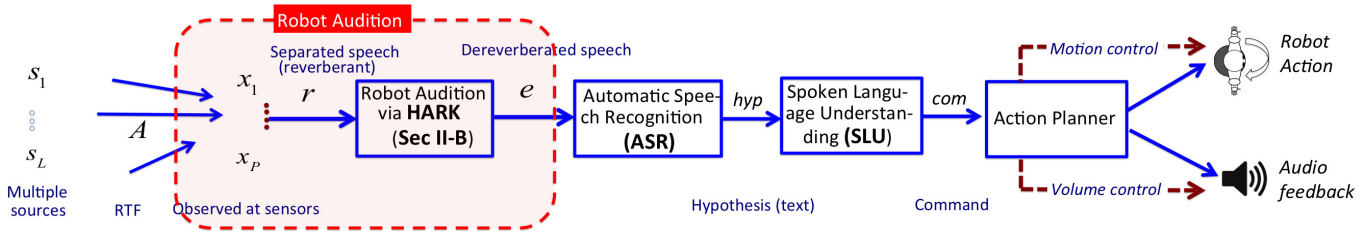


Fig. 3. Block diagram of a voice-based human-robot interactive system.

itself. We implemented a data-driven optimization scheme to extract parameters reflective of the dynamics between the speech and the room characteristics. We note that acoustic dynamics varies as perturbed by the change in position. In addition, the design methodology of the proposed method is hinged to the HMM (i.e., state sequence) which is an integral component of the ASR.

The concept of our approach is shown in Fig. 2. The recognized speech is treated not just a waveform but a sequence of sound units (i.e., phoneme HMMs). It is presumed that each sound unit is characterized by a unique frequency response that behaves differently given an acoustic excitation. In the same figure, it is shown that reflections inside the room drives an acoustic excitation that causes **temporal smearing** acting on the connected HMM sound units and within each HMM. Similarly, at each HMM, temporal smearing spans from one HMM state to the next. The smearing is hypothesized to be caused by the perturbation of the room acoustics due to changes in robot-speaker position. By modelling the smearing effect of reverberation, it is possible to compensate the changes in speaker positions objectively. In this paper we show the method of using a data-driven processing to empirically model the temporal smearing for dereverberation as a function of speaker position. Consequently, we improve the dereverberation performance of our proposed method using the ASR and SLU as metrics.

This paper is organized as follows; in Sec. II, we show the background of the previous reverberation model and the concept of dereverberation. The method in optimizing the temporal smearing for effective dereverberation is discussed in Sec. III. Experimental set-up with actual robot is discussed in Sec. IV followed by results and discussion in Sec. V. Finally, we conclude the paper in Sec. VI.

II. Background

A. Speech Communication-based Human-robot Interaction

Fig. 3 is an example of our human-robot interaction set-up. First, the robot audition framework based on HARK [17] is employed to process the multiple sound sources $S_1, ..., S_L$ as observed at microphones $x_1, ..., x_P$ into separated speech r. Then, enhanced to e via dereverberation. The dereverberated speech is used as input to the ASR system which outputs the hypothesis hyp. Consequently, the SLU system extracts the command com information from hypothesis. Lastly, a robot action is executed which includes motion and/or audio feedback. It is obvious in this figure that we need a robust robot audition system that can support speech communication in adverse conditions. This is vital in achieving a successful robot understanding. Thus, it is imperative that we compensate the dereverberation mechanism per change in speaker position. In reality, reverberation is not the only problem in attaining robust robot audition system. The following are other common problems

- Ego Noise (Noise from within the robot's moving parts)
- Directional Noise (External noise)
- Background Noise (External/Internal but additive in nature)
- Voice Activity Detection (Detecting speech segments)

Most of the problems above are already integrated in HARK, and in this paper we will focus only on the reverberation problem.

B. Robot Audition

Microphone array processing based on beamforming and blind separation described in [9][17] is employed to convert the multi-microphone observed signals $x_1, ..., x_P$ resulting to the separated reverberant signal $r(\omega)$. Moreover, we note that the RTF denoted by $A(\omega)$ is readily available during the microphone array processing [9][17]. However, $A(\omega)$ is assumed to be constant, but in real-world application this may not hold true any more especially when room size is factored in together with the robot and the objects inside the room. More specifically, room acoustics is more likely to change due to the acoustic perturbation caused by the changes in speaker positions. In the end, $A(\omega)$ needs to be updated. In our previous method [4][13], the smearing

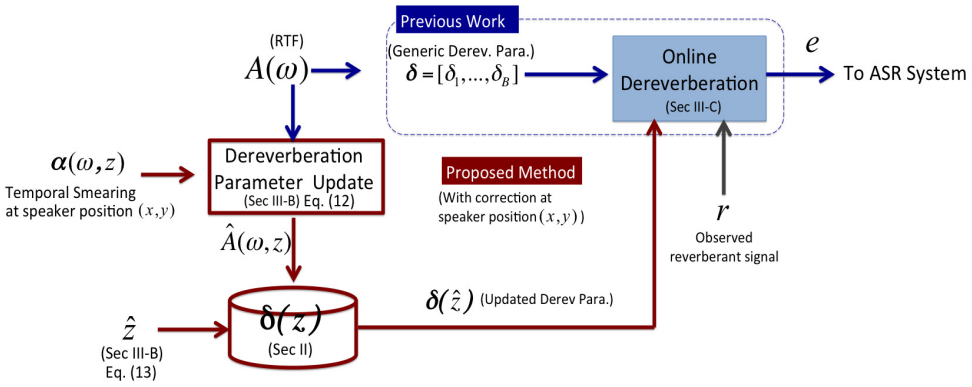


Fig. 4. Block diagram of the proposed method.

effect of reverberation is adopted from [15][5] and is solely dependent on the room transfer function (RTF) given as

$$\begin{aligned} r(\omega) &= A^E(\omega)c(\omega) + A^L(\omega)c(\omega) \\ &= \quad e(\omega) \quad + \quad l(\omega), \end{aligned} \tag{1}$$

where $r(\omega)$ is the reverberant speech model w.r.t. ω frequency, $c(\omega)$ is the clean speech, $A^E(\omega)$ and $A^L(\omega)$ are the early and late reflection components extracted from the full RTF $A(\omega)$. Both $A^E(\omega)$ and $A^L(\omega)$ are experimentally predetermined in [13]. $r(\omega)$ can be treated as the superposition of $e(\omega)$ and $l(\omega)$, known as the early and late reflection, respectively. In [13] we treat $l(\omega)$ as long-period noise which is detrimental to the ASR, and dereverberation is defined as suppressing $l(\omega)$ and recovering $e(\omega)$ estimate. The latter is further processed with Cepstrum Mean Normalization (CMN) during ASR. Eq. (1) simplifies dereverberation into a denoising problem, and through spectral subtraction (SS) [10], the estimate $\hat{e}(\omega)$ in frame-wise manner j is given as

$$|e(\omega,j)|^2 = \begin{cases} |r(\omega,j)|^2 - |l(\omega,j)|^2 \\ \quad \text{if } |r(\omega,j)|^2 - |l(\omega,j)|^2 > 0 \\ \\ \beta|r(\omega,j)|^2 \quad \text{otherwise.} \end{cases} \tag{2}$$

where β is the flooring coefficient. In real condition, $l(\omega, j)$ is unavailable, precluding the power estimate $|l(\omega, j)|^2$. A scheme in [13] shows a workaround to this problem, approximating $l(\omega, j)$ directly from the observed reverberant signal $r(\omega, j)$ through the error

$$E_m = \frac{1}{J} \sum_j \sum_{\delta_b \in B_q} |l(\omega,j) - \delta_b(\omega,j) r(\omega,j)|^2. \tag{3}$$

For the given set of bands $\boldsymbol{B} = \{B_1, \ldots, B_Q\}$, the suppression parameter δ_b is determined through minimum mean square error criterion in Eq. (3) via offline training discussed in [4][13]. The multi-band treatment improves error minimization as opposed to single-band. The new estimate $\hat{e}(\omega)$ through the modified SS becomes

$$|e(\omega,j)|^2 = \begin{cases} |r(\omega,j)|^2 - \delta_b |r(\omega,j)|^2 \\ \quad \text{if } |r(\omega,j)|^2 - \delta_b |r(\omega,j)|^2 > 0 \\ \\ \beta|r(\omega,j)|^2 \quad \text{otherwise.} \end{cases} \tag{4}$$

It is obvious that the dereverberation platform in Eq. (4) is dependent on the suppression parameter $\boldsymbol{\delta}$. Consequently, $\boldsymbol{\delta}$ depends on the RTF-centric reverberation model in Eq. (1). Although Eq. (1) is effective for waveform enhancement, it does not have any provision for HMM analysis (i.e., energy transfer in the HMM level) whenever the room acoustics is perturbed. Perturbation exists especially when the sound source (i.e. speaker) changes position relative to the robot. We note that in real scenario the relative position between the human and robot always changes. Thus, it is imperative that the original $A(\omega)$ needs to be updated as we cannot expect that the current acoustic perturbation is still reflective of the original $A(\omega)$. With no update capability, the dereverberation performance is very limited since the suppression parameter is frozen together with the RTF. The simplified block diagram of the previous and proposed methods is shown in Fig. 4. In the proposed method, the suppression parameter can be updated to $\boldsymbol{\delta}(\hat{z})$ depending on the joint dynamics of the room characteristics and the observed reverberant signal as characterized by $\alpha(\hat{z})$. Where $\hat{z}$ is the most probable HMM state sequence. Fig. 4 is explained in detail in the following section.

III. Methods

A. Database

The clean speech database used in the ASR is utilized to generate the reverberant database. The word-level text transcript is converted to phoneme-level transcript. The clean speech database is re-played using a loudspeaker inside a reverberant room and recorded by a microphone located at distance away from the loudspeaker. The newly recorded speech data becomes the reverberant database r. In this paper, we are interested on the basic sound units defined as the phonemes in our application. Thus, when referring to sound units, these basically come from the speech database itself.

B. Optimized Temporal Smearing Coefficients for suppression parameter Update

Suppose that the observed reverberant speech when processed by a filter is given as

$$o[n] = \sum_{m=0}^{M-1} \alpha_m \;\; r[n-m] \tag{5}$$

where r is the observed reverberant data and the temporal smearing filter which is

$$\boldsymbol{\alpha} = [\alpha_0, \alpha_1, ..., \alpha_{M-1}]^T, \tag{6}$$

is unknown. The length of the filter M samples can be indirectly associated to the extent of reverberation (i.e., reverberation time). It is hypothesised that $\boldsymbol{\alpha}$ charcterizes the joint acoustic perturbation due to reverberation and the changes in speaker position. The objective is to estimate $\boldsymbol{\alpha}$ in the context of the ASR system. Thus, the resulting estimate would capture the temporal smearing characteristics associated to the joint dynamics of the room characteristics (RTF) and the actual sound units spoken at an arbitrary position. We assume that $\boldsymbol{\alpha}$ is associated to a change in the speaker position (x, y) but we will drop the position notation (x, y) for simplicity. For now, the actual signal o is immaterial since we are interested with the ASR's output (hypothesis) which is given as

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \log \left(P(f^{(o)}(\boldsymbol{\alpha})|\boldsymbol{w})P(\boldsymbol{w})\right. \tag{7}$$

where $f^{(o)}(\boldsymbol{\alpha})$ is the extracted feature vector from the utterance, $\boldsymbol{w}$ is the phoneme-based transcript, $P(f^{(o)}(\boldsymbol{\alpha})|\boldsymbol{w})$ is the acoustic likelihood (i.e., using reverberant acoustic model) and $P(\boldsymbol{w})$ is due to the language (i.e., using language model). The latter can be ignored since phoneme-based transcript $\boldsymbol{w}$ is known, thus, argmax in Eq. (7) acts on $\boldsymbol{\alpha}$ which is rewritten as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \log P(f^{(o)}(\boldsymbol{\alpha})|\boldsymbol{w}). \tag{8}$$

In ASR, the total log likelihood in Eq. (8) when expanded [14] to include all possible state sequence in conjunction with the length of the smearing template is expressed as

$$\Gamma(\boldsymbol{\alpha}) = \sum_j \log P(f_j^{(o)}(\boldsymbol{\alpha})|\hat{s}_j), \tag{9}$$

where s_j is the state at frame j. Eq. (9) paves the formulation in analyzing the problem based on the HMMs in the form of state sequence. By using the ∇ operator, the total probability is maximized w.r.t the smearing coefficient in Eq. (6), thus,

$$\nabla_{\boldsymbol{\alpha}} \Gamma(\boldsymbol{\alpha}) \;\; = \left\{ \frac{\partial \Gamma(\boldsymbol{\alpha})}{\partial \alpha_0}, \frac{\partial \Gamma(\boldsymbol{\alpha})}{\partial \alpha_1}, ..., \frac{\partial \Gamma(\boldsymbol{\alpha})}{\partial \alpha_{M-1}} \right\}. \tag{10}$$

Assuming a Gaussian mixture distribution with mean vector μ_{jv} and diagonal covariance matrix $\boldsymbol{\Sigma}_{jv}^{-1}$, respectively. Eq. (10) can be shown similar to that in [8] as

$$\nabla_{\boldsymbol{\alpha}} \Gamma(\boldsymbol{\alpha}) = -\sum_j \sum_{\mathrm{v}=1}^{\mathrm{V}} \gamma_{jv} \frac{\partial f_j^{(o)}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \boldsymbol{\Sigma}_{jv}^{-1} (f_j^{(o)}(\boldsymbol{\alpha}) - \mu_{jv}). \tag{11}$$

where γ_{jv} is the posteriori of v mixture and j frame of the most likely HMM state. $\frac{\partial f_j{}^{(o)}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$ is the Jacobian matrix of the reverberant feature vector. The HMM-optimized smearing coefficients are obtained using [11][12] based on Eq. (11). In general, the HMM can generate Z-best most likely state sequences. Thus, from Eq. (8) we expand $\boldsymbol{\alpha}(z)$ corresponding to the Z possible HMM state sequence. In this manner we can capture the effect in the state sequence caused by the joint dynamics of the room characteristics and the sound excitation as perturbed by the change in speaker position.

The Z-best optimized smearing coefficients are used to update the readily available RTF $A(\omega)$ which is provided in the microphone array processing discussed in Sec. II. The RTF update done for all z is expressed as

$$\hat{A}(\omega, z) = \alpha(\omega, z) A(\omega) \tag{12}$$

where $\alpha(\omega, z)$ is the $z-th$ temporal smearing in frequency domain. Thus, several RTFs are generated using the update in Eq. (13). Then, suppression parameters $\boldsymbol{\delta}(z)$ are computed for each $\hat{A}(\omega, z)$ in the same manner as discussed in Eq. (3) in Sec. II, and these values are kept in the database. In the online mode, the acoustic likelihood of the observed reverberant data is filtered with the pre-computed $\alpha(z)$ for all z templates and the corresponding $\hat{z}$ is selected through

$$\hat{z} = \underset{z}{\operatorname{argmax}} P(f^{(\alpha(\omega,z))*r}|\boldsymbol{w}). \tag{13}$$

$\hat{z}$ signifies that the observed reverberant signal r is a close match to the corresponding $\alpha(\omega, \hat{z})$ in the acoustic likelihood criterion. Thus, its corresponding $\boldsymbol{\delta}(\hat{z})$ is selected as the updated suppression parameter.

C. Online Dereverberation

In the online mode, the system takes in as input the observed reverberant signal and select the optimal $\boldsymbol{\delta}(\hat{z})$ as described in Sec. III-B. $\boldsymbol{\delta}(\hat{z})$ is used as input for dereverberation. Specifically, the spectral subtraction in Eq. (4) is rewritten as

$$|\hat{e}(\omega, j)|^2 = \begin{cases} |r(\omega, j)|^2 - \delta_b(\hat{z})|r(\omega, j)|^2 \\ \quad \text{if } |r(\omega, j)|^2 - \\ \quad \delta_b(\hat{z})|r(\omega, j)|^2 > 0 \\ \\ \beta|r(\omega, j)|^2 \quad \text{otherwise.} \end{cases} \tag{14}$$

where $\delta_b(\hat{z})$ acts on the frame level j.

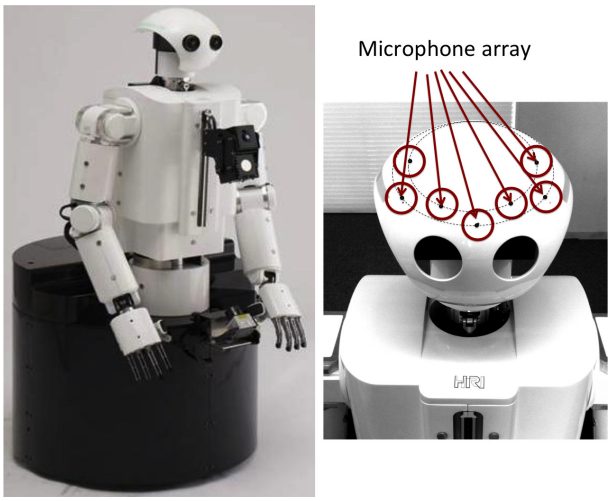


Fig. 5. HRI-JP humanoid robot "Hearbo".

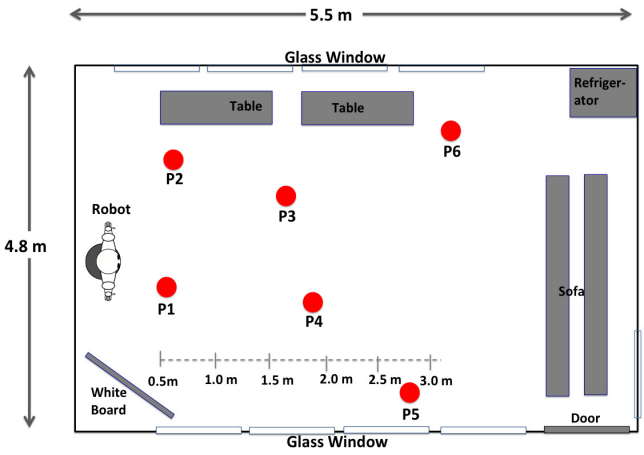


Fig. 6. Room set-up for testing (Room 4).

IV. Experiments with Robot

A. Humanoid Robot: Hearbo

The Honda Research Institute Japan's (HRI-JP's) humanoid robot named "Hearbo" is shown in Fig. 5. It has 20 degrees of freedom and its head is embedded with microphone array arranged in two concentric circles of different diameters. it is equipped with a robot audition software based on HARK [17] which implements microphone array methods for hands-free speech processing.

B. ASR and SLU Systems

The baseline acoustic model is a 3-state HMM based on Gaussian Mixture Models and trained using the World Street Journal corpus. The test data is composed of ten English speakers. Each person utters 20 utterances for each test position in $P1 - P6$ (see Figure 6). Hypothetically speaking, the test speakers may speak in freeform. However, the utterances for the actual testing are scripted to maintain uniformity and to avoid mistakes as these may impact the SLU performance.

The human-robot interaction setting re-enacts a sushi restaurant scene. The customer (speaker) may approach the robot at an unknown position (i.e., P1-P6) and engage via voice communication. In the course of the conversation, the speaker asks the robot questions about the variety of fish used in preparing the traditional Japanese dishes "Sushi" or "Sashimi". Upon recognition via the ASR system, the robot is tasked to translate the English fish name into its Japanese equivalence. Due to reverberation and the acoustic perturbation the observed reverberant speech is processed using our proposed method as shown in Figure 4 prior to ASR. Then the SLU system processes the output of the ASR system hyp to identify the fish name for the possible robot action. An example of the question from the customer would be, "Hearbo, we had Sweetfish yesterday for dinner. Can you tell me what it is called in Japanese ?". The robot should be able to identify that the fish in question is "Sweetfish" and be able to give its corresponding Japanese name. Part of the interaction is that, the robot automatically adjust its volume in accordance to its proximity with the speaker and being able to turn and face towards the speaker direction. In our experiment we provide both the ASR and SLU results to confirm whether the proposed method impacts both the ASR and SLU systems.

C. Room Condition

We conducted our experiment in four different room settings (Room 1- Room 4) with Reverberation Time (RT) of 80 ms., 240 ms., 900 ms. and 940 ms., respectively. Room 1 is the least reverberant while Room 4 exhibits the most effect of reverberation for having the longest RT among the four rooms. In this work, we only focus the effect of reverberation so the background noise has signal to noise ration of 20 dB only. An example of one of the rooms (i.e., Room 4 with RT = 940 ms.) is shown in Fig. 6. Test positions inside the room are denoted as P1-P6. Although the RT is different for each room, the test positions P1-P6 are purposely positioned at the same places for all of the four different rooms for uniformity. Thus, the robot-to-speaker distances are the same.

V. Results and Discussion

The ASR results in terms of word correct are shown in Table 1. The results are averaged over the four different rooms. Method (A) is the result when no enhancement was implemented while method (B) is the result based on Linear Prediction residual approach [1]. By exploiting the characteristics of the vocal chord, it is able to remove the effects of reverberation. The result in method (C) is based on wavelet extrema clustering [2]. Similar to that in [1] except that it operates in the wavelet domain to find and remove the effects of reverberation. Method (D) is based on adaptation by [16], Instead of suppression, this method minimizes the mismatch through adaptation of the feature vector. The method in (E) is the result based on the previous method [13][4] (Eq. (2)) employing the old reverberant model. The proposed method (F) is based on Eq. (14) employing the

TABLE I

ASR RESULTS AVERAGED ACROSS ALL ROOMS (ROOM 1-ROOM 4) IN WORD CORRECT RATE (%)

	P1	*P2*	*P3*	*P4*	*P5*	*P6*
(A) No Enhancement	90.0 %	84.1 %	74.2 %	69.5 %	43.9 %	27.3 %
(B) Based on LP Residuals [1]	90.2 %	86.1 %	77.0 %	72.2 %	58.3 %	42.4 %
(C) Based on Wavelet Extrema [2]	90.4 %	86.3 %	78.1 %	74.5 %	60.6 %	46.2 %
(D) Based on Feature Adaptation [16]	90.7 %	86.5 %	79.3 %	76.2 %	63.4 %	49.8 %
(E) Spectral Subtraction (Previous Reverberation Model) [4][13]	90.8 %	86.9 %	79.6 %	76.5 %	68.3 %	54.3 %
(F) Spectral Subtraction (Proposed Method)	**91.2 %**	**87.7 %**	**82.8 %**	**81.4 %**	**74.7 %**	**66.4 %**

TABLE II

SLU RESULTS AVERAGED ACROSS ALL ROOMS (ROOM 1-ROOM 4) IN CORRECTLY IDENTIFYING THE FISH NAME (%)

	P1	*P2*	*P3*	*P4*	*P5*	*P6*
(A) No Enhancement	100.0 %	94.0 %	83.0 %	78.0 %	35.0 %	10.0 %
(B) Based on LP Residuals [1]	100.0 %	94.0 %	85.0 %	80.0 %	61.0 %	30.0 %
(C) Based on Wavelet Extrema [2]	100.0 %	94.0 %	85.0 %	81.0 %	65.0 %	35.0 %
(D) Based on Feature Adaptation [16]	100.0 %	94.0 %	86.0 %	83.2 %	68.0 %	38.0 %
(E) Spectral Subtraction (Previous Reverberation Model) [4][13]	100.0 %	94.0 %	86.0 %	84.0 %	68.3 %	43.0 %
(F) Spectral Subtraction (Proposed Method)	**100.0 %**	**96.0 %**	**88.0 %**	**86.0 %**	**71.0 %**	**59.0 %**

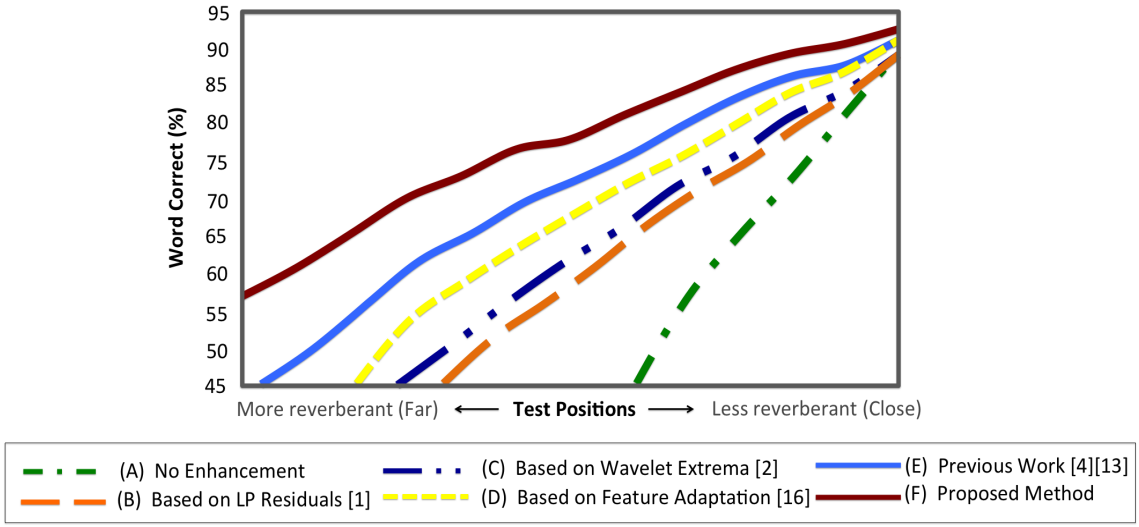


Fig. 7. Sorted ASR results using simulated data across Room 1- Room 5.

current reverberant model analysis that involves the notion of temporal smearing. In this table, we show that the propose method outperforms the existing methods and it is more effective farther distances. The adaptation based approach in [16] is only good in shorter reverberation time but performs poorly at longer reverberation time. This can be attributed to the fact that this method does not actually suppress the effects of dereverberation. We also show in Table 2 the results of the SLU system. This result confirms that the improvement in recognition performance attributed by the proposed method is translated in the machine understanding phase. Thus, the proposed method may positively impact interaction experience. In Fig. 7, we simulated the reverberant data inside the four different room by convolving a known RTF from the database and generate synthetic reverberant data. The purpose of this is to show the overall characteristics of the proposed method with more test data aside from the real recording in Table 1. We note that it is difficult to record different test points and synthetic reverberant data have been used and confirmed to show the same trend as real data. In this figure we concatenate and sort all the results from different rooms. We confirm the effectiveness of our proposed method in addition to that in Table 1.

The possible reasons why the proposed method fares better than the rest of the methods presented in this paper are: (1) the ability to update the suppression parameters reflective of the changes of the acoustic dynamics inside the room. It should be noted that depending on the acoustic excitation, acoustic room dynamics may change. (2) Formulation of the reverberation and optimization problems evolves in the HMM structure which is just proper since the dereverberation task is for the ASR system. This enables the processing of the acoustic waveform to better match the HMM-based ASR system. Lastly, (3) all of the optimization procedures are data-driven which results to a more realistic treatment of the effect of reverberation as opposed to just simply rely on the RTF.

VI. CONCLUSION

In this paper, we have shown the method of analyzing the reverberant model in an effective way that aids dereverberation for improved ASR performance. By analyzing the temporal smearing of the HMMs, we are able to incorporate the acoustic perturbation caused by the change in speaker position. We integrated it in the design process which is centered in the ASR system. This is very important because we have successfully expanded the traditional dereverberation method to environments in which robots are deployed. The proposed method is able to cope with demanding nature or human-robot communication such as the unpredictable change in speaker position. We have confirmed that the proposed method performs well in both real and synthetic data. Lastly, we confirmed the benefit of the proposed method is not just limited to the ASR system, more importantly it is able to improve the SLU performance as well. We note that the latter is a precursor of human-robot interaction experience. In our future work, we will consider the the effects of noise and investigate the prospect of expanding to deep neural networks (DNN).

REFERENCES

[1] B. Yegnanarayana and P. Satyaranyarana, "Enhancement of Reverberant Speech Using LP Residual Signals", *In Proceedings of IEEE Trans. on Audio, Speech and Lang. Proc.*, 2000.

[2] S. Griebel and M. Brandstein, "Wavelet Transform Extrema Clustering for Multi-channel Speech Dereverberation" *IEEE Workshop on Acoustic Echo and Noise Control*, 1999.

[3] K. Kinoshita , T. Nakatani and M. Miyoshi, Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation, *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2006.

[4] R. Gomez, K. Nakamura, and K. Nakadai, "Robustness to Speaker Position in Distant-Talking Automatic Speech Recognition" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2013.

[5] P. Naylor and N. Gaubitch, "Speech Dereverberation" *In Proceedings IWAENC*, 2005

[6] Akinobu Lee, *Multipurpose Large VocabularyContinuous Speech Recognition Engine*, 2001.

[7] S. Vaseghi "Advanced Signal processing and Digital Noise reduction", Wiley and Teubner, 1996.

[8] M. Seltzer, "Speech-Recognizer-Based Optimization for Microphone Array Processing *IEEE Signal Processing Letters*, 2003.

[9] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "Adaptive Step-size Parameter Control for real World Blind Source Separation" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008.

[10] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* ,1979.

[11] , "On numerical analysis of conjugate gradient method" *Japan Journal of Industrial and Applied Mathematics*, 1993.

[12] , W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes in C: The Art of Scientific Computing" *Cambridge University Press*, 1988 .

[13] R. Gomez and T. Kawahara, "Robust Speech Recognition based on suppression parameter Optimization using Acoustic Model Likelihood" *In Proceedings IEEE Transactions Speech and Acoustics Processing*, 2010.

[14] "The HTK documentation http://htk.eng.cam.ac.uk/docs/docs.shtml"

[15] E. Habets, "Single and Multi-microphone Speech Dereverberation Using Spectral Enhancement" *Ph.D. Thesis*, June 2007.

[16] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise" *Speech Communication*, pp 244-263, 2008.

[17] "HARK wiki http://winnie.kuis.kyoto-u.ac.jp/HARK/"

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-10

音源定位における能動耳介での動作の影響について

On effect of active pinnae motion to sound source localization

尾堂航, 公文誠
Wataru ODO, Makoto KUMON
熊本大学大学院自然科学研究科
Graduate School of Science and Technology, Kumamoto University
wataru@as.mech.kumamoto-u.ac.jp

Abstract

本研究では能動的に動作可能な反射板 (能動耳介) を備えた二つのマイクロホンを用いたロボットシステムにおいて, 能動耳介の動作が音源定位に与える影響を考える. 能動耳介は反射板の向きを変えて伝達特性を変化させ, 音源定位性能の向上を目指すシステムであるが, 能動耳介の動作時にエゴノイズが生じる. これまでこのエゴノイズが音源定位に与える影響は知られていなかった. 本研究では能動耳介の動作時におけるエゴノイズが音源定位に与える影響を明らかにする.

1 はじめに

私たち人間は, 外部の情報を得るために五感と呼ばれる感覚を用いて様々な情報を得ている. 特に聴覚においては, 左右の耳で聞こえた音の遅れや音の大きさの違いなどを脳で判断し, それを音源の方向として認識した結果, 方位角, 仰伏角, 距離などの情報を得ることができる. ロボットが私たちの身近で活動するためには, 人間の普段の生活環境を認識する必要がある. 従って人間のように音の情報によって周囲の環境を認識するシステムがロボットにとって求められる. ロボットにも音情報から音源の位置・方向を認識する音源定位と呼ばれる聴覚機能と, これを用いた環境を認識する能力が重要となる.

ロボットにおける音源定位については 3 つ以上のマイクロホンを用いるマイクロホンアレイによる研究が行われている. 例えば MUSIC 法やビームフォーマなどが良く知られた手法である [大賀 07]. マイクロホンの数が増えると音源定位が容易に行えるようになるが装置の大型化やマイクロホンを配置する空間の確保, 計算量の増加などシステムが複雑になるという問題点が挙げられる.

一方, 人間や動物は二つの耳で音の到来方向を認識することができ, 両耳間時間差 (Interaural Time Difference,ITD), 両耳間位相差 (Interaural Phase Difference,IPD), 両耳間レベル差 (Intersural Level Difference,ILD) などの音響特徴量を用いて音源定位を行うとされている [Garas 00]. ここから, ロボットも人間や動物と同様に音源定位に二つの耳を用いることが考えられる. このような二つのマイクロホンを用いたロボットの研究の一つに章ら [章 08] が提示する音源推定のための特徴量として ILD を用いた手法がある. これは事前に学習した特徴量の分布との相関を求めることで仰伏角や方位角を推定するものである. また西野ら [西野 07] はバイノーラル聴覚信号を用いて ILD を特徴量として単一ガウス分布で近似した音源方向モデルを構築することで仰伏角や方位角の推定を行っている. ところで人間や猫は頭部を動かすことによって音源定位能が向上することが知られている. 例えば, 猫頭部を固定すると水平方向の音源定位における分解能が劣化することが報告されており [Populin 98] 身体動作を伴う能動的な音源定位が重要であることが示唆されている. ロボットによる音源定位についても同様に能動的な作用を考えることは有用と考えられ, 金らは人工可動耳介を用いて追従動作を行うことで音源定位能力を向上させている.[金 12] また, 野田は確率的パラメータ推定を行い耳介の動作の下で正中面の音源定位を行っている.[野田 12] しかし, このような耳介の動作を行うには, モータ等で耳介を駆動することになりマイク近くで駆動騒音 (以下エゴノイズと呼ぶ) が生じることになり, 耳介動作そのものが音を用いた環境認識での妨げになる可能性がある. しかし, これまでの研究で能動耳介のエゴノイズが音源定位に与える影響について明らかになっていないため, 本研究では能動耳介の動作時におけるエゴノイズが音源定位に与える影響について調べることとした. 具体的には音声信号, 動作パターンの違いなどを考慮して数パターンの音収録を行い音源の方向を推定することで調査を行った.

2　能動耳介

本研究で用いる耳介を Fig.1,Fig.2 に示す．耳介は幅 70mm，奥行き 40mm，高さ 80mm となっており，マイクは耳介中央の奥行き 10mm の部分に配置してある．耳介を可動させるためのモータを含めると奥行きが 85mm となる．骨材となる部分はアクリル棒を用い，反射板となる部分は厚さ 1mm のアルミニウム板である．また，マイク正面に幅 30mm, 厚さ 4mm, 高さ 30mm の耳珠に相当する板を取り付けている．これは耳介による収録音への影響を際立たせるためである [本田 85]．耳介を円弧状の基台に取り付け，頭部にのせたものを Fig.3,Fig.4 に示す．

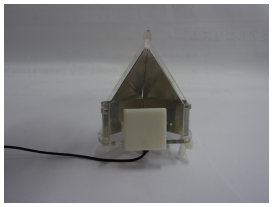


Figure 1: 正面図

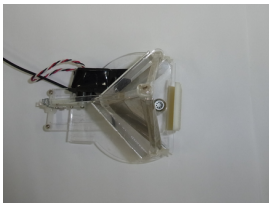


Figure 2: 上面図

Figure 3: 正面図

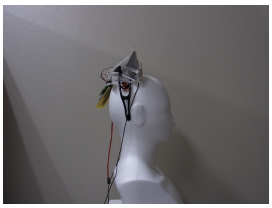


Figure 4: 側面図

耳介には RC サーボ 4ch が搭載されており，耳介一つにつき 2ch を用いる．アクリルによる骨材をモータで押し引きすることで上下方向の動きを可能にし, 左右の回転に関しては, 土台自体をモータにより回転させる (Fig.5). それぞれのサーボはマイコンによって制御され, マイコンへの指令値は Bluetooth 接続で外部の PC から行う構成になっている．Fig.5 から分かるようにマイク近くにモータが取り付けられており, 動作時には騒音を生じる.

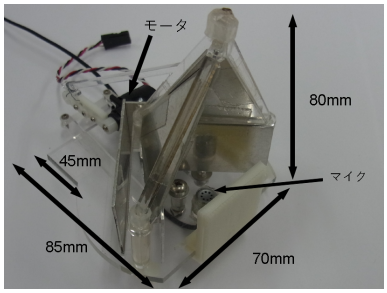


Figure 5: システム図

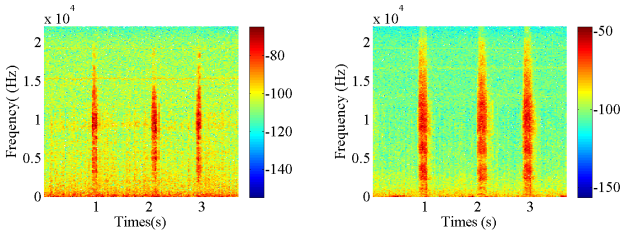
3　能動耳介動作時のエゴノイズ

3.1　エゴノイズの特徴

能動耳介の動作時に発生するエゴノイズを計測するため耳介を起こした姿勢と伏せた姿勢を繰り返し, 次の 2 つを動作パターンとして設定した.

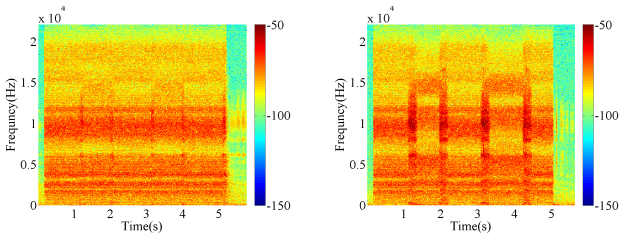
1. 耳介の一方のみを動作させる移動量が少ない動作パターン
2. 両耳介の可動範囲全体を移動する動作パターン

また, この動作を行いながら白色雑音または音楽をマイクにて収録した．受聴した音信号は, サンプリング周波数 44.1kHz で録音し, 以下では FFT 長 1024 点を 1 フレームとして処理をし, 周波数領域での信号を考える.



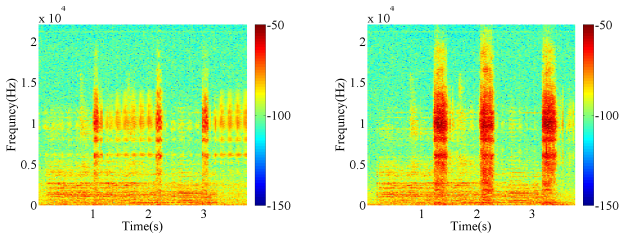
(a) 動作パターン 1　(b) 動作パターン 2

Figure 6: エゴノイズのみのスペクトログラム



(a) 動作パターン 1　(b) 動作パターン 2

Figure 7: 白色雑音を対象音とした時のスペクトログラム



(a) 動作パターン 1　(b) 動作パターン 2

Figure 8: 音楽を対象音とした時のスペクトログラム

Fig. 6 はエゴノイズのみを収録したスペクトログラムであり, 能動耳介から発生するエゴノイズは広い周波数帯域で瞬間的にパワーをもつ信号であることがわかる．さらに耳介の移動量によってノイズのパワーは異なるが, 帯域

や時間区間はほぼ共通しているため能動耳介のエゴノイズは耳介の移動量に依存しないと言える．Fig. 8 と Fig. 7 からエゴノイズのパワーは提示した対象音信号に比較して大きいことがわかる．これはエゴノイズの発生源である能動耳介のサーボモータが耳介の後方に設置してあり能動耳介の集音部に近いためであると考えられる．

3.2 エゴノイズ区間の検出

能動耳介のエゴノイズが短い時間に大きなパワーをもつ信号であることが分かったので, 雑音に影響される時間区間の検出は収録した音信号のパワーに注目すれば可能であると考えた．フレーム k において周波数 f の信号 $s_k(f)$ のパワーを $P_k(f)$ と書けば

$$P_k(f) = 10\log_{10} s_k^*(f)s_k(f) \tag{1}$$

と表せる．エゴノイズが全ての周波数帯域に現れることから $P_k(f)$ の合計を S_k とし

$$S_k = \sum_{f\in F} P_k(f) \tag{2}$$

を考える．ここで F は周波数点の全体集合を示す．適当な閾値 α の下で

$$S_k > \alpha \tag{3}$$

となるフレーム k はエゴノイズが生じていると判断すると考えられる．ノイズ源であるモータがマイクに非常に近く, 対象とする音信号に比べ, ノイズのパワーは顕著に大きいため容易に α を決定できる点は強調したい．

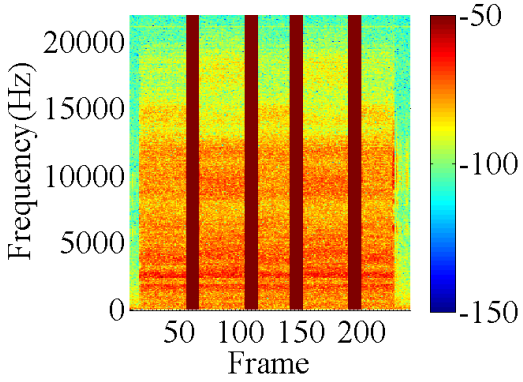


Figure 9: エゴノイズ除去後

実際に Fig. 8(b) に上記の検出法を適用し, ノイズ区間を除去した結果を Fig. 9 に示す．適切に機能していることが分かる．

4 両耳間レベル差と両耳間位相差を用いた音源定位

エゴノイズの影響を音源定位の観点から調べるため, まず定位法について説明する．

4.1 音源定位に用いる音響特徴量

周囲の環境による暗騒音や残響, ロボットの身体による反射や回折によりロボットが受聴する信号は原信号とは異なるものとなる．ある位置 $\boldsymbol{X}$ より得られる周波数を ω とし，耳介の姿勢を u とする．左右のマイクロホンの伝達関数のうちロボットの身体によるものは $H_l(\boldsymbol{X},u;\omega)$,$H_r(\boldsymbol{X},u;\omega)$, 環境によるものは $H_{le}(\boldsymbol{X};\omega)$,$H_{re}(\boldsymbol{X};\omega)$ と表すことができる．原信号 s_O に対してロボットの左右のマイクロホンで受聴する音信号をそれぞれ $s_l(\boldsymbol{X},u;\omega)$, $s_r(\boldsymbol{X},u;\omega)$ とすれば

$$s_l(\boldsymbol{X},u;\omega) = H_l(\boldsymbol{X},u;\omega)H_{le}(\boldsymbol{X};\omega)s_O(\omega) \tag{4}$$
$$s_r(\boldsymbol{X},u;\omega) = H_r(\boldsymbol{X},u;\omega)H_{re}(\boldsymbol{X};\omega)s_O(\omega) \tag{5}$$

の関係がある．もし環境の影響が $H_{le} \fallingdotseq H_{re}$ であれば両耳間レベル差 (Interaural Level Difference,ILD) を z_{ILD} と表すと

$$z_{\mathrm{ILD}} \equiv 20\log|H_l(\boldsymbol{X},u;\omega)| - 20\log|H_r(\boldsymbol{X},u;\omega)| \tag{6}$$

と近似でき z_{ILD} が u,ω の関数として $z_{\mathrm{ILD}}(u,\omega)$ となり, ロボットの身体のみの影響で特徴づけられる．

両耳間位相差は両耳間時間差と似た特徴量である．両耳間レベル差と同様に左右のマイクロホンの伝達関数を考えると原信号 s_O に対して以下のような両耳間位相差 (Interaural Phase Difference)z_{IPD} を得ることができる．

$$z_{\mathrm{IPD}} \equiv \angle H_l(\boldsymbol{X},u;\omega) - \angle H_r(\boldsymbol{X},u;\omega) \tag{7}$$

4.2 音源の尤度

音響特徴量の逆変換から直接, 音源の位置 (以下 ${}^{\mathrm{t}}\boldsymbol{X}$ と書く) を得ることは難しいため, 事前に周波数成分を十分に含んだ試験信号を与え, 位置と対応づけた音響特徴量を事前に計測し, あらかじめ測定された ILD,IPD のデータセットを規範として, 受聴信号と照らし合わせることで定位を行う．ロボットの姿勢を u, 周波数領域での ILD,IPD を表したベクトルをそれぞれ $\boldsymbol{z}_{\mathrm{ILD}}$,$\boldsymbol{z}_{\mathrm{IPD}}$ と表し音源定位を行うのに有効な周波数帯域を識別するインデックスベクトルを $\boldsymbol{z}_{\mathrm{ACT}}$ と表す．またこれらのベクトルの周波数 ω に対する要素を $\boldsymbol{z}_{\mathrm{ILD}}(\omega)$ などと書くとする．ロボットが受聴する信号 $s_l(\boldsymbol{X},u;\omega)$, $s_r(\boldsymbol{X},u;\omega)$ について, 下記のように z_{ACT} を定める．

$$z_{\mathrm{ACT}} = g(s_l(\boldsymbol{X},u;\omega), s_r(\boldsymbol{X},u;\omega)) \tag{8}$$

ここで g はインデックス関数であり,

$$g(x,y) = \begin{cases} 1 & \epsilon < x, \epsilon < y \\ 0 & (\text{上記以外}) \end{cases} \tag{9}$$

また ϵ は正の定数である．この操作によって ϵ より小さい値の受聴信号 $s_l(\boldsymbol{X},u;\omega)$,$s_r(\boldsymbol{X},u;\omega)$ を除外することが

できる．これらの観測などの情報の下での位置 $\boldsymbol{X}$ に音源の存在する ILD,IPD からの尤度 $l_{\mathrm{ILD}}(\boldsymbol{X}|u, \boldsymbol{z}_{\mathrm{ILD}}, \boldsymbol{z}_{\mathrm{ACT}})$, $l_{\mathrm{IPD}}(\boldsymbol{X}|u, \boldsymbol{z}_{\mathrm{IPD}}, \boldsymbol{z}_{\mathrm{ACT}})$ を

$$l_{\mathrm{ILD}}(\boldsymbol{X}|u, \boldsymbol{z}_{\mathrm{ILD}}, \boldsymbol{z}_{\mathrm{ACT}}) = \frac{\sum_{i=1}^{N} \boldsymbol{z}_{\mathrm{ACT}} \exp\left\{-10(\boldsymbol{z}_{\mathrm{ILD}}(\omega_i) - \boldsymbol{z}^d_{\mathrm{ILD}}(\boldsymbol{X}, u, \omega_i))^2\right\}}{\sum_{i=1}^{N} \boldsymbol{z}_{\mathrm{ACT}}} \quad (10)$$

$$l_{\mathrm{IPD}}(\boldsymbol{X}|u, \boldsymbol{z}_{\mathrm{IPD}}, \boldsymbol{z}_{\mathrm{ACT}}) = \frac{\sum_{i=1}^{N} \boldsymbol{z}_{\mathrm{ACT}} \exp\left\{-10(1-\cos(\boldsymbol{z}_{\mathrm{IPD}}(\omega_i) - \boldsymbol{z}^d_{\mathrm{IPD}}(\boldsymbol{X}, u, \omega_i))\right\}}{\sum_{i=1}^{N} \boldsymbol{z}_{\mathrm{ACT}}} \quad (11)$$

と定め，全観測からの音源位置の結合し尤度を $l(\boldsymbol{X}|u, \boldsymbol{z}_{\mathrm{ILD}}, \boldsymbol{z}_{\mathrm{IPD}}, \boldsymbol{z}_{\mathrm{ACT}})$ として表せば,

$$l(\boldsymbol{X}|u, \boldsymbol{z}_{\mathrm{ILD}}, \boldsymbol{z}_{\mathrm{IPD}}, \boldsymbol{z}_{\mathrm{ACT}}) = l_{\mathrm{ILD}}(\boldsymbol{X}|u, \boldsymbol{z}_{\mathrm{ILD}}, \boldsymbol{z}_{\mathrm{ACT}}) l_{\mathrm{IPD}}(\boldsymbol{X}|u, \boldsymbol{z}_{\mathrm{IPD}}, \boldsymbol{z}_{\mathrm{ACT}}) \quad (12)$$

のように与えるものとする．以下, この尤度に基づいて音源位置 ${}^{\mathrm{t}}\boldsymbol{X}$ を推定する．

4.3 ベイズ推定による音源定位

時刻 k でのロボットの姿勢を u_k, 得られた観測量を z_k, 音源方向を ${}^{\mathrm{t}}\boldsymbol{X}_k$ とする．また, 時刻 1 から k までの量を $u_{1:k}$ のように添字で表すことにする．耳介の傾きは u_k の関数として表現できるとする．

時刻 1 から k までの制御量 $u_{1:k}$ と観測量 $z_{1:k}$ が得られ時刻 k における音源方向 ${}^{\mathrm{t}}\boldsymbol{X}_k$ の存在についての信念が $B\left({}^{\mathrm{t}}\boldsymbol{X}_k|z_{1:k}, u_{1:k}\right)$ となっていたとする．${}^{t}\boldsymbol{X}_{k-1}$ が与えられたとするとモーションモデル h によって

$$B\left({}^{t}\boldsymbol{X}_k|z_{1:k-1}, u_{1:k}\right) = \int_{t\boldsymbol{X}_{k-1}} h({}^{t}\boldsymbol{X}_k|{}^{t}\boldsymbol{X}_{k-1}, u_k) B({}^{t}\boldsymbol{X}_{k-1}|z_{1:k-1}, u_{1:k-1}) d{}^{t}\boldsymbol{X}_{k-1} \quad (13)$$

となる. 本研究ではモーションモデルとして

$$h({}^{t}\boldsymbol{X}_{k+1}|{}^{t}\boldsymbol{X}_k) = \exp\left(-2({}^{t}\boldsymbol{X}_{k+1} - {}^{t}\boldsymbol{X}_k)^2\right) \quad (14)$$

を用いる．今, 制御量 u_{k+1} の操作の下で，新しい観測 z_{k+1} が得られたとするとベイズの推定法 [Thrun 05] から $k+1$ における音源位置に関する信念分布は

$$B\left({}^{\mathrm{t}}\boldsymbol{X}_{k+1}|z_{1:k+1}, u_{1:k+1}\right) = \frac{l\left(z_{k+1}|{}^{\mathrm{t}}\boldsymbol{X}_{k+1}, u_{1:k}, u_{k+1}\right) B\left({}^{\mathrm{t}}\boldsymbol{X}_{k+1}|z_{1:k}, u_{1:k}\right)}{\int_{\boldsymbol{X}} l\left(z_{k+1}|\boldsymbol{X}, u_{1:k}, u_{k+1}\right) B\left(\boldsymbol{X}|z_{1:k}, u_{1:k}\right) d\boldsymbol{X}} \quad (15)$$

と漸化的に求められる. これから, 初期時刻から $k+1$ まで式 (15) を観測ごとに繰り返し適用することで音源の存在位置に関する分布を求めることができる．また音源位置の最尤推定値を ${}^{\mathrm{t}}\hat{\boldsymbol{X}}_{k+1}$ とし次式で求める．

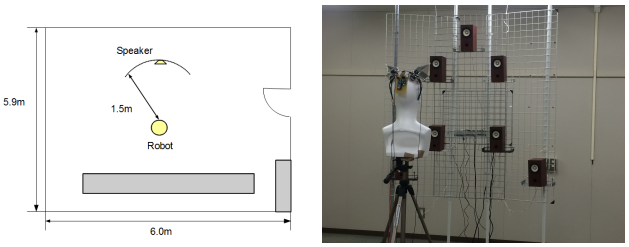
$${}^{\mathrm{t}}\hat{\boldsymbol{X}}_{k+1} = \arg\max B\left({}^{\mathrm{t}}\boldsymbol{X}_{k+1}|z_{1:k+1}, u_{1:k+1}\right) \quad (16)$$

5 音源定位におけるエゴノイズの影響

上述の定位法で音源位置を推定する際, 能動耳介の動作に伴うエゴノイズが与える影響について, 実際の実験データを基に考察する．

5.1 収録実験

収録は Fig.10 に示す奥行き 6.0m, 幅 5.9m, 高さ 2.5m の居室内で行った．音源として図の奥に見えるようなスピーカーを用いている．



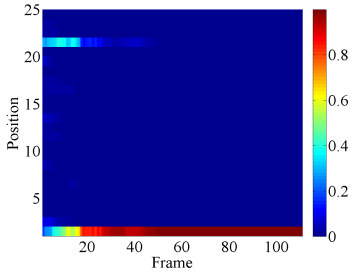
(a) 模式図　(b) 実験風景

Figure 10: 実験環境

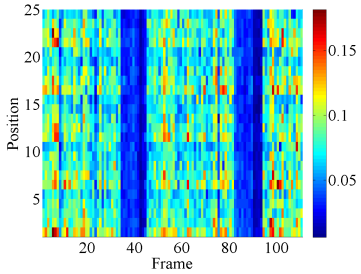
能動耳介を人形頭部に取り付け能動耳介の正面方向 1.5m, 正面高さ 1.5m を基準としロボットから見て方位角方向, 仰伏角方向ともに $-20°$ ～ $20°$ の $10°$ 刻みで計 25 点から白色雑音, 音楽をそれぞれ印加し計測を行い, さらにそれぞれの音源位置で耳介の姿勢を変化させながら収録を行った．姿勢は耳介を起こした姿勢と伏せた姿勢とを繰り返す動作パターンとした．

5.2 音源定位結果

音源が左下 (位置番号 1) にある時に収録したデータのスペクトログラムを Fig.11 に示し, このときの音源定位を行った結果を Fig.11 に示し縦軸は推定した音源位置のインデックス番号, 横軸は時間区間を示しており色によって信念 B を表している．



(a) 音源の存在確率



(b) 尤度

Figure 11: 定位結果

この Fig.11(a) 例では最初の 35 フレームまでにエゴノイズがなかったため, エゴノイズが生じた段階で, すでによく定位が行えており, その後のエゴノイズの影響があっても正解音源位置の検出が可能であることがわかる. しかし Fig.11(b) の 40 フレーム付近,90 フレーム付近の時間区間は尤度の値が他の時間区間と比べて低い値を示し, 明確な音源方向は得られていない. これらの時間区間はエゴノイズが発生した時間区間と一致することから, 事前分布の状態によってはエゴノイズが音源定位に悪影響を与えている可能性がある. そこで, これらのフレーム付近として, エゴノイズ前の 5 フレームとノイズ後の 20 フレームに着目し,

1. エゴノイズ区間の前後フレームとエゴノイズ区間を用いて音源定位を行う場合

2. エゴノイズ区間では式 (13), 式 (14) に替えて, 一切更新を行わない次のモデル

$$p\left({}^{t}\boldsymbol{X}_{k+1}|z_{1:k+1},u_{1:k+1}\right)=p\left({}^{t}\boldsymbol{X}_{k}|z_{1:k},u_{1:k}\right) \quad (17)$$

を適用し, エゴノイズ区間の前後の時間区間では式 (13), 式 (14) 用いて音源定位を行う場合

3. 2 と同様にエゴノイズ区間をスキップするがエゴノイズ区間中にも音源の存在確率分布の更新にモーションモデル式 (14) を用いて音源定位を行う場合

の 3 つについて検討することとした. 定位性能の評価には理想的な信念の分布と推定で得られた分布間の KL 情報量を用いた. KL 情報量は

$$D_{\mathbf{KL}}(R||Q)=\sum_{i}R(i)\log_2\frac{R(i)}{Q(i)} \quad (18)$$

と定義され, 本研究では R に正解位置の確率が 1 でそれ以外が 0 となる確率分布, Q に音源の信念分布 B を与えた.

Table 1: KL 情報量の変化

	1 の場合	2 の場合	3 の場合
KL 情報量の平均値	3.17	2.77	3.11

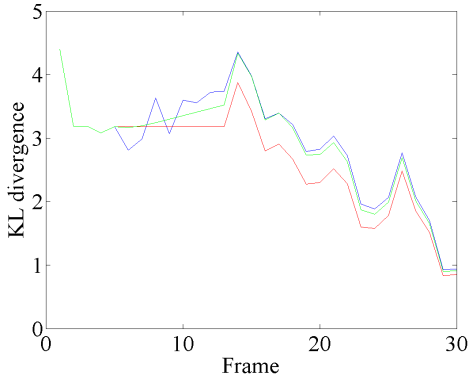


Figure 12: KL 情報量

この場合 D_{KL} が小さい程性能が良いことを示す. Fig.12 に正解音源位置 1 番での結果を示す. 横軸が時間区間, 縦軸が KL 情報量, 青は 1 の場合, 赤は 2 の場合, 緑は 3 の場合を示している. 5 フレームより後のエゴノイズ区間では 1,3 の場合は KL 情報量が増加し定位性能が低下しているが 2 の場合はエゴノイズ区間をスキップしているため KL 情報量が増加せず 1,3 の場合よりも良好な定位性能を維持している. このような評価を 25 点の異なる音源位置に対して実験し, その KL 情報量の平均を Tab.1 にまとめた. Tab.1 よりエゴノイズ区間を音源定位に用いない 2 の場合が KL 情報量が最も小さく音源定位性能が良い. このことはエゴノイズ区間での観測が信念の分布を理想的な分布から遠ざけていることを示しているが能動耳介の音源定位性能を下げていることが言える.

6 まとめ

能動耳介動作時のエゴノイズが音源定位性能に与える影響を調べた. エゴノイズ区間を音源定位の際に取り除くことで KL 情報量の増加を防ぐことができるため, エゴノイズが音源定位性能を下げることが能動耳介システムでも確認できたが他の音源や動作パターンについても検証を行うことが必要であると考えている.

参考文献

[大賀 07] 大賀 寿郎, 山崎 芳男, 金田 豊: 音響システムとディジタル処理, コロナ社, 2007.

[Garas 00] Garas J.: Adaptive 3D Sound Systems, Kluwer, 2000.

[章 08] 章 忠，井和　章，三宅　哲夫，今村　孝，堀畑　聡：バイノーラルモデルを用いた音源方向推定，日本機械学会論文集 C 編， Vol.74-739, pp. 642-649, 2008.

[西野 07] 西野　隆典, 井上直哉, 伊藤克亘, 武田一哉：両耳間音圧差の包絡を用いたガウス分布に基づく音源方向推定， 日本音響学会学会誌 C 編， Vol.63 no.1, pp. 3-12, 2007.

[Populin 98] Luis C. Populin , Tom C. T. Yin： Pinna Movements of the Cat during Sound Localization，Journal of Neuroscience， 18(11)， pp. 4233-4243, 1998.

[金 12] 金天海, 中臺一博, 辻野広司：ウェアラブル人工耳介-音追従動作による音源定位能力の向上-， 日本ロボット学会第 30 回記念学術講演会， 3D1-1, 2012.

[野田 12] 野田佳孝 公文誠： 二つの能動耳介による正中面内の音源方向推定，第 13 回システムインテグレーション部門講演会 (SI2012), pp1643-1646, 2012

[本田 85] 本田　学：耳珠のはたらき，耳鼻臨床， 78. 増 1, pp789-801, 1985.

[Thrun 05] S. Thrun, W.Burgard, and D.Fox： Probabilistic Robotics， MIT Press， 2005,

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

SIG-AI-Challenges home page (WWW): http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/