人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302

ＡＩチャレンジ研究会 (第３８回)

Proceedings of the 38th Meeting of Special Interest Group on AI Challenges

CONTENTS

◇ 【基調講演】ロボットの自律移動機能と音環境理解 1
加賀美聡，鮫島一平，佐々木洋子，畑尾直孝，トンプソン・サイモン，仁瓶雄真，江川拓良 (産業技術総合研究所デジタルヒューマン工学研究センター)
◇ ノンパラメトリックベイズモデルを用いた雑音ロバストな音響イベント同定 3
中村圭佑 ，ランディ・ゴメス，中臺一博 (HRI-JP)
◇ Semi-Blind Infinite NMF を用いた動作雑音抑圧手法の提案とその評価 9
手塚太貴（東京工業大学大学院），吉田尚水（東京工業大学大学院），中臺一博 （東京工業大学大学院/HRI-JP）
◇ Hands-free Speech Recognition Robust to distance and Azimuth in Robot Application 16
Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto and Kazuhiro Nakadai (HRI-JP)
◇ 【基調講演】野鳥の歌コミュニケーション理解への試み 22
鈴木麗璽（名古屋大学大学院），Charles E. Taylor, Martin L. Cody（UCLA）
◇ 複数のマイクロホンアレイを用いた理科室における音源アクティビティの分析 28
石井カルロス寿憲，Jani Even，塩見昌裕，萩田紀博 (ＡＴＲ知能ロボティクス研究所)
◇ ガウス回帰に基づく両耳間レベル差の補間 34
木元大輔，尾堂航，公文誠 (熊本大学)
◇ Combining Steered Response Power with 3D LIDAR scans for building sound maps 40
Jani Even, Yoichi Morales, Jonas Furrer, Carlos Toshinori Ishi, Norihiro Hagita (ATR-IRC)
◇ ロボット聴覚ソフトウェア HARK を用いたクイズの同時回答を識別するロボット司会者の設計と実装 45
西牟田勇哉，平山直樹，大塚琢馬，杉山治，糸山克寿，奥乃博 (京都大学大学院)
◇ ホースの伸び縮みによるマイク位置の変化を許容するマイクロホンアレイを用いたホース型ロボットの姿勢推定 51
坂東宜昭，大塚琢馬，糸山克寿（京都大学大学院），中村圭佑 (HRI-JP)，昆陽雅司，田所諭（東北大学大学院），中臺一博（東京工業大学大学院/HRI-JP)，奥乃博（京都大学大学院）

日　時　2013 年 12 月 6 日　　場　所　早稲田大学 西早稲田キャンパス 55 号館 S 第 4 会議室
Waseda University, Tokyo, Dec. 6, 2013

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

ロボットの自律移動機能と音環境理解

Autonomous Robot Navigation Functions with Auditory Environmental Mapping

○加賀美　聡、鮫島　一平、佐々木　洋子、畑尾　直孝、トンプソン・サイモン、
仁瓶　雄真、江川　拓良（産業技術総合研究所デジタルヒューマン工学研究センター）
*Satoshi KAGAMI, Yoko SASAKI, Ippei SAMEJIMA, Naotaka HATAO, Simon THOMPSON,
Yuma NIHEI, Takuro EGAWA (Digital Human Research Center, AIST, Japan)

s.kagami@aist.go.jp

Abstract—This paper describes autonomy navigation functions for mobile robots including auditory functions that authors are working on. Mapping, localization, static obstacle finding, moving human finding & prediction, path planning, path following, and sound source mapping are denoted.

1. はじめに

本報告では、自律移動するロボットのために必要な自律知能として、地図作成、位置認識、障害物発見、人追跡、経路計画、経路制御、音地図作成などの手法を組み合わせて実現する手法について紹介する。光学センサとしては LIDAR を、音センサとしては、著者らが開発してきた全方位低サイドローブマイクアレイを利用している。Fig.1 に日本科学未来館用に開発した Peacock を示す。最上部にマイクアレイを搭載し、LIDAR は中段におくことにより、目立たなくすることと、なるべく下方の検出領域を拡大させることを目指している。外装はタッチセンサにより構成され、触るとロボットが停止する機構となっている。

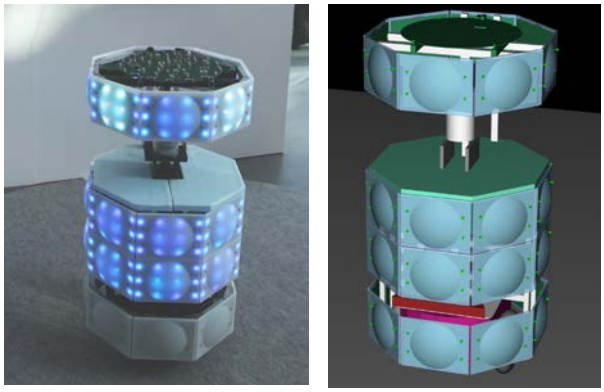


Fig.1. Mobile Robot Peacock equipped with LIDAR and microphone array

2. 自律移動のための諸機能

自律移動を行う際に、地図やその地図座標系での位置を利用者と共有できないと、利用者にとって意味のある行動をとることができない。ここでは LIDAR を利用して、人の頭上の高さで二次元地図をまずつくり、その二次元地図を初期値として三次元地図を作成している。日本科学未来館３F のフロア図と作成した二次元地図を Fig.2 に示す。位置認識も同様に人の頭上の高さを利用して行い、多数の来館者があった際にも、位置認識を安定に行えるようにしている[1)]。

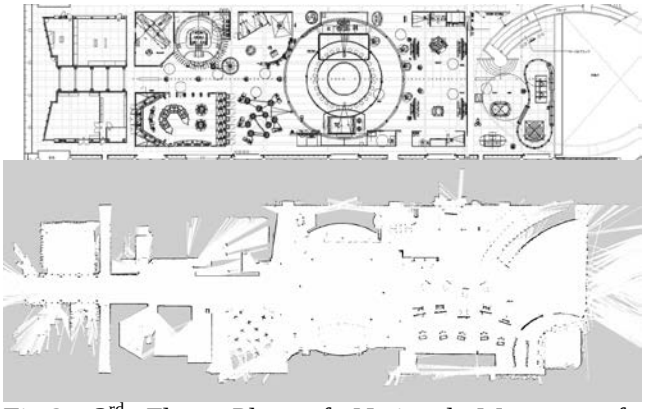


Fig.2. 3rd Floor Plan of National Museum of Emerging Science and Innovation (up) and obtained map using LIDAR (bottom)

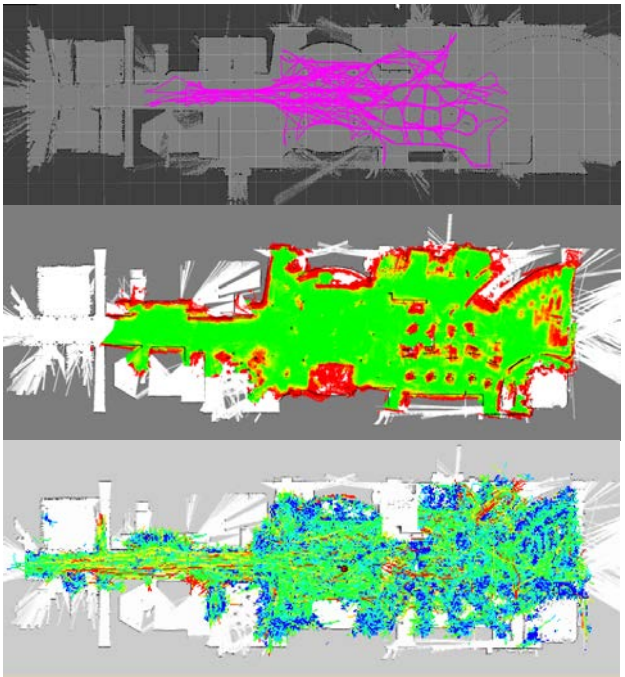


Fig.3. Localization result of about 2.5km run(up), detected static obstacle map(center), and detected walking human trajectories of 120,000 people. (bottom)

Fig3（上)図は、3 時間で約 2.5km の走行を行っ

た際の自己位置認識結果を示している。この実験において、particle filter の共分散は、角度で最大 0.07[rad]、平均 0.032[rad]、xy の面積で最大：567 [cm²]、平均 73[cm²]であった。

次に Fig3(中)図は、静止障害物の存在確率を示している。フロア内には、椅子、掲示板、子供が展示物を見るために上るスツールなどが存在しており、それらを地図に登録したものである[2)]。

最後に Fig3（下）図は、1 日観測したときの移動している人の軌跡のべ12万人分をプロットしたものである。色は速度を表している[2)]。

3. 音地図作成

ビームフォーミングの効果を最大化するために、32ch のマイクの最適配置問題を探索的に求めた。約 16,000 通りの配置候補の中で、全周波数帯域に渡ってメインローブ幅が小さく、サイドローブとの感度差が大きくなる配置を選択することにより、サイドローブゲインで全周波数帯域の平均で-16.8[dB]の感度の配置が得られた。Fig.4　に感度分布を示す[3)]。

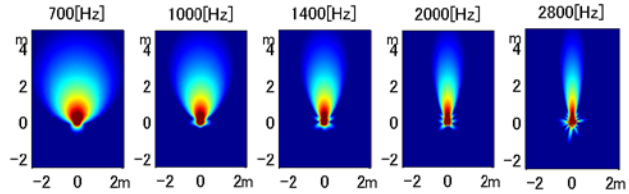
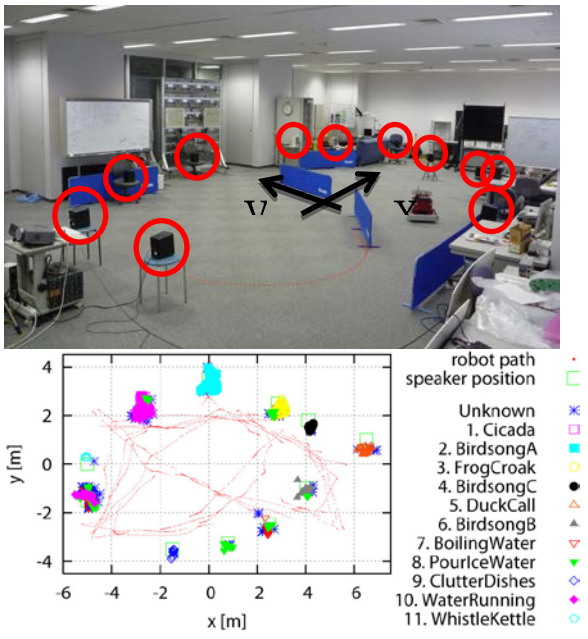


Fig.4. Directivity pattern of devleoped 32ch microphone array that focus on upward direction.

各瞬間には、定位した角度がわかるので、bearing only SLAM を実現する方法として、RANSAC を用いた motion stereo 法により、音源地図を得る方法を提案した[4)]。本手法を用いて、11 音源の地図を作成した結果を Fig5 に示す。



4. まとめ

本稿で示した peacock は、G 空間 Expo2013 で未来館を 3 日間にわたって自律走行を行った。また筆者らのグループでは同様の手法で、屋外を走行する segway[5)]、prius[6)]、建設機械[7)]と、さまざまな車両を自律移動させてきている。講演ではそれらの個別の案件についても紹介したい。

参考文献

1) 江川拓良、鮫島一平、仁瓶雄真、Simon Thompson、畑尾直孝、椛澤光隆、加賀美聡、竹村裕、溝口博：3 次元 LIDAR を用いた 2 次元環境地図の作成手法とそれを利用したナビゲーション，SI2013。

2) 鮫島 一平，仁瓶 雄真，畑尾 直孝，加賀美 聡，竹村 裕，溝口 博，大崎 章弘：日本科学未来館におけるサービスロボットのための人環境情報地図の構築、第 18 回ロボティクスシンポジア予稿集，pp.270--277,，Mar.，2013.

3) 佐々木 洋子，加賀美 聡，溝口 博：移動ロボット搭載用 32ch マイクロホンアレイの設計と精度評価、第 24 回日本ロボット学会学術講演会予稿集，pp.1B19，岡山大学，岡山市，Sep.，2006.

4) Yoko Sasaki, Satoshi Kagami, Hiroshi Mizoguchi : Multiple Sound Source Mapping for a Mobile Robot by Self-motion Triangulation, Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS2006), pp.380--385, China, Oct., 2006.

5) Simon Thompson, Satoshi Kagami, Masafumi Okajima : An Autonomous Mobile Inspection Robot for an Electric Power Sub-station, Proceedings of 10th International Conference on Informatics in Control, Automation and Robotics, pp.300--306, Iceland, Jul., 2013.

6) Satoshi Kagami, Simon Thompson, Ippei Samejima, Tsuyoshi Hamada, Shinpei Kato, Naotaka Hatao, Yuma Nihei, Takuro Egawa, Kazuya Takeda, Hiroshi Takemura, Hiroshi Mizoguchi : Autonomous Vehicle Navigation by Building 3D Map and by Detecting Human Trajectory using LIDAR, Proceedings of the 1st IEEE International Conference on Cyber-Physical Systems, Networks, and Applications, Taiwan, Aug., 2013.

7) 黒沼 出，浜本 研一，大塩 真，片村 立太，三浦 悟，加賀美 聡，サイモン トンプソン：放射線環境下における建設機械の自動運転システム、第 31 回日本ロボット学会学術講演会論文集，pp. 1 H3-01，東京，Sep.，2013.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-02

ノンパラメトリックベイズモデルを用いた雑音ロバストな音響イベント同定

Noise-robust Acoustic Event Identification Based on a Nonparametric Bayesian Model

中村 圭佑, ゴメス ランディ, 中臺 一博

Keisuke NAKAMURA, Randy GOMEZ, 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, r.gomez@jp.honda-ri.com, nakadai@jp.honda-ri.com

Abstract

本稿では，実環境応用のための雑音ロバストな音響イベント同定について述べる．既存の GMM などのパラメトリックモデルを用いた同定手法は，目的音に雑音や残響が混入した際の学習環境と音環境のミスマッチによって性能が劣化する問題があった．そこで本稿は Latent Dirichlet Allocation と Nested Pitman-Yor Process に基づくノンパラメトリックベイズモデルを用いて，大量の音環境データから音響イベント同定の統計的モデルを構築する手法を提案する．提案手法により，未知雑音が存在する環境下の音響イベント同定において，既存法より 5-18 pts の性能向上が確認できた．

1 序論

実環境での音を媒介としたシーン理解（音環境理解[1]）に関する研究が盛んに行われている．実環境における音環境は，音声だけでなく音楽や環境音などが含まれるため，音環境理解を実現するためには，特定の音に特化しない一般の音の理解（一般音理解）が必要不可欠である．音響イベント同定とは，音響信号の中からイベント（音源）を検出し，その種類や名前といった高次シンボルの抽出（同定）を行う一般音理解を実現するための要素技術であり，近年 ICASSP2013 での Special Session や WASPAA2013 での D-CASE Challenge [2] などで盛んに研究が行われるようになった．音響イベント同定における近年の主要問題は，音声に比べて時間的にも周波数的にも多様な特性を持つ音（反復音，突発音，音楽などの混合音など）に対応しうる特徴量抽出 [3; 4] と識別器設計 [5-22] であるといえる．多様な音に対応した識別器実現のためには，周波数だけでなく時間を陽に考慮した設計が重要となり，これまで，非負値行列因子分解を用いた手法 [7; 8; 20]や，重み付き有限状態トランスデューサを用いた手法 [13]，スペクトログラムを用いた手法 [14; 16]，隠れマルコフモデルを用いた手法 [15]，特徴量ヒストグラムを用いた手法 [19]，音声認識に倣った手法 [3; 21; 22]などが提案されてきた．これらの手法は，ガウス混合モデルを用いた手法 [5] などの周波数特性のみを用いた手法に比べ，時間を扱えるようになったことから，音声（話者同定）だけに限定しない多様な音に対応できるようになった．しかし，これらのパラメトリックモデルを用いた手法は，様々な音に対する最適なモデルを考慮することが難しいという課題がある．短い突発音や長い音楽など，各音源の長さや複雑さはそれぞれ異なるはずであり，パラメトリックモデル構築の際にあらかじめ決めなければならないモデルパラメータ（信号長，フレーム長，状態数，N グラムの次数など）は音の長さや複雑さに合わせて調節可能であることが望ましく，既存法は音源毎に異なるモデルを考慮しうるほど表現能力が十分でない．また，パラメトリックモデルを用いる場合，音環境が学習に用いた環境と適合しない場合は性能が低下してしまう問題があり，雑音や残響などが混入する実環境下の応用において課題を残している．

本稿ではこれらの，1) 音によって異なる長さや複雑さの考慮，2) 音環境と学習環境のミスマッチ問題に取り組み，実環境ロバストな音響イベント同定を実現することを目的とする．

1) に対して，本稿は音声認識に倣った音響イベント同定にノンパラメトリックベイズモデルを導入する．これまでの音声認識に倣った手法[3; 21; 22]は，音の長さをある程度表現することが可能であるが，パラメトリックモデルを用いているため，2) の問題解決が十分になされていない．また，モデルパラメータが固定されているため，音源毎の音の長さや複雑さに合わせた最適なモデルを得ることが難しかった．そこで，本稿はノンパラメトリックベイズ法の一つである Nested Pitman-Yor (NPY) 過程[26]を一般音の音響イベント同定モデル生成のために適用し，

大量の音環境データから，音響イベント同定の統計的モデルを構築する．これにより，任意の長さのセグメント（単語）と N-gram 言語モデルの次数を教師無し学習で推定でき，音の長さや複雑さをモデルに反映することができる（2.3 章）．これまでにもノンパラメトリックベイズモデルを用いた音響イベント同定手法は提案されてきた[18; 19; 20]が，同定に用いる信号の時系列長を固定していたため，音源毎の音の長さを陽に考慮することが困難であった．提案法は時系列長が可変なモデルを用いるため，音の長さをより柔軟に表現することが可能である．

2) に対し，2 つのアプローチを提案する．まず，Latent Dirichlet Allocation (LDA) [24] を一般音に適用し，音響特徴量を符号化する際に必要な音の基本単位（符号）を雑音ロバストとなるように選択することで，実環境下の残響や雑音で生じる音環境と学習環境とのミスマッチを吸収する（2.2 章）．2 つ目に，音の基本単位の相互距離に基づくあいまい検索を導入して音響イベント同定の雑音ロバスト性向上を図る（2.4 章）．

提案手法を学習環境とミスマッチする雑音環境下における音響イベント同定に適用し，その有効性を示す．

2　提案手法

本章では，音響特徴量抽出について触れた後，ノンパラメトリックベイズモデルを用いた手法の詳細について述べる．

2.1　音響特徴量抽出

ある音源からのモノラル信号入力に対する短時間フーリエ変換 $u_\tau(\omega)$ を，目的音 $s_\tau(\omega)$ と雑音 $n_\tau(\omega)$ が混合した以下のモデルとして定義する．

$$u_\tau(\omega) = s_\tau(\omega) + n_\tau(\omega) \tag{1}$$

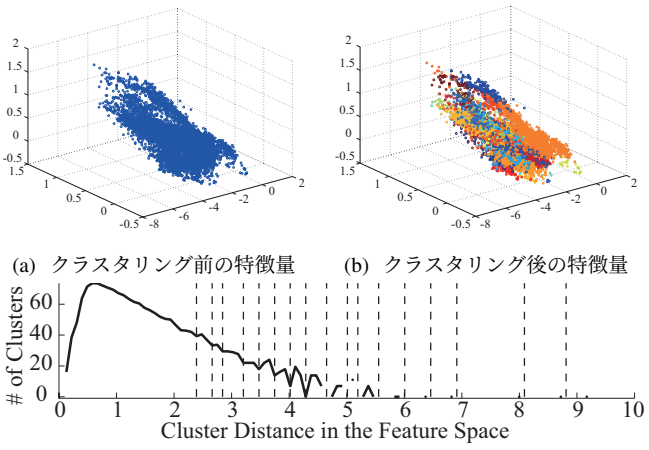
ここで，τ はフレーム番号を表す．$u_\tau(\omega)$ から音響特徴量を抽出し，それを $x_{d\tau}$ と表す．ここで，$u_\tau(\omega)$ には，D 個の音響イベントが含まれるとし，d $(1 \le d \le D)$ はそのインデックスとする．また，d 個目の音響イベントを構成するフレーム総数を N_d とする．ゆえに，$1 \le \tau \le N_d$ に対して，$x_d = \{x_{d1}, \ldots, x_{dN_d}\}$ となる．

2.2　音響イベントの基本単位の推定

音響イベントの基本単位の推定は，音響特徴量のクラスタリングを行う際に，クラスタリングの閾値を LDA によって最適に設定することによって実現する．

2.2.1　凝集型階層クラスタリング

音声認識における音声信号の音素のように，音響イベントにおける一般音の基本単位を「音ユニット」と定義し，音響イベントを音ユニットに分解する．音ユニットを定義するため，音響特徴量の凝集型階層クラスタリング



(c) ヒストグラムと極小点

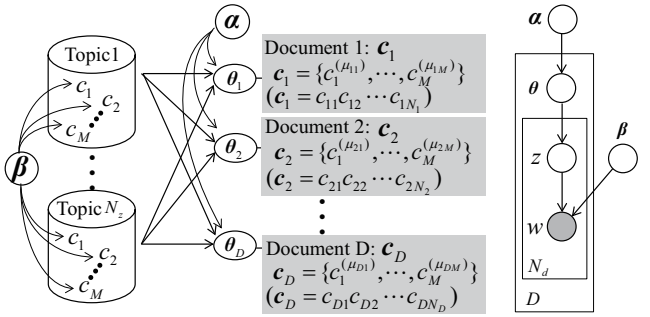
Figure 1: 凝集型階層クラスタリングの例

[23] を行う．ここで，便宜上，$u_\tau(\omega)$ に含まれる音響イベントのフレーム特徴量 $x_{d\tau}$ を時系列に並べ，インデックスを振り直し，x_n $(1 \le n \le N)$ と表記する．凝集型階層クラスタリングを Figure 1(a) のように x_n に対して行う．ここで $N = \sum_{d=1}^{D} N_d$ である．距離関数 Δ とクラスタ間距離は，クラスタ重心のユークリッド距離として定義した．したがって，i 番目と j 番目のクラスタの距離関数は以下で定義される．

$$\Delta_{ij} = \left\| \frac{1}{N_i} \sum_{n \in i} x_n - \frac{1}{N_j} \sum_{n \in j} x_n \right\| , \tag{2}$$

ここで，N_i と N_j はそれぞれ，i 番目と j 番目のクラスタに属するサンプル数である．全ての i と j $(i \neq j)$ の組の中で Δ_{ij} が最小となる組を凝集させることをクラスタ数が十分に小さくなる（1 になる）まで繰り返す．音ユニットは得られた各クラスタに対して定義される．クラスタ間距離に対するクラスタ数のヒストグラムを Figure 1(c) に示す．横軸は Δ_{ij} の距離範囲を 100 に分割したビンを表す．縦軸はそれぞれのビンに対応した距離範囲において統合されたクラスタ数を表す．Figure 1(c) は点線で示された極小を複数持ち，これらの極小となる距離のいずれかで分割したクラスタから音ユニットを構成する．Figure 1(a) と 1(b) にクラスタリング前の特徴量分布と，ヒストグラムが極小となる距離でクラスタリングした特徴量分布を示す．図では，x_n を特異値分解を用いて三次元に次元削減を行ったものをプロットしている．$C = \{C_1, C_2, \ldots, C_M\}$ を得られたクラスタの重心の集合とする．ここで，M はクラスタ数を表す．また，m 番目のクラスタ C_m に対応する音ユニットを c_m と定義する $(1 \le m \le M)$．従って，音ユニットの集合は，$c = \{c_1, c_2, \ldots, c_M\}$ となる．$x_{d\tau}$ に対応した音ユニット（以降，$c_{d\tau}$ と表記する）は，$m = \underset{1 \le m \le M}{\operatorname{argmin}} ||C_m - x_{d\tau}||$ を満たす c_m として決定される．最終的に，d 番目の音響イベントの音ユニット系列は以下のように符号化できる．

$$c_d = c_{d1} c_{d2} \ldots c_{dN_d} , \tag{3}$$



(a) Illustration of LDA (b) Graphical Model

Figure 2: LDA の生成過程とグラフィカルモデル

ここで，c_d と $c_{dn}(1 \le n \le N_d)$ は，次節の LDA における文書と単語とみなして処理を行う．

2.2.2 LDA に基づく音ユニットの選択

雑音ロバストな音ユニットを得るためには，Figure 1(c) が極小となる複数の距離候補から最適なものを選択することが求められる．直感的には，音ユニット同士の距離が近い場合は，音の変化に対する感度が向上するものの，定常音に対して符号化された系列が揺れてしまう場合があり，遠い場合は雑音ロバストな音ユニットが得られるものの，異なる音響イベントを表現しうるほどの感度を満たせない可能性がある．そこで，最適な距離候補の選択に LDA [24] を用いる．LDA は文書モデルの一種で，N_z 個の潜在トピック ($z = \{z_1, z_2, \ldots, z_{N_z}\}$) でコーパス上の文書が表せると仮定した確率的生成モデルである．本稿では，LDA を 式 (3) に適用する．すなわち，コーパスは D 個の文章からなる文章集合 W ($W = \{c_1, c_2, \ldots, c_D\}$) を持ち，$d$ 番目の文書は，音ユニットから定義される N_d 個の単語 ($c_d = c_{d1}c_{d2}\ldots c_{dN_d}$) から構成されるとする．ここで，語彙数は M となり ($c = \{c_1, c_2, \ldots, c_M\}$)，音ユニットの種類数と一致する．LDA のため，単語 c_m $(1 \le m \le M)$ に対して，d 番目の文書中の c_m の個数を上付き文字で $c_m^{(\mu_{dm})}$ と表す．ここで，μ_{dm} は d 番目の文書中に存在する単語 c_m $(1 \le m \le M)$ の数，すなわち $N_d = \sum_{m=1}^{M} \mu_{dm}$ である．LDA では，z を生成するための N_z 次元の確率 ($\theta_d = \{\theta_{d1}, \theta_{d2}, \ldots, \theta_{dN_z}\}$) がディリクレ分布 $\mathrm{Dir}(\theta_d|\alpha)$ に従うと仮定する．W を生成するための確率は以下で表される．

$$p(W|\alpha,\beta) = \prod_{d=1}^{D} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{k=1}^{N_z} p(z_{dk}|\theta_d) p(c_{dn}|z_{dk},\beta) \right) d\theta_d ,$$

ここで，α と β が LDA で推定するパラメータとなる．$\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_{N_z}\}$ は $\mathrm{Dir}(\theta_d|\alpha)$ のためのパラメータであり，$\beta \in \mathbb{R}^{\mathbb{N}_F \times \mathbb{M}}$ は，トピック $z_k(1 \le k \le N_z)$ の中の語彙 $c_m(1 \le m \le M)$ のユニグラムの確率 $p(c_m|z_k)$ である．Figure 2 に生成過程のイメージとグラフィカルモデルを示す．α と β の推定のため，本稿では変分ベイズを用いた．

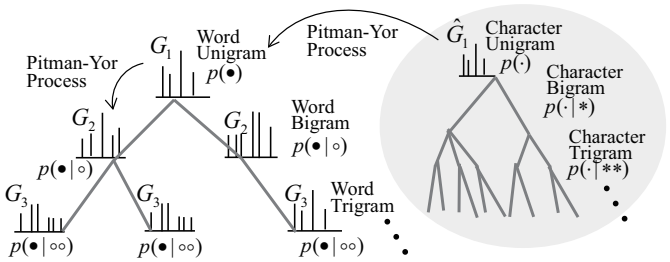


Figure 3: NPY 過程での単語・文字 N グラムモデル

最適な距離の選択のため，Figure 1(c) の N_l 個の極小に対して，それぞれ式 (3) の c_d を算出し，LDA を行う．$\beta = [\beta_1, \ldots, \beta_M]$ を各語彙（音ユニット）に対応した列ベクトルとする．i 番目と j 番目の列ベクトル β_i と β_j の向きが十分に近い場合，それらに相当する音ユニットである c_i と c_j はトピック空間内で同じものを表すと考えられるため，統合する必要がある．従って，N_l 個の候補の中で，これらの列ベクトルの分散が最大となるように距離を選択することで，最適な音ユニットを選択する．具体例を以下に示す．

1) $x_{d\tau}$ の凝集型階層クラスタリングを行う．
2) クラスタ間距離-クラスタ数のグラフを算出し，N_l 個の極小を得る．
3) $1 \le n \le N_l$ に対して以下を繰り返す．
   - n 番目の極小となる距離でクラスタ C を形成する．
   - 全ての d に対して音ユニット系列 c_d を計算する．
   - c_d を用いて LDA の β を推定する．
4) β の列ベクトルの分散が最大となる n を選択する（$\hat{n}$ とする）．
5) $\hat{n}$ 番目の極小に対応する音ユニット $\hat{c} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_M\}$ を用いて音ユニット系列 $\hat{c}_d = \hat{c}_{d1}\hat{c}_{d2}\ldots\hat{c}_{dN_d}$ を求める．

2.3 NPY 過程によるノンパラメトリックな時間統合

自然言語処理では，単語が経験的にわかっているので，音素（音声信号の基本単位）を単語ごとに区切ることができる．一方，非音声を含む一般音の音響イベントでは，そのような分割を経験的な知見から，事前情報として得ることは難しい．事前情報無しに分割を行うため，自然言語の形態素解析に用いられる NPY 過程 [26] を用いる．NPY 過程は，言語を単語 N グラムと文字 N グラムのネスト構造でモデル化する．二つの N グラムモデルの推定には，未知語に対してロバストな Hierarchical Pitman-Yor (HPY) 過程 [25] を用いる．HPY 過程はディリクレ過程の階層的な拡張である．単語 N グラムモデルでは，単語列 $h = w_{t-n}, \ldots, w_{t-1}$ の次の単語 w の生成確率が以下で表される．

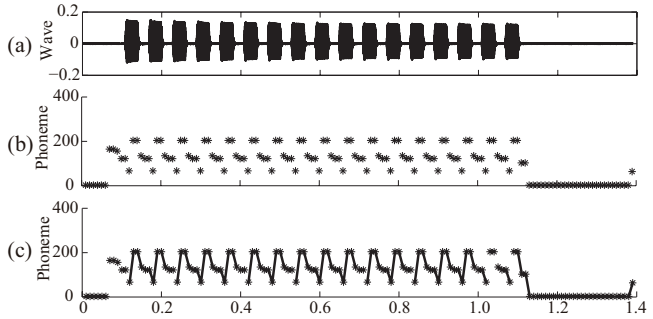


Fig.4.1. Repetitive sound : Phone ring

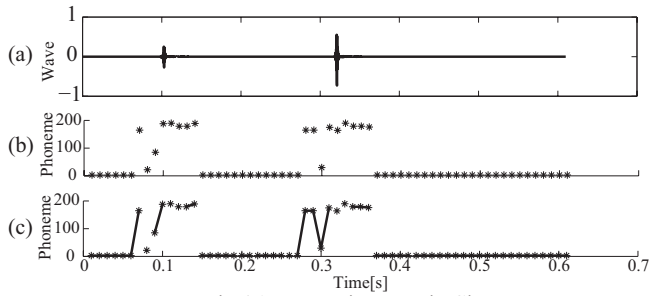


Fig.4.2. Percussive sound : Clap

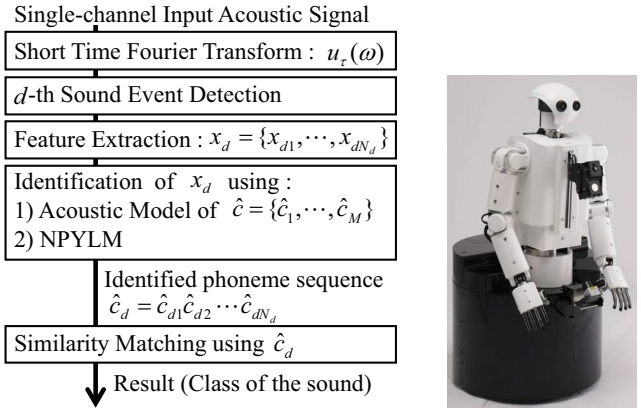
Figure 4: NPY 過程の例: (a) 音響信号, (b) 分割前音ユニット系列, (c) 分割後音ユニット系列

$$p(w|h) = \frac{\gamma(w|h) - \eta t_{hw}}{\xi + \gamma(h)} + \frac{\xi + \eta t_{hw}}{\xi + \gamma(h)} p(w|h') \,, \qquad (4)$$

ここで，$h' = w_{t-n-1}, \ldots, w_{t-1}$ は h から前単語 w_{t-n} を除いた文字列である．ゆえに，$p(w|h')$ は $(n-1)$ グラムの系列 h' から w を生成する確率となる．$\gamma(w|h)$ は h の w の数であり，$\gamma(h) = \sum_w \gamma(w|h)$ である．t_{hw} は h' から生成された w の数であり，$t_h = \sum_w t_{hw}$ である．η と ξ は HPY 過程のパラメータであり，Gibbs sampling を用いて推定する．

HPY 過程は次数を一つ減らした $p(w|h')$ を $p(w|h)$ の推定に用いるため，単語ユニグラムの指定に使用する基底測度 $p(w|h')$ を事前知識として与えられた語彙から得る．しかし，セグメントの情報が与えられていない場合は基底測度を得ることができない．そこで NPY 過程では文字 N グラムを階層的に設け，基底測度を HPY 過程を用いて推定することでこの問題を解決する．本稿における NPY 過程は $\hat{c}_d$ を用いて文字 N グラムと単語 N グラムを Figure 3 のように同時に推定する．ここで，2.2.2 節の LDA では各音ユニットが単語として定義されたが，NPY 過程における単語 w は音ユニット系列として定義される．

Figure 4 に電話の着信音（定常的な反復音:Figure 4.1(a)）と拍手（突発音:Figure 4.2(a)）から得られた音ユニット系列と単語列を示す．横軸と縦軸はそれぞれ時間フレーム τ と音ユニット番号 $\hat{c}_{dN_d}$ を表す．Figure 4.1(b) と Figure 4.2(b) に分割前の音ユニット系列を点線で，Figure 4.1(c) と Figure 4.2(c) に分割後の単語系列を実線で示す．図より，電話の着信音は繰り返しを単位として同じ単語に，拍手は突発音の前後で一単語として区切られたことがわかる．NPY 過程で得られた単語を用いることで，音響信号の時間情報を考慮した分割を実現することができる．



(a) 処理の流れ　　(b) ロボット

Figure 5: 処理の流れとハードウェア

2.4 あいまい検索による雑音補償

NPY 過程を一般音に適用することによる問題は，もともと符号化された自然言語を対象とするため，雑音が混入したことによる入力系列の揺れを扱うことができないことである．こうした揺れを補償するため，本稿は音ユニットの相互距離に基づくあいまい検索を導入する．

前節の凝集型階層クラスタリングから得られた C に対して，クラスタ重心の相互距離（ユークリッド距離）を計算し，あるクラスタから N_Δ 番目に近いクラスタまでを同じクラスタ，つまり同じ音ユニットとみなして処理を行う．例えば，x_d が $\hat{c}_d = c_1c_2c_3$ と推定されたとする．$N_\Delta = 1$ で，(c_1, c_2) と (c_3, c_4) の組が近いと判定された場合，あいまい検索では，これらの音ユニットを入れ替えた $c_1c_1c_3$, $c_2c_1c_3$, $c_1c_2c_3$, $c_2c_2c_3$, $c_1c_1c_4$, $c_2c_1c_4$, $c_1c_2c_4$, $c_2c_2c_4$ も $\hat{c}_d$ として用いる．従って，長さ N_d の $\hat{c}_d$ を同定する場合は，$N_d^{N_\Delta+1}$ 個の候補系列が検索対象となる．

3 評価実験

提案手法と既存手法の音響イベント同定の性能比較を行う．Figure 5(a) に音響イベント同定の処理の流れを示す．評価では Figure 5(b) のロボットの頭部額の位置に設置されたマイクを用いて，音響イベントを 1m の距離で発生させて録音を行った．ロボットは残響時間 0.2 秒の部屋の中央に配置した．音響信号は 16bit，16kHz でサンプリングした．音響特徴量抽出のフレーム長とシフト長は 512, 160 とした．同定に用いるモデルの学習と評価は，Table 1 に示されるデータセットを用いて 5 分割交差検定を行った．手法の雑音ロバスト性を評価するため，学習は雑音の無い環境のデータのみを用いて行った．評価では，実環境下の同定性能評価のため，音声認識で一般的に用いられるバブル雑音とロボット背面のファンの雑音を用い，これらの雑音の SN 比を変化させて行った．SN 比は SNR[dB] = $20\log_{10}(\pi_s/1-\pi_s)$ $(0 \le \pi_s \le 1)$ と定義した．

ここで，音響イベント信号 $s_\tau(\omega)$ と雑音 $n_\tau(\omega)$ は

Table 1: 学習・評価用データセット

Speech	Dataset : ATR dataset (216 words by 5 male and 5 female speakers) # of cls : 2 (male and female)
Music	Dataset : RWC-MDB-G [28] (32 genres of music for approx. 5 minutes) # of cls : 32 (ex. popular, ballad, etc.)
Environment	Dataset : RWCP [29] (92 kinds of sounds for approx. 4 minutes) # of cls : 92 (ex. phone ring, clap, etc.)

$u_\tau(\omega) = \pi_s s_\tau(\omega) + (1-\pi_s) n_\tau(\omega)$ となるよう混合した．

π_s={1,0.95,0.9,0.85,0.8,0.7,0.5,0.3} とし，SNR[dB] = $\{\infty, 12.8, 9.5, 7.5, 6.0, 3.7, 0.0, -3.7\}$ となった．

音響特徴量には 41 次元の Mel Scale Log Spectrum(MSLS) [27]を用いた (13 次元 MSLS + Δ + Δ^2 + ΔE + Δ^2E).

3.1 LDA に基づく音ユニットの雑音ロバスト性

提案法による音ユニットの雑音ロバスト性を評価するため，音ユニットを用いて学習した GMM(GMM-D) と，手動ラベルを用いて学習した GMM (GMM-S[5]) の同定性能を比較した．GMM は雑音の無いデータを用いて学習した混合数 16 のものを用いた．

手動ラベル (GMM-S) では，x_d の全区間において $\hat{c}_d = c_d c_d \ldots c_d$ と一様にラベル付けを行い，音ユニットの種類数は Table 1 のデータベースの正解クラス数と同じ 126(=2+32+92) とした．一方 GMM-D では，音ユニットは LDA によって抽出されるため，音ユニットの種類数は自動的に決められ，本評価では 96 種類となった．また，ラベル付けは自動で行われる．評価には各音響イベントの平均フレーム正解率 (Frame Correct Rate (FCR)) を用いた．

Figure 6 の GMM-S と GMM-D を比較すると，SNR = ∞ に対しては GMM-S が GMM-D より高い性能を示した．GMM-D では音ユニットが自動的に決められるため学習データに対して過剰適合してしまったと考えられる．クリーン環境における性能向上は今後の課題である．一方，SNR ≠ ∞ では，GMM-D が GMM-S より 3-9 pts 性能が高いという結果が得られた．このことから提案法によって自動的に抽出された音ユニットの雑音ロバスト性を確認できた．

3.2 NPY 過程に基づく言語モデルとあいまい検索

最後に，2.2 節で得られた音ユニット系列を用いて，NPY 過程によって得られた言語モデルを音響イベント同定に適用した．言語モデルの語彙数は 351 となった．言語モデルの有効性検証のため，前節の GMM から，モノフォン HMM の音響モデルへ拡張した．ここで状態数と混合数はそれぞれ 1，16 とし，提案する音ユニットに基づいて学習を行った．まず，音響モデルを拡張したことによる影響を確認するため，Figure 6 の MONO-D にユニグラム

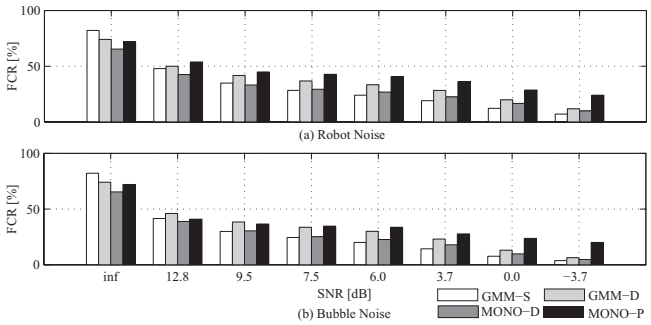


Figure 6: 平均フレーム正解率による同定結果

の言語モデルを用いた時の結果を示した．モノフォンモデルは GMM に比べて次状態の音ユニットへの遷移を抑制するため，GMM-D より性能が劣化したと考えられる．

次に，MONO-D と同じ音響モデルを用いて，ユニグラム言語モデルと後段処理に NPY 過程とあいまい検索を適用した時の結果を Figure 6 に示す．3.1 節で議論した GMM-D と同じ問題が影響し，クリーンな環境においては，MONO-P は GMM-S ほど高い性能を示さなかった．一方，クリーン環境以外の全ての場合において，MONO-P は GMM-S や GMM-D に比べて，それぞれ 5-18 pts, 5-13 pts 高い性能を示しており，雑音ロバスト性の向上を確認できた．実環境での音響イベント同定では，未知の雑音環境での動作が求められるため，未知雑音へのロバスト性が高いことは，実環境への適用性が高いと言える．

4 結論

本稿は雑音存在下の一般音の音響イベント同定について述べた．一般音の音響イベント同定を実環境応用する際に問題となる，1) 音によって異なる長さや複雑さの考慮と，2) 音環境と学習環境のミスマッチ問題に取り組んだ．1) に対し，音声認識に倣った音響イベント同定に対して，NPY 過程に基づくノンパラメトリックベイズモデルを導入した．2) に対し，LDA に基づく雑音ロバストな音ユニットの抽出と，音ユニットの相互距離に基づくあいまい検索を提案した．実環境収録データで評価実験を行った結果，一般的な GMM と比較し，学習環境とミスマッチする雑音環境下で 5-18 pts の性能向上が得られ，提案法の有効性を示すことができた．

参考文献

[1] D. Rosenthal and H. G. Okuno, “Computational Auditory Scene Analysis”, Lawrence Erlbaum Associates, Mahwah, New Jersey, pp. 399+xiii, 1998.

[2] D. Giannoulis *et al.*, “Detection and classification of acoustic scenes and events: an IEEE AASP challenge”, in *IEEE WASPAA*, 2013.

[3] X. Zhuang *et al.*, "Real-world acoustic event detection", *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[4] L. Ballan *et al.*, "Deep networks for audio event classification in soccer videos", in *ICME*, pp. 474–477, 2009.

[5] D. A. Reynolds *et al.*, "Robust text-independent speaker identification using gaussian mixture speaker models", *IEEE TSAP*, vol. 3, no. 1, pp. 72–83, 1995.

[6] K. Nakamura *et al.*, "Intelligent Sound Source Localization and Its Application to Multimodal Human Tracking", in *Proc. of IEEE/RAS IROS*, pp. 143-148, 2011.

[7] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection", in *Proc. of IEEE WASPAA*, pp. 69–72, 2011.

[8] M. L. Chin *et al.*, "Audio event detection based on layered symbolic sequence representations" in *Proc. of ICASSP*, pp. 1953–1956, 2012.

[9] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments", *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[10] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection", in *Proc. of 19th EUSIPCO*, pp. 1307–1311, 2011.

[11] B. Schauere *et al.*, " "Wow!" Bayesian surprise for salient acoustic event detection", in *Proc. of ICASSP*, pp. 6402–6406, 2013.

[12] K. H. Lin *et al.*, "Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization", in *Proc. of ICASSP*, pp. 2277–2280, 2012.

[13] D. Rybach *et al.*, "Silence is golden: Modeling non-speech events in WFST-based dynamic network decoders", in *Proc. of ICASSP*, pp. 4205–4208, 2012.

[14] Y. Sasaki *et al.*, "Daily sound recognition using Pitch-Cluster-Maps for mobile robot audition", in *Proc. of IEEE/RAS IROS*, pp. 2724-2729, 2009.

[15] V. Ramasubramanian *et al.*, "Continuous audio analytics by HMM and Viterbi decoding", in *Proc. of ICASSP*, pp. 2396–2399, 2011.

[16] C. Bauge *et al.*, "Representing environmental sounds using the separable scattering transform", in *Proc. of ICASSP*, pp. 8667–8671, 2013.

[17] M. Espi *et al.*, "A tandem connectionist model using combination of multi-scale spectro-temporal features for acoustic event detection", in *Proc. of ICASSP*, pp. 4293–4296, 2013.

[18] Y. Sasaki *et al.*, "Nested Infinite Gaussian Mixture Model for Environmental Audio Signal Recognition", in Proc. of SIG-Challenge 2012, B202-07.

[19] T. Nakamura, T., T. Nagai, and N. Iwahashi, "Multimodal categorization by hierarchical dirichlet process", in *Proc. of IEEE/RAS IROS*, pp. 1520–1525, 2011.

[20] Y. Ohishi *et al.*, "Bayesian Semi-supervised Audio Event Transcription based on Markov Indian buffet Process", in *Proc. of ICASSP*, pp. 3163–3167, 2013.

[21] S. Chaudhuri, M.Harvilla, and B. Raj, "Unsupervised Learning of Acoustic Unit Descriptors for Audio Content Representation and Classification", in *Proc. of INTERSPEECH*, pp. 2265-2268, 2011.

[22] A. Kumar *et al.*, "Audio event detection from acoustic unit occurrence patterns", in *Proc. of ICASSP*, pp. 489–492, 2012.

[23] W. H. Press *et al.*, *Numerical Recipes in C: the Art of Scientific Computing*, 2nd ed., Cambridge University Press, 1998.

[24] D. M. Blei *et al.*, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[25] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes", in *Proc. of ICCL and ACL*, vol. 44, pp. 985–992, 2006.

[26] D. Mochiahshi *et al.*, "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling", in *Proc. of the Joint Conf. of ACL and AFNLP*, vol. 1, pp. 100–108, 2009.

[27] Y. Nishimura *et al.*, "Noise-robust speech recognition using multiband spectral features", in *Proc. 148th Acoustical Soc. of America Meet.*, San Diego, CA, no. 1aSC7, 2004.

[28] M. Goto, "Development of the RWC Music Database", in *Proc. of ICA*, pp. 553–556, Apr. 2004.

[29] S. Nakamura *et al.*, "Sound Scene Database in Real Acoustic Environments", in *Proc. of Oriental COCOSDA Workshop*, pp. 103–106, 1998.

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-B302-03

Semi-Blind Infinite NMF を用いた動作雑音抑圧手法の提案とその評価

Semi-Blind Infinite Non-negative Matrix Factorization for Ego-motion Noise Suppression

◯手塚太貴 [1] 吉田尚水 [1*] 中臺一博 [1,2]

Taiki TEZUKA　Takami YOSHIDA　Kazuhiro NAKADAI

1 東京工業大学　情報理工学研究科, 2 (株) ホンダ・リサーチ・インスティチュート・ジャパン

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

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

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

Abstract

ロボット聴覚機能の実現における課題としてロボットの動作雑音が挙げられる. 本論文では, ノンパラメトリックベイジアンモデルの一種である Semi-Blind Infinite Non-negative Matrix Factorization (SB-INMF) を提案し, これを用いた新たな動作雑音抑圧手法を報告する．SB-INMF は動作雑音抑圧に関節角の情報が不要，動作雑音と目的音を線形分離するため，スペクトル減算よりも歪みが少ないといった特長がある．二種類の実ロボットを用いて評価実験を行った結果, 関節角情報を用いた従来の雑音テンプレートによる動作雑音抑圧よりも良好な雑音抑圧結果を得ることができた.

1　はじめに

ロボットの聴覚機能，つまり「ロボット聴覚」[1] は人とロボットの音声コミュニケーションを実現する上で重要な技術であり，2000 年に提唱されて以降精力的に研究が行われている．この研究の狙いはロボットに取り付けたマイクロホンを用いて聴覚機能を実現することである．これまでにもバイノーラル [2–7], マイクアレイ [8–11], マルチモーダル [2, 12] そしてユビキタスセンサ [13–15] など様々なアプローチによるロボット聴覚システムが報告されている．ロボット聴覚の重要な要素に音源定位や音源分離，音声認識技術が挙げられるが，ロボットの動作雑音はこれらの実現における大きな課題である．ロボットは荷物の運搬やダンスなど様々な動作中であっても，人間と音声を介したコミュニケーションが可能でなければならない．しかし，ロボットの動作中は必ず動作雑音が発生し，ロボットの聴覚機能を妨げてしまう．そのため，動作雑音を抑圧することの重要性が論じられるとともにこれまでに様々な手法が提案されてきた.

動作雑音抑圧手法には大きく二種類のアプローチがある．１つ目はセンサを用いて動作雑音を測定するアプローチ，２つ目は関節角のような動作雑音と相関のある情報から動作雑音を推定するアプローチである．１つ目のアプローチとして，中臺らは，ロボット外装の内外にマイクロホンを設置することで，ロボットの内外を区別する音響的身体性を構築し，内部雑音が小さい時間のみ音源定位を行う手法を報告している [1]．しかし, 音源分離のように，内部雑音の有無にかかわらず常に処理が必要な場合には対応が難しい．その他には，ロボットの機体内部に内部雑音の検知用のセンサを取り付け，*Frequency-Domain Blind Signal Separation(FD-BSS)* を適用することで動作雑音を推定する手法が提案されている [16]．しかしこれらの手法は音声収録用のマイクロホン以外のセンサを必要とし，性能の面で重要となるセンサの配置について議論されていない．そして，これらのアプローチでは動作雑音を推定するために追加したマイクロホンやセンサが必要となるが，これによりシステムが複雑になり，結果として計算コストが大きくなってしまうという問題がある.

２つ目のアプローチは，関節駆動により発生する動作雑音と関節角に強い相関関係があるという事実に基づいている．例えば, 伊藤らは， Sony AIBO を用いて，関節角度や位置を入力として動作雑音を推定するニューラルネットワークを構築し，推定雑音をスペクトル減算 [18] することにより，音声認識性能が向上できることを報告している [17]．しかし，シミュレーション実験しか行っておらず，残響など実環境ならではのファクターが加わった場合の有効性は不明である．また，西村らは，動作コマンドに対応した雑音テンプレートを構築し，これをスペクトル減算をする手法を提案した．さらに, スペクトル減算によって生じる歪みに対応するため，ミッシングフィーチャ理論を用いて音声認識性能を向上させた [3]．しかし，この手法は

* 現在 株式会社東芝　研究開発センター　知識メディアラボラトリー勤務

Table 1: 変数表記

意味	表記
マイクロホンの数	$N_{mic} \in \mathbb{N}$
周波数ビン数	$N_f \in \mathbb{N}$
因子数	$N_k \in \mathbb{N}$
基底数	$K \in \mathbb{N}$
観測信号のサイズ	$N_s \in \mathbb{N}$
観測信号のパワースペクトル	$Y \in \mathbb{R}_{+0}^{N_f+1}$
動作雑音のパワースペクトル	$X \in \mathbb{R}_{+0}^{N_f+1}$
基底	$F \in \mathbb{R}_{+0}^{N_f \times N_k}$
アクティベーション	$Z \in \mathbb{R}_{+0}^{N_s \times N_f}$
ゲイン	$\theta \in \mathbb{R}_{+0}^{N_k}$

$\mathbb{R}_{+0}$ は正の実数, $\mathbb{N}$ は自然数を表す.

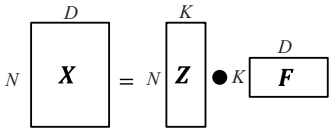


Figure 1: 因子モデル

動作コマンド単位の手法であるため，複数の動作を組み合わせた複雑な動作への対応は難しい．Ince らは，雑音テンプレートを処理フレーム単位で構築する手法を考案し，音源定位，分離，音声認識それぞれに提案手法が有効性であることを報告している [19]．しかし，この手法は動作が同じであれば, 必ず同じ動作雑音が発生すると仮定しており，実環境においてこの仮定が必ずしも成り立つとは言えない．このため, 動作雑音の推定に誤りが生じ，スペクトル減算によって音声認識性能が悪化する原因となる．そこで，本稿ではノンパラメトリックベイジアンモデルである *Semi-Blind Infinite Non-negative Matrix Factorization (SB-INMF)* を提案し，関節角など他のセンサの情報を使わない新たな動作雑音抑圧手法の実現を試みる．本研究で提案する SB-INMF では収録した音信号から直接動作雑音の推定を行うため前述のテンプレート法で発生する推定誤差は生じない．また，線形過程により動作雑音と目的音を分離することが可能であり，スペクトル減算を用いる手法よりも歪みが少ない動作雑音抑圧が可能である．

2 SB-INMF による動作雑音抑圧

表 1 に本論文で用いる記号の定義を示す．動作雑音は主に関節が駆動することで発生する．本稿では，動作雑音は各関節からの動作雑音を組み合わせて表現できるものと考える．このような場合，図 1 の様な因子モデル (LFM) を用いて動作雑音を表現することができる．ここで，基底 $\boldsymbol{F}$ は動作雑音の周波数方向の特徴を表し, アクティベーション $\boldsymbol{Z}$ は動作雑音の時間方向の特徴を表す.

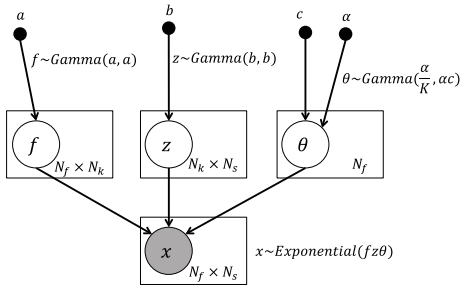


Figure 2: INMF のグラフィカルモデル

具体的には LFM として，モノラル信号の音源分離手法でよく用いられる非負値行列因子分解 (NMF) を用いる．各々の関節から生じる動作雑音パワースペクトルは非負値，かつ加法性であると見なし，収録信号に含まれる動作雑音と目的音を NMF を用いて線形分離する．NMF による音源の分離を行う際, 観測信号の確率分布として一般に以下の指数分布が用いられる [20].

$$x_{nd} \sim \text{Exponential}\left(\sum_k z_{nk} f_{kd}\right), \tag{1}$$

ここで $x_{nd} \geq 0, z_{nk} \geq 0, f_{kd} \geq 0$ は各々行列 $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{F}$ の要素である．NMF による動作雑音の因子モデル推定を行う際に，次に示す問題点を考慮する必要がある．①動作雑音と関節角の情報との間に完全な相関関係は成り立たない．②NMF によって推定された動作雑音の基底の数は関節の数と必ずしも一致しない，すなわち, 適切な基底の数は未知である．これらの問題を扱うために本研究では NMF による因子モデルを拡張した無限因子モデル (ILFM) を導入する．つまり，NMF を無限因子モデルが扱えるモデルように, *Infinite Non-negative Matrix Factorization(IMNF)* [21] に拡張する．これにより因子モデルが基底の候補数を上限なく保有する事を可能とし，かつ目的とするデータを表現するのに最適な基底の数を機械的に推定することが可能になる.

本論文ではノンパラメトリック確率過程の一つであり，スパースな学習を可能とするガンマ過程を用いて無限因子モデルを構築する．動作雑音の INMF を以下の様に定式化する.

$$f_{kd} \sim \text{Gamma}(a, a), \tag{2}$$
$$z_{nk} \sim \text{Gamma}(b, b), \tag{3}$$
$$\theta_k \sim \text{Gamma}(\alpha/K, \alpha c), \tag{4}$$
$$x_{nd} \sim \text{Exponential}\left(\sum_k \theta_k z_{nk} f_{kd}\right) \tag{5}$$

ここで，a, b, c はガンマ分布のパラメータである．このモデルは各基底に対応した非負値のゲイン θ_k を導入し，不必要な基底のゲイン値を 0 にすることで，最適な基底のみを用いた推定が可能である．なお，このモデルでは非負値行列の各要素の分布がガンマ過程に基づくものであると仮定している．そのため，このモデルは *Gamma Process Non-negative Matrix Factrization(GaP-NMF)* [22] と呼ばれる．

図 2 に式 (2)–(5) に対応したグラフィカルモデルを示す．白と灰色の円はそれぞれ隠れ変数と観測変数を表す．黒い点は与えられたパラメータを表し，複数ノードがある場合はプレートを用いて表す．x が観測信号とした動作雑音に対応する．

このモデルでは, 動作雑音を入力すれば INMF によって，動作雑音の因子モデルを得る事ができる．しかし, 実際の入力信号には音声を始めとする目的音が含まれているため，INMF を適用すると動作雑音と目的音の基底が得られる．このため得られた基底のうち, どの基底が動作雑音に対応し，どの基底が目的音に対応するのかを判別することが難しい．この問題を解決するために *Semi-Blind INMF(SB-INMF)* を提案する．

$$x_{nd} \sim \text{Exponential}\left(\sum_k \theta_k z_{nk} f_{kd} + \sum_l \tilde{\theta}_l \tilde{z}_{nl} \tilde{f}_{ld}\right), \quad (6)$$

ここで θ_k，z_{nk}，および f_{kd} は動作雑音に対応し，$\tilde{\theta}_l$，$\tilde{z}_{nl}$，および $\tilde{f}_{ld}$ は目的音に対応する．動作雑音の基底 f_{kd} を予め与えることで，動作雑音のアクティベーション z_{mk}，ゲイン θ_k，及び目的音の因子モデル $(\tilde{\theta}_l,\ \tilde{z}_{nl},\ \tilde{f}_{ld})$ を推定する．これにより，動作雑音と音声が混合した信号入力に対しても，動作雑音が抑圧された音声を得ることが可能になる．この手法による動作雑音抑圧はスペクトル減算のような非線形過程を踏まない雑音抑圧のため，目的音の歪みが少ない．INMF のパラメータ推定には GaP-NMF と同様の変分ベイズを用いる [22].

$$\begin{aligned}
q(z, f, \theta) &\approx q(z)q(f)q(\theta) && (7)\\
q(z) &= \text{GIG}(z; \gamma^{(z)}, \rho^{(z)}, \tau^{(z)}) && (8)\\
q(f) &= \text{GIG}(f; \gamma^{(f)}, \rho^{(f)}, \tau^{(f)}) && (9)\\
q(\theta) &= \text{GIG}(\theta; \gamma^{(\theta)}, \rho^{(\theta)}, \tau^{(\theta)}) && (10)\\
\text{GIG}(y; \gamma, \rho, \tau) &= \left(\frac{\rho}{2\tau}\right)^{\gamma/2} \frac{\exp\{(\gamma-1)\log y - \rho y - \tau/y\}}{\mathcal{K}_\gamma(2\sqrt{\rho\tau})} && (11)
\end{aligned}$$

GIG は一般化逆ガウス分布であり，そして γ, ρ, 及び τ は GIG のパラメータである．$\mathcal{K}_\gamma$ は第二ベッセル関数である．なお z，f，θ，γ，ρ 及び τ の添字は簡単化のため省略した．

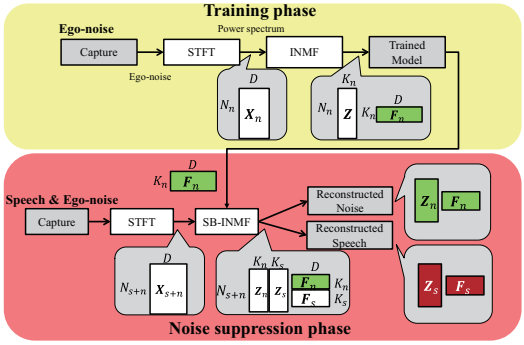


Figure 3: 動作雑音の推定及び抑圧システム

3 動作雑音抑圧システム

前節にて提案した SB-INMF に基づく動作雑音抑圧システムを図 3 に示す．図 3 の上部は INMF に基づく動作雑音の学習過程を表す．マイクロホンで収録した動作雑音に短時間フーリエ変換 (STFT) を行う．得られた動作雑音信号のスペクトル $\boldsymbol{X}_n(N_s \times N_f)$ に対して INMF を行い，N_{k1} 個の動作雑音の基底 $\boldsymbol{F}_n$ を得る．また，SB-INMF による動作雑音推定及び抑圧の過程を表す．目的音と動作雑音を含む信号入力に対し，STFT を行うことでスペクトル $\boldsymbol{X}_{s+n}$ を得る．次に，事前に学習した動作雑音のモデル $\boldsymbol{F}_n$ を既知の基底として SB-INMF を行う事で $\boldsymbol{X}_{s+n}$ に含まれた目的音の基底 $\boldsymbol{F}_s$ 及びアクティベーション $\boldsymbol{Z}_s$ を得る．最終的に $\boldsymbol{F}_s$，$\boldsymbol{Z}_s$ より目的音のスペクトルを得る．また，動作雑音も同様に，与えた基底 $\boldsymbol{F}_n$ 及びそのアクティベーション $\boldsymbol{Z}_n$ によって得ることができる．

4 評価実験

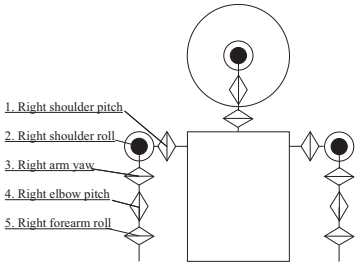
4.1 音響データ収録

提案手法の評価を行うために 2 つのヒューマノイド・ロボットを用いて音響データの収録を行った．

4.1.1 ヒューマノイドロボット

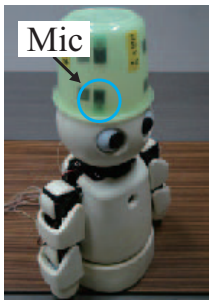
実験には Hearbo と Robovie-W の 2 つのヒューマノイド・ロボットを使用した．Hearbo は頭部に 8 個のマイクロホンが取り付けてあるが，収録にはそのうちの額にあるマイクロホン 1 つを用いた．また，Hearbo は計 34 自由度を持っているが，本研究の実験ではそのうち右腕の 5 つの関節: 右肩ピッチ角 (J1)，右肩ロール角 (J2)，右腕ヨー角 (J3)，右肘ピッチ角 (J4)，右前腕ロール角 (J5) を駆動させた．Robovie-W は市販されている小型ロボットであり, 頭部には音声収録用に 8 個のマイクロホンをつけた帽子が取り付けてあるが，実験ではそのうち 1 つを使用した．Robovie-W は計 17 自由度を持っているが, 本実験ではそのうち 5 つの関節: 腰ヨー角, 肩ロール角 (左右), 肘ピッチ角 (左右) を駆動させた．Robovie-W は各関節に指令値を送ることで制御できるが, 機体から関節角の状態

a) 写真

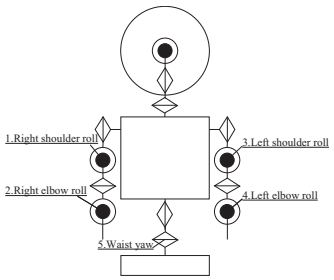


b) ジョイントとリンクの構成

Figure 4: ヒューマノイド・ロボット Hearbo



a) 写真



b) ジョイントとリンクの構成

Figure 5: ヒューマノイド・ロボット Robovie-W

を読み取ることは不可能である．従って, このロボットでは動作情報から動作雑音を推定する動作雑音抑圧手法を適用することが不可能である．

4.1.2 収録条件

Hearbo を用いた実験では以下の条件 D1-D6 の動作雑音及び音声データを収録した．音声及び動作雑音の収録は 4m×7m×3m の部屋で行った．Hearbo は部屋の中心に, 発話者 (スピーカ) は Hearbo の正面から 1.2m 離れた位置に配置した．

D1 J1 から J5 までの 5 個の関節をそれぞれ 20 秒ずつ駆動させ, 動作雑音を発生させた．

D2 J1 を 40 秒間駆動させ, 後半の 20 秒間にスピーカーから音声を流した．

D3 2 つの関節 (J1 と J2) を 40 秒間駆動させ, 後半の 20 秒間にスピーカーで音声を流した．

D4 3 つの関節 (J1,J2,J3) を 40 秒間駆動させ, 後半の 20 秒間にスピーカーで音声を流した．

D5 4 つの関節 (J1 – J4) を 40 秒間駆動させ, 後半の 20 秒間にスピーカーで音声を流した．

D6 5 つの関節 (J1 – J5) を 40 秒間駆動させ, 後半の 20 秒間にスピーカーで音声を流した．

D7 Hearbo を動かさずに 20 秒間スピーカーから音声を流した．

また,Robovie-W を用いた実験では以下の条件 R1,R2 の動作雑音及び発話を収録した．音声及び動作雑音の収録は 10m×10m×3m の部屋で行った．Robovie-W は円形のテーブル載せ，発話者 (人間) は Robovie-W の正面から 2.0m 離れた位置に配置した．

R1 5 つの関節を 5 秒間ランダムに動かし, 動作雑音を発生させた

R2 R1 の動作を行った状態で 2 秒間発話を行う

4.2 評価方法と結果

提案手法を以下の 4 つの実験により評価を行った．

1. 提案手法のパラメータ評価実験
2. 動作雑音推定・抑圧実験

まず実験 1 では SB-INMF に使用されているパラメータの変化が動作雑音抑圧結果に対しどのような影響を与えるか評価し，得られた結果から実験 2 で使用するパラメータを決定する．実験 2 では次の 3 段階に分けて実験を行う．①INMF による動作雑音の再構成実験，②動作雑音と音声の合成音を対象とした動作雑音抑圧実験，③実収録した動作雑音を含む音声を対象とした動作雑音抑圧実験．なお, 実験 2 ～ 4 では提案手法との比較に動作雑音性能が高いことが報告されている Ince らが提案したテンプレートベースの雑音抑圧手法 [19] を用いた．

4.2.1 提案手法のパラメータ評価実験

実験 1：式に示した基底のパラメータ a とアクティベーションのパラメータ b を 10^{-4}～10^{0} の間で変化させながら動作雑音抑圧を行う．Hearbo と Robovie-W について実験を行い，学習データには各々D1 と R1，テストデータには D1 と D7 を足し合わせ音響信号と R1 にクリーンな音声を足し合わせた音響信号を使用した．そして得られた結果を *Signal-to-Inference Ratio(SIR)*, *Signal-to-Distortion Ratio(SDR)*, そして *Signal-to-Noise Ratio(SNR)* により評価を行った．

SNR(SNR$_1$) を以下のように定義する．

$$\mathrm{SNR}_1 = 20\log_{10}\left(\frac{\sum_f\sum_\omega |X(\omega,f)|^2}{\sum_f\sum_\omega ||Y(\omega,f)| - |X(\omega,f)| - |\hat{N}(\omega,f)||^2}\right) \tag{12}$$

ここで ω と f はそれぞれ周波数ビンと時刻フレームを表し，X,Y は雑音を含まない音声及び動作雑音と音声の合成音のスペクトルを表す．また,SIR は目的音以外の妨害音 (動作雑音，環境雑音) による歪みを評価する指標，SDR は目的音の線形歪み，非線形歪み，そして上記の妨害音による歪みを総合的に評価する指標となっている．推定した音声信号 $\hat{x}(t)$ が式 (13) の様に分解できると仮定する．

$$\hat{x}(t) = x_{true}(t) + e_{noise}(t) + e_{artif}(t) \tag{13}$$

ここで $x_{true}(t)$ を真の音声成分，$e_{noise}(t)$ を動作雑音の成分，$e_{artif}(t)$ をいずれにも寄与しない成分とする．この時 SIR，SDR は式 (14),(15) で与えられる．

$$\mathrm{SIR} = 10\log_{10}(\frac{||x_{true}(t)||^2}{||e_{noise}(t)||^2}), \tag{14}$$

$$\mathrm{SDR} = 10\log_{10}\left(\frac{||x_{true}(t)||^2}{||e_{noise}(t)+e_{artif}(t)||^2}\right) \quad (15)$$

なお，SIR と SDR の計算は MATLAB のツールボックス”BSS Eval[1] ”を使用した．

結果： 得られた結果を図 6,7 に示す. 縦軸, 横軸ともにログスケールとなっている．Robovie の結果を見ると，SNR, SIR，SDR 全て共通して基底のパラメータによる影響の大きいことがわかる．しかし，各指標のカラーマップを比較すると SIR と SDR の分布は近い傾向にあるのに対し，SNR は全く異なる分布であることがわかる．また，Hearbo の結果を見ると SNR と SIR には Robovie-W と同じような傾向が見られるのに対し，SDR は異なりほぼ均一な値となってしまっていることがわかる．これは D7 には音声の他に環境雑音含まれており，推定した音声では環境雑音が抑圧されている (後の実験 3 参照) ため，SIR の値が高くても SDR の値が向上しなかったと考えられる．これらの結果を踏まえ，実験 2 〜 4 で用いるパラメータは SNR, SIR, SDR の各々の値がある程度補償されるものを使用した．具体的には Robovie の場合は $a = 10^{-2}$,$b = 10^{-2}$, Hearbo の場合は $a = 10^{-1.5}$,$b = 10^{-1.5}$ とした．

4.2.2 動作雑音推定・抑圧実験

Hearbo を用いた実験では D1 を用いて INMF による動作雑音の基底の学習を行った．また, 比較手法の雑音テンプレートのデータベース作成に実験 2,3 では D1 を，実験 3 では D2-D6 の最初の 10 秒間を使用した．

Robovie-W を用いた実験では R1 を用いて INMF による動作雑音の基底の学習を行った．

実験 2.1： Hearbo では D1,Robovie-W では R1 をテストデータとして与え, 動作雑音の推定が可能であるか評価した．指標には *Noise Estimation Error(NEE)* を用いて評価を行った．

$$\mathrm{NEE} = 20\log_{10}\left(\frac{\sum_f\sum_\omega |N(\omega,f)|^2}{\sum_f\sum_\omega (|N(\omega,f)| - |\hat{N}(\omega,f)|)^2}\right) \quad (16)$$

N 及び $\hat{N}$ はそれぞれオリジナルと推定した動作雑音スペクトルを表す.

実験 2.2: Hearbo では D1 に D7 を, Robovie-W では R1 に雑音を含まない音声を加算した音響信号をテストデータとして与え動作雑音抑圧を行った．評価指標には実験 1 と同様に SNR_1,SIR,SDR を用いた．

実験 2.3： Hearbo の場合は D2-D6 を,Robovie-W の場合は R2 をテストデータに用いた．評価は SNR, 実験 2.3 では正解となる音声及び動作雑音は分からない．そこで SNR を改良した SNR_2 を以下に定義し，評価にはこれを用いた．

[1] http://bass-db.gforge.inria.fr/bss eval/

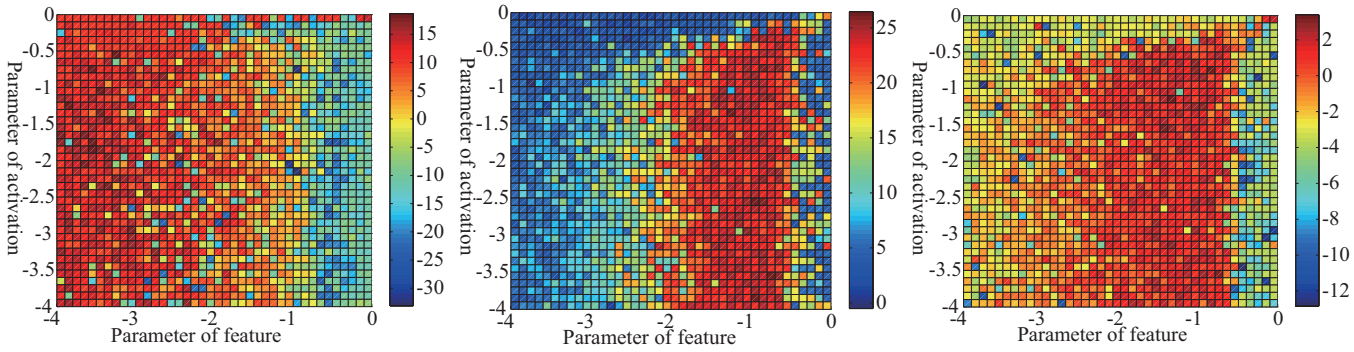
$$\begin{aligned}\mathrm{SNR}_2 &= 20\log_{10}\left(\frac{\sum_f\sum_\omega \left|\hat{S}(\omega,f)\right|^2}{\sum_f\sum_\omega \left||Y_N(\omega,f)| - |\hat{N}(\omega,f)|\right|^2}\right) \\ &\quad -20\log_{10}\left(\frac{\sum_f\sum_\omega |Y_S(\omega,f)|^2}{\sum_f\sum_\omega |Y_N(\omega,f)|^2}\right) \quad (17)\end{aligned}$$

Y_S 及び Y_N は各々入力信号の雑音部分と音声部分を示す．また,$\hat{S}$ は提案手法により推定した音声を示す．なお, テンプレートを用いた手法では $\hat{S}$ の推定ができないため, $|Y_S(\omega,f)| - |\hat{N}(\omega,f)|$ を代わりに使用した．

結果： 実験 2.1,2.2 の結果を表 2 に示す．表 2 の提案手法における基底の数はテンプレートを用いた雑音抑圧手法におけるテンプレートの数に対応する．提案手法とテンプレートを用いる手法で推定結果を比較すると，提案手法では推定に用いる基底の数は Hearbo の場合は 7 つ,Robovie-W の場合は 5 つのみであるが，テンプレート数が 300 個を超える場合よりも NEE の値が良いことがわかる．

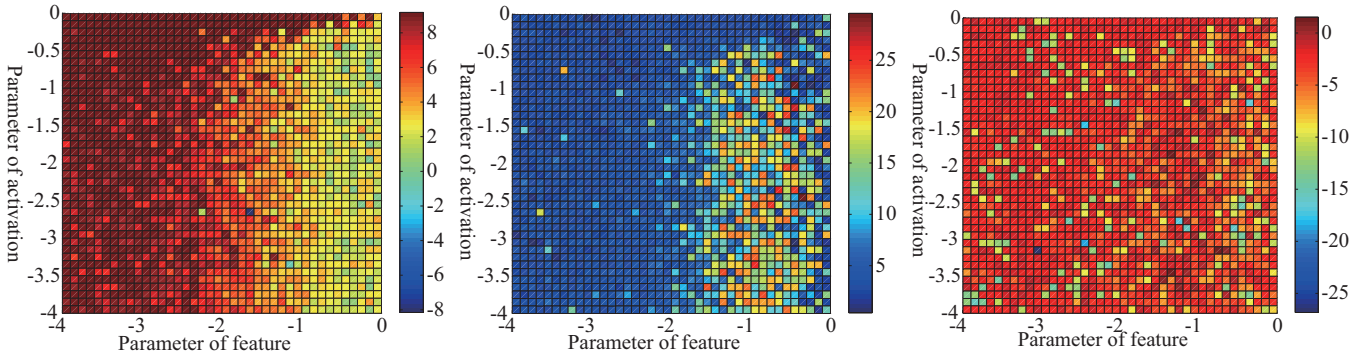
Hearbo を用いた実験 2.2 において提案手法では音声に対し基底数を 2 つ，雑音の基底と合わせて 9 つの基底が推定された．その抑圧性能を SNR_1 で評価すると，テンプレートの数が 1,022 の場合とほぼ同等の結果であった．また，SIR，SDR による評価では，テンプレート数に関わらずテンプレートを用いた手法よりも提案手法が良い結果が得られた．基底やテンプレートの数が多くなればなるほど，計算コストは大きくなる．そのため，少ない基底数で雑音の推定と抑圧が可能であるということは大きな利点である．図 8 に Hearbo を用いた実験 2.2 で推定した各信号及びその元信号のスペクトログラムを示した．図 8a) は,8b) に示した動作雑音と 8c) の音声を足し合わせた合成音で, 右が各信号を提案手法を用いて再構成した信号である．スペクトログラムからも動作雑音, 音声が推定できている事がわかる．また，実験 1 でも記述したが再構成した音声には環境雑音が含まれていない．これは動作雑音の基底の学習の際，動作雑音に含まれた環境雑音によって環境雑音の基底が学習され，既知の基底に含まれていたためと考えられる．Robovie-W は関節角情報を得られないためテンプレートを用いた手法との比較は不可能であるが，表 2 及び図 9 から Hearbo の場合と近い結果が得られていることがわかる．これらの結果から，動作雑音及び音声は提案手法によって推定可能であることがわかる．

実験 2.3 の結果を表 3 に示す．表 3 の”#of templates”はテンプレートを用いた動作雑音抑圧において，SNR_2 が最大となった時に使用したテンプレートの数を示している．表 3 より,D2-D6 全ての場合において提案手法のほうがテンプレートを用いた手法より SNR_2 が良い結果が得られた．学習データとは異なるテストデータを用いたオープン



a)SNR b)SIR c)SDR

Figure 6: 実験 1 （Robovie-W）



a)SNR b)SIR c)SDR

Figure 7: 実験 1 (Hearbo)

テストでも効果が確認できたことから，今回の提案手法がロバストであることがわかった．Robovie-W を用いた実験の結果は SNR_2 は 3.9dB であった．この値は Hearbo の場合よりも小さい値となったが，Robovie-W と話者の距離が Hearbo の場合よりも離れていたことが原因と考えられる．また，テンプレートベースの手法を用いた場合，実験 2.2 でのクローズテストでは非常に多くのテンプレート数を用いた時良い結果が得られた．しかし，実験 2.3 のオープンテストではテンプレートの数が少ないほうが良い結果が得られた．これは，同じ動作であっても常に同じ動作雑音が発生するとは限らない，すなわち完全な相関関係では無いためである．

5 おわりに

本論文では単一のマイクで関節角や他のセンサーの情報を用いない新たな動作雑音抑圧手法を提案し，その評価を行った．具体的には，*Infinite Non-negative Matrix Factrization* によって予め動作雑音の基底を学習し，*Semi-Blind INMF* によって動作雑音とた目的音を分離する手法を提案した．そして実際に 2 種類のロボットを用いた動作雑音抑圧実験を行い，提案手法の有効性を示した．今後は提案手法によって動作雑音が抑圧された音声による音声認識を行う予定である．

謝辞

本研究の一部は科研費（24118702，22700165）の補助を受けた．

参考文献

[1] K. Nakadai *et al.* Active audition for humanoid. *AAAI 2000*, pp. 832–839.

[2] K. Nakadai *et al.* Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. *Speech Communication*, Vol. 44, pp. 97–112, 2004.

[3] Y. Nishimura *et al.* Speech recognition for a humanoid with motor noise utilizing missing feature theory. *Humanoids 2006*,IEEE pp. 26–33. IEEE.

[4] T. Rodemann *et al.* Sound localization for humanoid robots — building audio-motor maps based on the hrtf. *IROS 2006*,IEEE/RSJ pp. 1171—1176..

[5] J. Hornstein *et al.* Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. *IROS 2006*,IEEE/RSJ pp. 860—865.

[6] J. Hornstein *et al.* Spectral cues for robust sound localization with pinnae. *IROS 2006*,IEEE/RSJ pp. 386-391.

[7] A.Portello *et al.* Active binaural localization of intermittent moving sources in the presence of false measurements. *IROS 2012*,IEEE/RSJ pp. 3294—3299.

[8] J.-M. Valin *et al.* Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In IEEE, editor, *ICRA 2004*,IEEE-RAS, pp. 1033—1038.

[9] Y.Sasaki *et al* . Spherical microphone array for spatial sound localization for a mobile robot. *IROS 2012*,IEEE/RSJ pp. 713-718.

Table 2: 実験２，３の結果

Robot	Hearbo							Robovie W
	Proposed	Template-based						Proposed
# of feat. /templ.	7 (ego-noise) 2 (speech)	31	98	303	1,022	3,115	8,431	5(noise) 3(speech)
NEE (dB)	9.4	8.0	8.2	8.0	9.9	12.3	24.6	9.3
SNR_1 (dB)	7.2	6.1	6.3	5.9	7.4	8.7	14.6	7.6
SIR (dB)	11.0	1.4	2.6	2.6	3.0	2.3	2.2	17.0
SDR (dB)	1.3	-0.8	1.1	1.1	1.3	0.8	0.8	2.0

Table 3: 実験４の結果

Robot	Dataset	Proposed	Template-based	
		$SNR_2(dB)$	$SNR_2(dB)$	# of templates
Hearbo	D2(J1)	5.9	2.5	21
	D3(J1+J2)	5.2	3.1	8
	D4(J1–J3)	6.3	3.2	12
	D5(J1–J4)	3.5	-0.76	244
	D6(J1–J5)	5.3	2.4	45
Robovie-W	R2	3.9	N/A	N/A

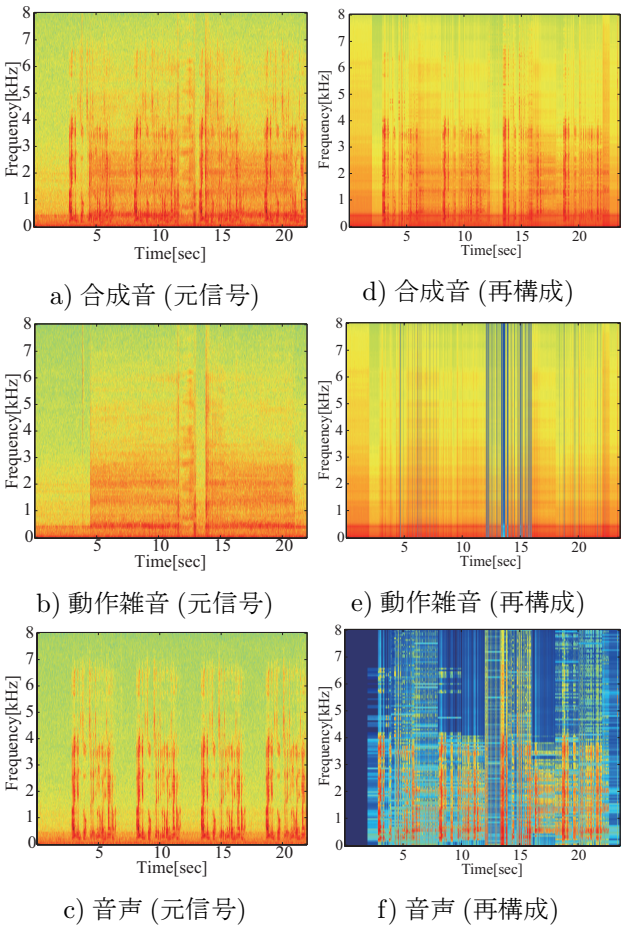


Figure 8: Hearbo での動作雑音抑圧

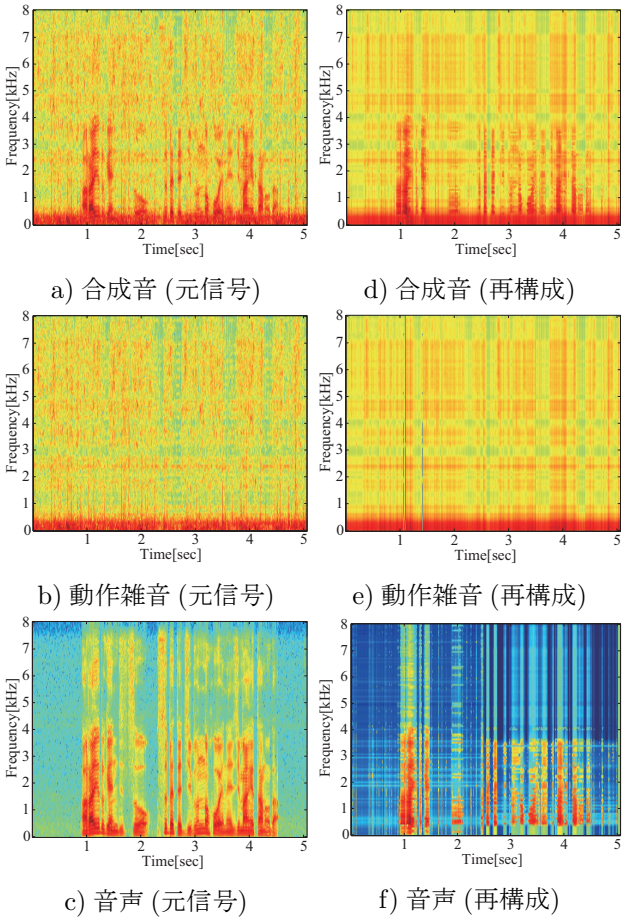


Figure 9: Robovie-W での動作雑音抑圧

[10] S.Yamamoto *et al.* Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech. *ASRU-2007*, pp. 111–116. IEEE, Dec. 2007.

[11] H.Saruwatari *et al.* Two-stage blind source separation based on ica and binary masking for real-time robot audition system. *IROS 2005*, pp. 209–214. IEEE, 2005.

[12] T.Yoshida and K.Nakadai . Active audio-visual integration for voice activity detection based on a causal bayesian network. *Humanoids 2012*,IEEE pp. 370-375.

[13] K.Nakadai *et al.* Sound source tracking with directivity pattern estimation using a 64ch microphone array. *IROS 2005*,IEEE/RSJ pp. 196-202.

[14] F. Perrodin *et al.* Design and calibration of large microphone arrays for robotic applications. *IROS 2012*,IEEE/RSJ pp. 4596-4601.

[15] J.Even *et al.* "Combining laser range finders and local steered response power for audio monitoring. *IROS 2012*,IEEE/RSJ pp. 986-991.

[16] J.Even *et al.* "Semi-blind suppression of internal noise for hands-free robot spoken dialog system. *IROS 2009*,IEEE/RSJ pp. 658-663.

[17] A. Ito *et al.* Internal noise suppression for speech recognition by small robots. *Eurospeech 2005*, pp. 2685–2688.

[18] S. F. Boll. A spectral subtraction algorithm for suppression of acoustic noise in speech. *ICASSP 1979)*, IEEE, pp. 200–203.

[19] Gokhan Ince *et al.* Incremental learning for ego noise estimation of a robot. *IROS 2011*, IEEE/RSJ, pp. 131–136.

[20] S. Abdallah and M. D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. *ISMIR 2004*, pp. 10–14, 2004.

[21] M. N. Schmidt, M.Mørup. Infinite non-negative matrix factorization. *EUSIPCO*, 2010.

[22] M. D. Hoffman *et al.* Bayesian nonparametric matrix factorization for recorded music. *ICML 2010*, pp. 439–446.

Hands-free Speech Recognition Robust to distance and Azimuth in Robot Application

Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto and Kazuhiro Nakadai
Honda Research Institute Co. Ltd., Japan

***Abstract*— In this paper we present two methods in addressing the changes in radial position and azimuth, respectively, relative to the robot and speaker. In the case of the former, room transfer function (RTF) estimation is employed via waveform-level compensation to reflect the change in power caused by the change of radial position to the RTF. In addition, acoustic model-level compensation is also used to complement the effect of robustness to changes in radial position. Finally, equalization is utilized to mitigate the effects of the change in the azimuth. All of the processes in the two methods are in accordance to maximizing the automatic speech recognition (ASR) performance for effective human-robot communication. Experimental evaluation in real environment condition confirms the robustness in recognition performance when used in hands-free human-robot communication.**

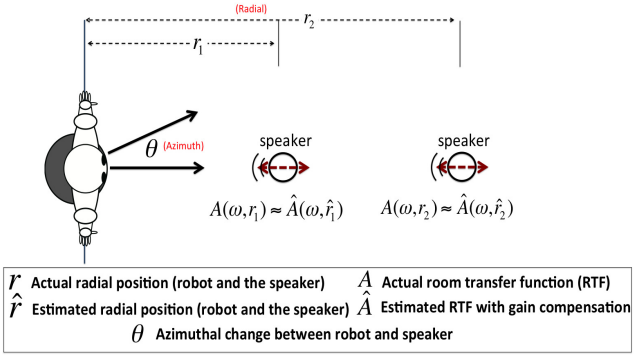


Fig. 1. Room scenario: Change in speaker's radial and azimuthal position.

I. Introduction

Automatic speech recognition (ASR) is one of the most important component in hands-free human-robot communication. Before robots execute any speech-based command, the speech acoustic signal has to be converted into text using the ASR and then, processed by an intelligent system for machine understanding. As the acoustic speech signal travels through free space inside an enclosed room, the observed signal at the microphone is distorted due to reflections known as reverberation. Late reverberation leads to a significant deterioration of the ASR performance. Dereverberation is therefore needed in a robust ASR system. In robot applications, this problem is further complicated since we cannot control the position of the speaker. When a speaker changes its radial position relative to the robot (r_1 to r_2 and vice versa) as shown in Fig. 1, the reverberant speech power observed at the microphones also changes, resulting in a mismatch to the acoustic model used in the ASR system. Thus, the dereverberation algorithm should adapt by estimating the change in location $\hat{r}$ and RTF $\hat{A}$ to minimize the effect of mismatch.

We have previously proposed an enhancement algorithm that automatically detects the reverberation time (RT) inside the room, then synthetically generate a new RTF from the pre-measured one [1][2]. The estimated RTF is used to enhance the reverberant speech by removing its late reflection component in conjunction with our ASR-based dereverberation scheme [1][2]. Although this method works, the assumption is very conservative. First, it assumes that only RT plays a significant role in describing the RTF. This assumption is only valid in symmetric rooms with no occlusions (i.e., chairs, tables, etc.). In real environments where robots are dispatched, it is fair assumption that the room will be filled with objects and speakers. Second, it does not take into consideration the change in speech power. It assumes that the speaker is stationary, which is unrealistic because we cannot control the speaker's radial position. Lastly, it is not immune to changes in the angular position (azimuth) of the speaker relative to the robot. These causes a major problem as the ASR system is very sensitive to the change in speech power (mismatch). We define radial position as the absolute distance between the speaker and the robot while the azimuth is the angle from a reference position (robot) to the speaker.

First, we will show a method that significantly improves the RTF estimation of our previous work [1][2]. The new method is capable of compensating the change in speaker's radial position through room RTF compensation. In effect, the variation in speech power which is crucial in ASR is also considered in the new RTF estimate. As a result, we can achieve a robust ASR performance in realistic environments where robots are deployed. The proposed compensation is done in two-way synergetic processes, accounting for both the waveform and acoustic model aspects that affect ASR performance. In the waveform level, the RTF is compensated in a manner that reflects possible changes in speaker's radial position. This focuses on the impact of the speaker's power variation to the RTF waveform. Consequently, the acoustic model-level compensation connects the waveform-level compensation to the ASR, by adopting the criterion used by the ASR system in estimating the RTF. This guarantees that the estimated RTF translates to ASR performance improvement, when used in conjunction with our ASR-based dereverberation scheme. Secondly, we will show the method in addressing changes in the azimuth via equalization.

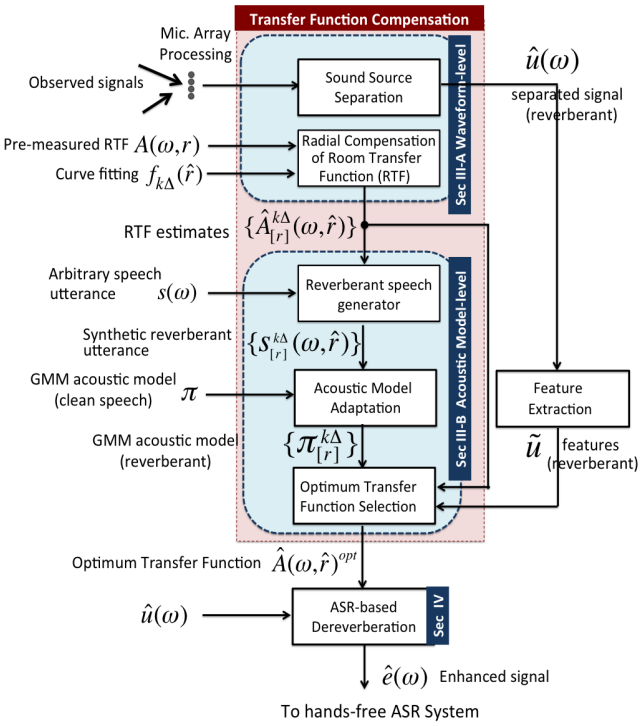


Fig. 2. Proposed RTF estimation for ASR-based dereverberation.

The organization of the paper is as follows; in Section II, the background of RTF estimation for ASR application is introduced. Then, in Section III, we show the method of our proposed RTF estimation with radial compensation, followed by equalization scheme for azimuthal change in Section IV. In Section V, we discuss the experimental set-up, together with recognition results using real reverberant data collected in a human-robot communication environment in Section VI. We will conclude this paper in Section VII.

II. BACKGROUND

A number of dereverberation approaches use readily available RTF through room impulse response measurement [3] [4]. Due to the dynamics inside the room, the generalization of the RTF becomes unrealistic in real environments. This arises the need of physically measuring several RTFs inside the room, which is impractical. Thus, RTF estimation becomes an interesting research topic. There are a number of RTF estimation techniques focusing on waveform accuracy. Although this is important, the requirement for RTF estimation is different when used in hands-free ASR applications. For example, when RTF is used in dereverberation for ASR, we are not interested in accurately modeling the reverberant speech but in estimating the late reflections, in which RTF is not the sole-determining factor. In ASR, we convert speech waveform to models (i.e., Hidden Markov Models (HMMs)). This process requires a conversion of a rich signal information to a more watered-down representation. Each phone HMM represents a short speech segment with a duration of 30-100 msec, and each state captures information about a distribution of spectral parameters. With this perspective, the HMMs' description of speech is of low resolution, compared to the RTF, with respect to time and frequency. Thus, for ASR application, it may be sufficient to use an RTF estimate instead of the accurate RTF [6]. In our previous work [1][2] we adopted an RTF estimator based on the premise that the multiple reflections of sound can be described by a decaying acoustical energy [5] given as,

$$A^2(l) \approx e^{(6\ \ln(10)/RT)\ l}, \tag{1}$$

where l is the discrete time sample, and RT is the reverberation time. From Eq. (1) we can easily derive the RTF's frequency domain equivalence $\hat{A}(\omega)$, where ω denotes frequency domain. We note that this RTF estimate does not take into consideration the distance between the speaker and the robot. This model is applicable only if there is not much perturbation inside the room. Moreover, it fails when the speaker moves along the radial axis or changes angular position relative to the robot because it does not model the variation in speech power as a function of position change, which the ASR is very sensitive to. Thus, we need a new RTF estimation method that takes care of the change in the power of the reverberant speech to minimize mismatch in the ASR system.

There is no guarantee whether an estimated RTF (even using the most accurate ones) would result to an improvement in the ASR performance when used in conjunction with any dereverberation scheme for hands-free human-robot communication [1][2]. RTF estimation should be based on the practical requirements for which it is to be utilized. In the case of hand-free human-robot ASR application, the design criterion should not be based on the waveform accuracy of the RTF estimate. It is prudent to adopt the criterion used by the ASR system (i.e., acoustic model likelihood criterion) as part of the RTF estimation criterion, which is the fundamental objective of this paper.

III. ROBUSTNESS TO RADIAL POSITION

The proposed RTF estimation method is shown in Fig. 2. First, the microphone array signals are processed resulting in $\hat{u}(\omega)$. Then, RTF is estimated by compensating the pre-measured RTF $A(\omega,r)$ together with the curve fitting function derived offline $f_{k\Delta}(\hat{r})$. The step-size increment $k\Delta$ generates a set of RTFs $\{A_{[r]}^{k\Delta}(\omega,\hat{r})\}$. Using these RTFs, together with an arbitrary clean speech utterance $s(\omega)$, a set of synthetic reverberant data are generated $\{s_{[r]}^{k\Delta}(\omega,\hat{r})\}$. Consequently, the clean acoustic model π (Gaussian Mixture Model (GMM)) is adapted using the generated reverberant data resulting to adapted models $\{\pi_{[r]}^{k\Delta}\}$. Then, the optimum RTF is selected by evaluating the likelihood scores using the features $\tilde{u}$ of the separated reverberant signal $\hat{u}(\omega)$. The corresponding k that maximizes the likelihood score is used to select among $\{A_{[r]}^{k\Delta}(\omega,\hat{r})\}$ the optimal RTF $\hat{A}(\omega,\hat{r})^{opt}$. This optimal RTF is then used in conjunction with the ASR-based dereverberation scheme.

A. Acoustic Waveform Compensation

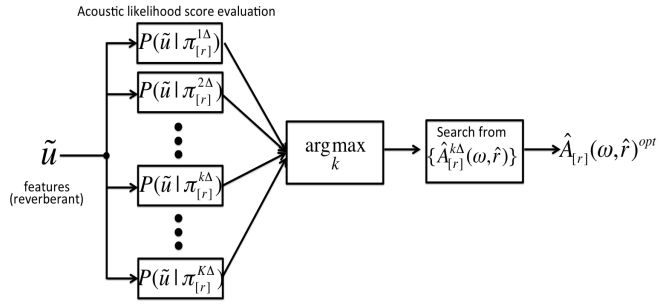


Fig. 3. Optimum RTF selection based on maximum likelihood estimation.

1) Microphone Array Sound Separation: Suppose that there are N sources and M ($\geq N$) microphones. Let $\boldsymbol{u}(\omega)$ denote the input acoustic signals of N sources in frequency domain, described as $\boldsymbol{u}(\omega) = [u_1(\omega), \cdots, u_N(\omega)]^T$, where T represents the transpose operator. $\boldsymbol{x}(\omega) = [x_1(\omega), \cdots, x_M(\omega)]^T$ is the vector containing the signals received by M microphones. The model for microphone array signal processing is described as follows:

$$\boldsymbol{x}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{u}(\omega) + \boldsymbol{n}(\omega), \tag{2}$$

where $\boldsymbol{A}(\omega) \in \mathbb{C}^{M\times N}$ is a *Transfer Function (TF)* matrix between a microphone array and sound sources; $\boldsymbol{n}(\omega)$ denotes an additive noise, which is assumed to be statistically independent of $\boldsymbol{u}(\omega)$.

The sound sources are spatially separated by a hybrid algorithm of beamforming and blind separation called *Geometrically constrained High-order Decorrelation based Source Separation (GHDSS)*. The input vector $\boldsymbol{x}(\omega)$, $\hat{\boldsymbol{u}}(\omega)$ is used to define by $\hat{\boldsymbol{u}}(\omega) = \boldsymbol{V}(\omega)\boldsymbol{x}(\omega)$ in frequency domain, where $\boldsymbol{V}(\omega)$ stands for the separation matrix. GHDSS updates $\boldsymbol{V}(\omega)$ so that it can correctly estimate $\boldsymbol{u}(\omega)$ in Eq. (2) by $\hat{\boldsymbol{u}}(\omega)$. In order to estimate $\boldsymbol{V}(\omega)$, GHDSS introduces two cost functions, that is, separation sharpness (J_{SS}) and geometric constraints (J_{GC}):

$$J_{SS}(\boldsymbol{V}(\omega)) = \|\phi(\hat{\boldsymbol{u}}(\omega))\hat{\boldsymbol{u}}^H(\omega) - \mathrm{diag}[\phi(\hat{\boldsymbol{u}}(\omega))\hat{\boldsymbol{u}}^H(\omega)]\|^2$$
$$J_{GC}(\boldsymbol{V}(\omega)) = \|\mathrm{diag}[\boldsymbol{V}(\omega)\boldsymbol{A}(\omega) - \boldsymbol{I}]\|^2$$

where $\|\cdot\|^2$ indicates the Frobenius norm, diag$[\cdot]$ is the diagonal operator, and H represents the conjugate transpose operator. For a nonlinear function, $\phi(\hat{\boldsymbol{u}}(\omega))$, we selected a hyperbolic-tangent-based function [7] in this paper. Since the best $\boldsymbol{V}(\omega)$ is always changing in the real world, $\boldsymbol{V}(\omega)$ is adaptively updated as described in [8]. Consequently, the separated signal $\hat{u}(\omega)$ that satisfies all the criterion is achieved from the vector $\hat{\boldsymbol{u}}(\omega)$ in the same manner in [8].

2) Radial Compensation of Room Transfer Function: Let $A(\omega, r)$ denote pre-measured RTF between any of the microphones in the array and sound sources. Here, r is the distance between a microphone in the array and the sound sources. Our objective is to estimate $A(\omega, \hat{r})$ by radial compensation, where $\hat{r}$ is the radial distance of an estimated point. The following conditions are assumed:

a) The sound sources are in a far field. This implies that the phase of the RTF is static with respect to the change in distance.
b) The amplitude of $|A(\omega, \hat{r})|$ decays exponentially with $\hat{r}$.

Under these assumptions, RTF of unknown distance can be estimated as follows:

$$\hat{A}_{[r]}(\omega, \hat{r}) = f(\hat{r})A(\omega, r)\ , \tag{3}$$

where $\hat{A}_{[r]}(\omega, \hat{r})$ is an estimated RTF at $\hat{r}$ using pre-measured $A(\omega, r)$. $f(\hat{r}) \in \mathbb{R}$ is the exponential gain function of $\hat{r}$. Since $f(\hat{r})$ is unknown, $f(\hat{r})$ is obtained as a priori information based on a nonlinear curve fitting using measured RTFs. Specifically, $f(\hat{r})$ is described by

$$f(\hat{r}) = \frac{\alpha_1}{\hat{r}} + \alpha_2, \tag{4}$$

where α_1 and α_2 are the estimated fitting parameters.
The steps for the radial compensation are as follows:

1) Measure a limited amount of RTFs along the radial axis with the microphone array with different r, denoted as $A(\omega, r_{[i]})$, where $r_{[i]}$ is a measured point, and i_r is the number of measured points.
2) Obtain mean amplitude of RTFs over frequency bins of $A(\omega, r_{[i]})$ by

$$\bar{A}(r_{[i]}) = \frac{1}{p_h - p_l + 1}\sum_{p=p_l}^{p_h}\left|A(\omega_{[p]}, r_{[i]})\right|\ , \tag{5}$$

where p_h and p_l are the indexes of the maximum and minimum frequencies respectively.
3) Obtain α_1 and α_2 through nonlinear curve fitting as follows:

$$\boldsymbol{F}_x = \begin{bmatrix} \frac{1}{r_{[1]}} & 1 \\ \vdots & \vdots \\ \frac{1}{r_{[i_r]}} & 1 \end{bmatrix}, \boldsymbol{F}_y = \begin{bmatrix} \bar{A}_m(r_{[1]}) \\ \vdots \\ \bar{A}_m(r_{[i_r]}) \end{bmatrix},$$
$$[\alpha_1, \alpha_2]^T = \left(\boldsymbol{F}_y^T\boldsymbol{F}_y\right)^{-1}\boldsymbol{F}_y^T\boldsymbol{F}_y \tag{6}$$

4) Select a reference RTF $A_m(\omega, r)$ among the pre-measured RTFs.
5) $\hat{A}_{[r]}(\omega, \hat{r})$ is estimated by Eq. (3) with α_1 and α_2 in Eq. (6).

For practical reasons, it is desirable to allow more degrees of freedom for the compensated RTF, Eq. (4) is allowed to deviate within a close neighbourhood in a discrete step-wise manner $k\Delta$ for $k = 1, .., K$. This covers the uncertainty of the RTF estimate which proved to be effective in our previous work [1][2]. Thus, Eq. (4) becomes

$$f_{k\Delta}(\hat{r}) = k\Delta\ \frac{\alpha_1}{\hat{r}} + \alpha_2,$$

where k is an integer and Δ is a constant value derived experimentally. Thus, Eq. (3) is rewritten as

$$\hat{A}_{[r]}^{k\Delta}(\omega, \hat{r}) = f_{k\Delta}(\hat{r})A(\omega, r)\ .$$

By introducing $k\Delta$, we are able to generate a set of RTF estimates $\{\hat{A}_{[r]}^{k\Delta}(\omega, \hat{r})\}$ within a close neighbourhood $\hat{A}_{[r]}(\omega, \hat{r})$.

The selection of the most probable (optimum) RTF is done through acoustic model likelihood score evaluation discussed in Sec III-B-3.

B. Acoustic Model-level Compensation

1) Reverberant Speech Generator: Using an arbitrary speech utterance $s(\omega)$ and transfer function $\hat{A}_{[r]}^{k\Delta}(\omega,\hat{r})$, we synthesize reverberant speech data

$$s_{[r]}^{k\Delta}(\omega,\hat{r}) = s(\omega)\hat{A}_{[r]}^{k\Delta}(\omega,\hat{r}). \tag{7}$$

For a set of transfer functions $\{\hat{A}_{[r]}^{k\Delta}(\omega,\hat{r})\}$, we generate the corresponding set of synthetic reverberant utterances $\{s_{[r]}^{k\Delta}(\omega,\hat{r})\}$. The synthesized reverberant speech is used for acoustic model adaptation.

2) Acoustic Model Adaptation: Prior to acoustic model adaptation, a Gaussian Mixture Model (GMM) π is trained using a clean speech database (i.e., the database used in our ASR). The GMM is composed of four states which guarantee that the temporal smearing caused by the reverberation is captured by the model (i.e., in the states). Since we are only interested in the reverberant effects of the sound and not its phonetic meaning, the GMM is designed to be phoneme, speaker and gender independent. Thus, we can use any arbitrary utterance. Model adaptation is implemented through Maximum Likelihood Linear Regression (MLLR) which is effective when dealing with small amount of data [9]. MLLR can adapt the means and covariance of π. Using the synthesized reverberant speech $\{s_{[r]}^{k\Delta}(\omega,\hat{r})\}$ as adaptation data, the MLLR adapted mean for every r is expressed as

$$\boldsymbol{\mu}_{m_c}^{k\Delta} = \boldsymbol{W}_{m_c}\boldsymbol{\zeta}_{m_c}\,, \tag{8}$$

where $\boldsymbol{W}_{m_c}$ is the transformation matrix while $\boldsymbol{\zeta}_{m_c}$ refers to the extended mean vector. The subscript m_c denotes the mixture m at class c. The adapted covariance is given as

$$\boldsymbol{\Sigma}_{m_c}^{k\Delta} = \boldsymbol{B}_{m_c}^{T}\boldsymbol{H}_c\boldsymbol{B}_{m_c}\,, \tag{9}$$

where $\boldsymbol{B}_{m_c}$ is the inverse of the Choleski factor

$$\boldsymbol{B}_{m_c} = \boldsymbol{C}_{m_c}^{-1}\,. \tag{10}$$

Both the transformation matrix $\boldsymbol{W}_{m_c}$ and the linear transformation $\boldsymbol{H}_c$ are calculated using the adaptation data $s_{[r]}^{k\Delta}(\omega,\hat{r})$ discussed in Sec. III-B-1. The adapted GMM $\pi_{[r]}^{k\Delta}$ has the corresponding means $\boldsymbol{\mu}_{m_c}^{k\Delta}$ and variance $\boldsymbol{\Sigma}_{m_c}^{k\Delta}$, respectively.

3) Optimum Transfer Function Selection: This process shown in Fig. 3 is crucial as it takes into consideration the overall contribution of the different variables/processes operating in the entire system in the selection of the optimal RTF. It connects the different modules, such as the waveform-level compensation, the ASR mechanism and the reverberant signal $\hat{u}(\omega)$ altogether. As a result, the RTF choice guarantees optimal ASR performance when used in conjunction with the ASR-based dereverberation scheme. Using the GMM adapted models $\pi_{[r]}^{k\Delta}$ discussed in Sec III-B-2, we identify the corresponding k associated with each model that best matches the actual reverberant signal $\hat{u}(\omega)$. The spectral features $\tilde{u}$ of $\hat{u}(\omega)$ is extracted and used to evaluate the likelihood scores based on the acoustic likelihood criterion, which is identical to the one used in the ASR system. The corresponding k that results in the highest likelihood score is used to select the optimal RTF.

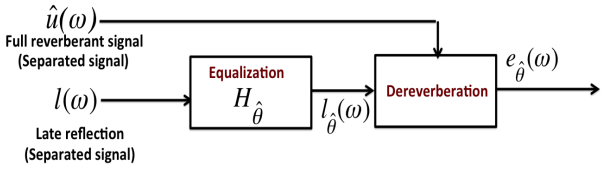


Fig. 4. Robustness to azimuthal change via equalization.

The ASR-based dereverberation in [1][2] is adopted. In the offline training mode (left figure), we replaced the previous RTF estimator based on RT (see Sec II) with the proposed method discussed in Sec III. After selecting $\hat{A}(\omega,\hat{r})^{opt}$, we extract the late reflection component of the RTF $\hat{A}_l(\omega,\hat{r})^{opt}$ [1][2]. Then, the reverberant signal $\hat{u}(\omega)$ is simulated using $\hat{A}(\omega,\hat{r})^{opt}$ and the clean speech database $c(\omega)$ in which we extract the late reflection approximation $\hat{l}(\omega)$. In the same manner, using $\hat{A}_l(\omega,\hat{r})^{opt}$ and $c(\omega)$ we can simulate the real late reflection $l(\omega)$. Consequently, $\{\delta_1,...,\delta_B\}$ is optimized with the objective of minimizing the error between $l(\omega)$ and $\hat{l}(\omega)$. The optimized weighting parameter $\{\delta_1,...,\delta_B\}^{opt}$ guarantees the actual dereverberation (right figure) to work even when using only the $\hat{l}(\omega)$ and not the actual $l(\omega)$. Dereverberation through the modified SS [2] is obtained through

$$|e(\omega)|^2 = \begin{cases} |\hat{u}(\omega)|^2 - \delta_b|\hat{l}(\omega)|^2 \\ \quad \text{if } |\hat{u}(\omega)|^2 - \delta_b|\hat{l}(\omega)|^2 > 0 \\ \\ \beta|\hat{u}(\omega)|^2 \quad \text{otherwise,} \end{cases} \tag{11}$$

where δ_b is the multi-band weighting parameters optimized through an offline training scheme. The multi-band treatment improves error minimization as opposed to single-band. In the actual dereverberation, these parameters are used together with $\hat{u}(\omega)$ to recover $\hat{e}(\omega)$ for ASR.

IV. Robustness to Azimuth Change Via Late Reflection Equalization

In theory, multiple unique RTFs are needed to match the corresponding change in azimuthal orientation θ for each channel (i.e., $\boldsymbol{A}_\theta(\omega)$). This is because when θ changes, the acoustical dynamics inside the room is perturbed as the concentration of speech power changes as a function of θ. In short, the late reflection also varies with θ in reverberant environments like the one in our work in [15]. However, it is impractical to measure all possible θ variations since it requires a corresponding RTF measurement for all M microphones. To mitigate this, we employed an equalization scheme, by dealing with the source-separated late reflection $l(\omega)$ instead of the multi-channel RTF characteristics. This scheme simplifies the supposed complicated analysis of

TABLE I

AVERAGED WORD RECOGNITION RATE (%) (RT= 240 MSEC AND ROOM 2 RT = 640 MSEC)

Methods (Room 1)	0.5 m	1.0 m	1.5 m	2.0m	2.5 m
A. No processing	79.1%	73.2%	57.6%	35.3%	20.6%
B. Blind Dereverberation	80.3%	76.9%	65.6%	49.6%	37.7%
C. Previous Method (Sec II)	**81.2%**	**78.3%**	**71.3%**	**55.7%**	**46.1%**
D. Waveform Compensation (Sec III-A)	81.6%	79.4%	73.1%	57.2%	50.8%
E. Waveform and Acoustic Model Comp. (Sec III-A&B)	82.3%	81.2%	75.8%	60.7%	55.4%
F. Waveform and Acoustic Model Comp. + Equalization (Sec III-A&B + Sec IV)	**82.9%**	**82.3%**	**77.5%**	**62.7%**	**57.8%**
Methods (Room 2)	0.5 m	1.0 m	1.5 m	2.0m	2.5 m
A. No processing	31.8%	15.6%	0.40%	-8.10%	-20.2%
B. Blind Dereverberation	41.9%	33.4%	20.6%	10.0%	0.90%
C. Previous Method (Sec II)	**45.0%**	**38.3%**	**26.9%**	**16.1%**	**7.40%**
D. Waveform Compensation (Sec III-A)	46.3%	41.1%	32.5%	23.5%	16.5%
E. Waveform and Acoustic Model Compensation (Sec III-A&B)	48.4%	43.8%	36.7%	28.4%	22.1%
F. Waveform and Acoustic Model Comp. + Equalization (Sec III-A&B + Sec IV)	**50.1%**	**46.4%**	**38.9%**	**30.7%**	**25.9%**

the effect of the azimuthal orientation with respect to the multi-channel RTFs into simple single channel filtering. The equalized late reflection signal becomes

$$l_\theta(\omega) = l(\omega)H_\theta. \tag{12}$$

where $l(\omega)$ is the separated late reflection using a generic (unmatched) RTF while H_θ is the equalizer.

H_θ is a filter derived experimentally during the offline mode by analyzing the response of the late reflection as a function of the actual azimuthal change θ. Suppose that $l_\theta(\omega)$ is the actual late reflection with a corresponding multi-channel RTF $\boldsymbol{A}_\theta(\omega)$. The filter design involves the poles positioning method on a logarithmic frequency grid based on [12][13]. The target response is set to $l_\theta(\omega)$ and H_θ for $\{\theta_1,\ldots,\theta_g,\ldots,\theta_G\}$ are derived by properly positioning the poles to achieve the target response $l_\theta(\omega)$ [14]. Note that the target response $l_\theta(\omega)$ was preprocessed via smoothing to avoid direct inversion problems [14]. With an effective θ selection procedure similar to that in Fig. 3 the equalization process [15], azimuthal change is compensated via filtering.

Dereverberation based on [2] is given as

$$|e_{\hat{\theta}}(\omega)|^2 = \begin{cases} |\hat{u}(\omega)|^2 - H_{\hat{\theta}}(\omega)|l_{\hat{\theta}}(\omega)|^2 \\ \quad \text{if } |\hat{u}(\omega)|^2 - H_{\hat{\theta}}(\omega)|l_{\hat{\theta}}(\omega)|^2 > 0 \\ \beta|\hat{u}(\omega)|^2 \quad \text{otherwise.} \end{cases} \tag{13}$$

where $\hat{u}|(\omega,t)|^2$ is the power of the separated reverberant signal ($\hat{u}|(\omega,t)|^2 \approx r|(\omega,t)|^2$) and $l_{\hat{\theta}}|(\omega,t)|^2$ is the separated late reflection power. We note that the equalization process is key to the hybrid approach as it eliminate δ in the Eq. (13). In our previous method [16], the dependence on the δ parameter was the stumbling block towards the utilization of multi-channel processing since the optimization of δ is computationally expensive for multi-channel signals. This limitation is rectified in the proposed method.

V. EXPERIMENTAL SET-UP

A. Training and Testing Database for ASR

The training database is from the Japanese Newspaper Article Sentence (JNAS) corpus. The open test set consists of 200 test utterances coming from 24 speakers. Recognition experiments are carried out on the Japanese dictation task with 20K-word vocabulary . The language model is a standard word trigram model. The acoustic model is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. Test experiment is conducted using actual human-robot communication set-up. The microphone array is embedded on the head of the robot. In the experiment, we used different occlusions such as table, chairs, etc. (real environment setting).

Real reverberant data are recorded inside two different reverberant rooms (Room 1 and Room 2) with RT of 240 ms and 640 ms respectively. Six different radial axes at different azimuth θ selected randomly. We considered five radial location points $r_{[i]} = \{0.5\text{m}, 1.0\text{m}, 1.5\text{m}, 2.0\text{m}, 2.5\text{m}\}$, $1 \le i \le i_r = 5$ are for testing. Each location point consists of 200 test utterances. These are then processed with our proposed methods (i.e. Sec. III and Sec. IV).

VI. ASR RECOGNITION RESULTS

In Table I, the method (A) shows the performance when the reverberant test data are not processed (no dereverberation), together with an acoustic model matched on the test data condition. Method (B) shows the performance using a blind dereverberation scheme that does not require any RTF estimation to carry out dereverberation [11]. The method (C) is the performance when using our previous RTF estimation [1][2]. (D) is the result when using the proposed RTF estimation method discussed in Sec III-A where only the waveform-level compensation is in effect. The method (E) shows the result when both the waveform-level and acoustic model-level compensation are in effect in estimating the RTF (both Sec III-A&B). Lastly, the method in (F) shows the performance when equalization in Sec. IV is implemented on top of the radial compensation method in Sec. III.

Although the method (B) performs better than method (A), it is less effective than methods that use RTF information. We note that (B) operates blindly (no RTF is required). It is confirmed in (D) that the proposed waveform-level RTF estimation outperforms the previous method in (C). This is due to the fact that in the proposed method (Sec III-A),

we are able to address the variation of speech power as a function of the speaker's position. Moreover, the proposed RTF estimator performs best when optimal RTF selector through acoustic model likelihood criterion (Sec III-B) is employed together with the waveform-level compensation (Sec III-A) in (E).

The recognition performance disparity between Room 1 and Room 2 is attributed to the different RT. The latter has larger RTF with more occlusions inside. Also, we emphasize that we experiment on a large vocabulary continuous dictation task and not on isolated word recognition. Unlike the latter, continuous dictation is more susceptible to reverberation due to long-duration utterances which generates more reflections. Moreover, isolated word recognition task has a very small vocabulary (hundred words) and does not consider insertion and deletion errors. In our case, we used 20K-words. Thus, recognition rate is always higher in isolated word recognition task compared to the continuous dictation task. The negative recognition values are attributed to the insertion and deletion errors.

VII. CONCLUSION

We have presented a method that compensates the variation of the speaker's radial position and azimuth relative to the robot, respectively. In the case of compensation the effect of the change in the radial position, we have shown the effect of designing the RTF estimator in conjunction with the ASR system. This results in an RTF estimate that is more tailored to achieving optimal ASR performance. In the case of azimuthal change, we have shown a method based on equalization which further mitigates the effect of mismatch. Currently, these two methods are implemented independently and in the future, we will consider the joint optimization of the two methods.

REFERENCES

[1] R. Gomez, T. Kawahara, "Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer" *In Proceedings of Interspeech*, 2009.

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

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

[4] Y. Huang, J. Benesty, and J. Chen, "Speech acquisition and enhancement in a reverberant, cocktail-party-like environment" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008

[5] H. Kuttruff, "Room Acoustics" *Spon Press*, 2000

[6] 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.

[7] H. Sawada *et al.*,"Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. of ICASSP 2002*, 2002.

[8] 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.

[9] C.J.Leggeter and P.C. Woodland,"Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models" *In Proceedings of Computer Speech and Language*, 1995

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

[12] B. Bank, "Direct Design of Parallel Second-order Filters for Instrument Body Modeling", *In Proceedings of the International Computer Music Conference*, 2007.

[13] J. Laroche and J-L. Meillier, "Multichannel Excitation/Filter Modeling of Percussive Sounds with Application to the Piano" *In Proceedings IEEE Transactions Speech and Audio Processing*, 1994.

[14] B. Bank, G de Poli and L. Sujbert, "A Multi-rate Approach to Instrument Body Modeling for Real-time Sound Synthesis Spplications" *In Proceedings of 112th AES Conventiona*, 2002.

[15] R. Gomez, K. Nakamura and K. Nakadai, "Dereverberation Robust to Speaker's Azimuthal Orientation in Multi-channel Human-Robot Communication" *In Proceedings of the IEEE IROS*, 20013.

[16] R. Gomez, K. Nakamura and K. Nakadai "Hands-free Human-Robot Communication Robust to Speaker's Radial Positition" *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-05

野鳥の歌コミュニケーション理解への試み

Understanding bird communication with songs

鈴木 麗璽
名古屋大学 大学院情報科学研究科
Charles E. Taylor, Martin L. Cody
Dept. Ecology and Evolutionary Biology, University of California, Los Angeles

Abstract

We report on the current state of our research on temporal partitioning to avoid soundscpace overlap by bird communities, based on information theoretical analyses, computational evolutionary experiments, and preliminary recording using HARK.

1 はじめに

野鳥の歌行動の調査と分析を行う我々の研究グループ[1] では，近年生物音響学において発展しているマイクロフォンアレイを用いた観測[1]に基づく野鳥の観測手法の開発と分析[2; 3; 4; 5]や，複雑な構造を持つ歌のフレーズ間の遷移ネットワーク構造の分析[6]，歌のアート作品への応用[7]など，幅広い観点で研究活動を行っている．現在，さらなる展開として，野鳥の歌行動における時間的な相互作用ダイナミクスに注目し，観測と分析を行っている．本稿では，北カリフォルニアにおける野鳥の歌の録音を題材にした時間的重複回避行動の多様性に関する予備的分析と，その適応的意義に関する仮説形成のための進化モデルと分析[8]について紹介する．また，HARK を用いた野鳥の歌の録音と分析に関する予備的検討について報告する．

2 野鳥の歌重複回避行動に関する予備的解析

2.1 野鳥の歌重複回避行動

限られた資源をいかにして効率的に共同利用するかは生物・社会・人工物に遍在する普遍的な問題である．生物においては，資源利用・探索戦略の適応性は最適採餌理論をはじめとして生態学における主要なテーマとして論じられてきた．その中でも，単一の共有資源の同種・異種間での利用の競合を避けるために利用時間が分割される現象（temporal resource partitioning）は，様々な種と時間スケールにおいて観察されている．

現在，我々は野鳥の歌行動における時間的重複回避に注目している．ある種の鳥（鳴禽類）は，近隣個体への縄張りの主張や繁殖期における異性に対するアピールのために，比較的長い鳴き声である歌（もしくはさえずり）を歌う[9]．特に，早朝の森では数多くの鳥が短い時間に同時に歌う傾向があり，夜明けの合唱と呼ばれている．多くの場合，各個体は周期的に歌行動を繰り返すが，その際，近傍の他個体と同時に歌うのを避ける場合があることが知られており[10; 11; 12; 8]，これには自身の歌をより伝わり易くする働きがあると考えられている．例えば，Cody と Brown による先駆的な研究[10]では，Wrentit と Bewick's Wren が早朝に一分毎に歌を歌う頻度を計測したところ，歌う頻度に周期的な変動があり，そのピークが両種で逆位相になっていることが報告された．また，Popp らは，Wood Thrush，Eastern Wood-Pewee，Great Crested Flycatcher，Ovenbird の間で一方の種の歌が他方の直後の歌行動に与える影響をプレイバック実験で調査し，どの種も重複回避の傾向があることや，特に Ovenbird は他種の歌の直後に歌う傾向がある一方で Wood Thrush はその傾向がないことなど，歌毎の短期間の相互作用においても種間に多様な違いがあることを示した[13]．最近では，熱帯雨林など多くの種が共存する環境においても，特に近い周波数帯域の歌を歌う種が多い種間で時間的重複回避が顕著に観察されること[12]などが報告されている．我々は，このような音空間の時分割による効率的な資源利用は，すべてではないにせよ，野鳥において広く存在するであろうと考えている．

2.2 カリフォルニアの野鳥における歌重複回避行動

このような野鳥の群集における重複回避行動がどのような仕組みで実現されているかを明らかにするため，現在，我々の主要な調査地域である北カリフォルニアの野鳥の歌

[1] http://artsci.ucla.edu/birds/

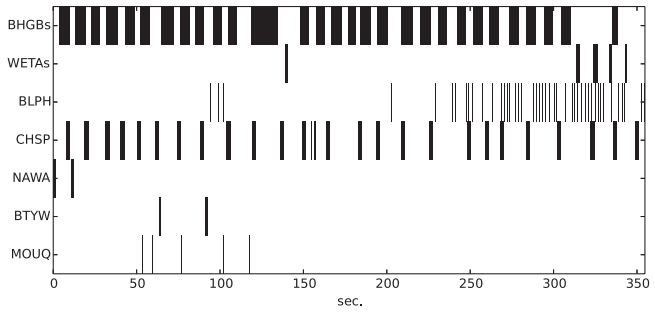


Figure 1: カリフォルニア州アマドール郡での録音における歌行動の時間的ダイナミクス．Black-headed Grosbeak (*Pheucticus melanocephalus*), WETA: Western Tanager (*Piranga ludoviciana*), Black Phoebe (*Sayornis nigricans*), Chipping Sparrow (*Spizella passerina*), Nashville Warbler (*Oreothlypis ruficapilla*), Black-throated Gray Warbler (*Setophaga nigrescens*), Mountain Quail (*Oreortyx pictus*).

重複回避行動の予備的調査と分析を行っている．本稿では，Figure 1 に示す録音を題材に具体例を紹介する．同図は，カリフォルニア州アマドール郡の森林において朝 6 分弱にわたって行われたモノラル録音を分析し，各種がいつどのくらいの長さの歌を歌っているかを可視化したものである．横軸が時間，縦軸のそれぞれが種を表す．各矩形は 1 つの歌[2] に対応し，その個体がその時間に歌っていたことを示している．この録音では全 7 種について各 1 個体ずつの歌行動が観測された．

同図から，種毎に歌行動は大きく異なることがわかる．まず，歌自体に違いがあることがわかる．例えば，BHGB の歌は全体として他の種と比べて長い．これは，BHGB の歌が短いフレーズの組み合わせから構成されており変化に富んだ長い歌であるためである．一方，BLPH はパルスのようなとても短い歌を歌う．また，歌うタイミングにも大きな違いがあることがわかる．BHGB や CHSP は継続的に一定の周期で歌うことを繰り返しているように見えるが，BLPH や BTYW は限られた期間において集中的に歌っているようにも見える．このように，野鳥はその歌や歌い方に様々な特徴があり，種特異的で生得的なものもあれば個体学習によって幼少期に獲得されたり変化し続けるものもある．このような多様な特徴を持つ野鳥の群集において，各種が歌を介してどのように影響し合っているかを明らかにしたいというのが我々の大きな目的である．

そこでまず，Figure 1 の時系列データにおいて全体として重複回避行動が観察されるかどうかについて調べた．

[2] BHGB と WETA については，1 つの歌が短い複数のフレーズから構成されているが，今回の分析では一連のフレーズ全体を 1 つの歌とみなすこととした．

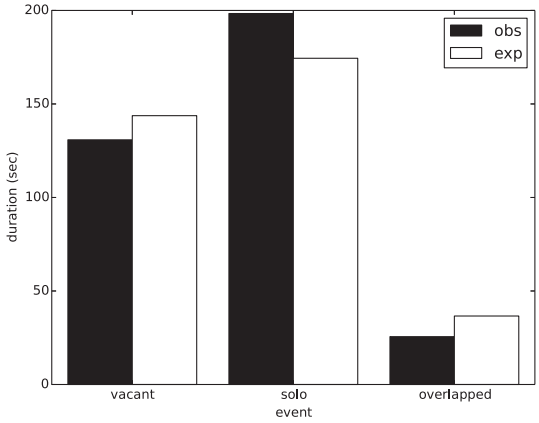


Figure 2: Figure 1 の録音における重複回避行動.

各種が歌っていた総時間と録音全体の長さを用いて，各種がランダムなタイミングで歌ったと仮定したとき，いずれかの種が単独で歌っている時間の期待値を計算することができる．Figure 2 は，この方法で算出した，どの種も歌っていない時間（vacant），1 種のみが歌っている時間（solo），2 種以上が歌っている時間（overlapped）それぞれの期待値と，Figure 1 の実測値との比較を示したものであり，有意な差があることを確認している（カイ二乗検定，χ^2= 7.77, P=0.0206）．同図から，期待値に比べて実測値では一種のみで歌う時間が長く，2 種以上が同時に歌う時間の実測値が短いことがわかり，歌行動の重複回避の傾向が確認できる．つまり，ここに居合わせた 7 種の個体群全体の相互作用の結果として，効率的な音空間の時分割利用が実現されていることが推測される．

2.3 非対称な種間相互作用の情報論的分析

上記のような重複回避は種間のどのような相互の影響の結果，実現されているのだろうか．次に，各種の行動が他種の直近の歌行動に依存しうると仮定し，複数の時系列間の非対称な影響を定量化する指標である移動エントロピー[14]を用いて分析を行った．時系列 $Y_t = \{y_t\}_{t=1,2,\ldots}$ から $X_t = \{x_t\}_{t=1,2,\ldots}$ への移動エントロピー $T_{Y\to X}$ とは，時系列 X_t の時刻 t から m 期間分の最近の状態 $x_t^m = \{x_t, \ldots, x_{x-m-1}\}$ が与えられた時の時刻 $t+1$ での状態 x_{t+1} の不確かさ（＝エントロピー）が，もう一方の時系列 Y_t の l 期間分の最近の状態 $y_t^l = \{y_t, \ldots, x_{y-l-1}\}$ も同時に与えられた際にどれほど減少するかを示したものであり，この意味で，Y_t から X_t に $T_{Y\to X}$ だけ情報の流れがあると表現する．本分析では，算出されたエントロピーの実質的な影響を把握するために，移動エントロピーと，影響元の時系列がランダムであった場合の移動エントロピーの差分である有効移動エントロピー[15]を計算し，前者が後者に対して有意に大きい（t 検定，$P < 0.001$）場

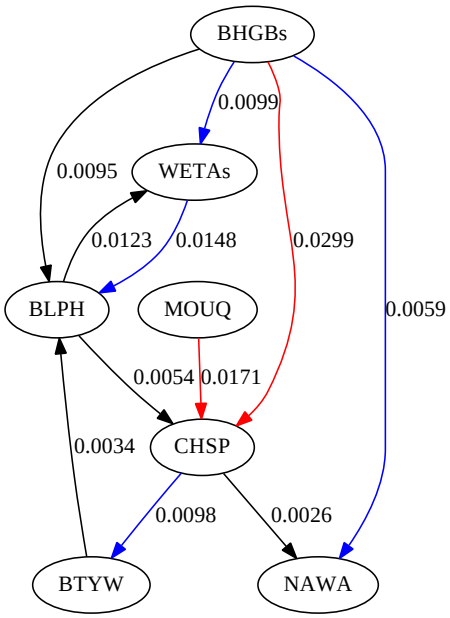


Figure 3: Figure 1 における種間の有効移動エントロピー．

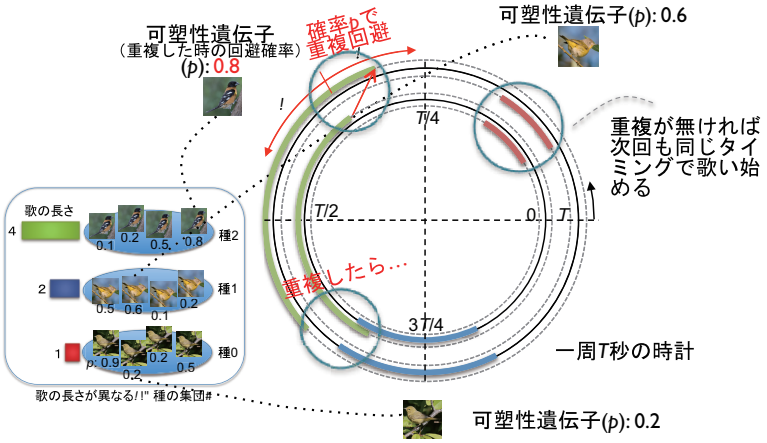


Figure 4: 歌重複回避行動の共進化モデル．歌の長さの異なる複数種の集団から 1 個体ずつ選んで複数のグループをつくる．各グループでは，各個体は基本的には共有する一定の周期で歌行動を繰り返すが，他種の歌との重複に応じて重複回避行動（歌い出すタイミングの微調整）を可塑性遺伝子として持つ回避確率 p で行う（$1-p$ で調整しない）．自身が単独で歌った時間を適応度とし，各種において適応度の高い個体ほど多く次世代に子孫を残す遺伝的アルゴリズムで各種の集団を進化させる．図は重複したとき次回のタイミングを回避確率でわずかにずらす場合の例である．

合にのみ情報の流れがあるものとした[3]．

Figure 3 は，種間の有効移動エントロピーを有向グラフで表現したものである[4]．例えば，BHGB から CHSP への有効移動エントロピーが 0.0299 であることを示している．同図から，有意な影響がある関係は限定されており，また，種間で非対称な情報の流れがあることがわかる．特に，BHGB は 4 種に対してリンクが出ているのに対し，BHGB 自身に向かうリンクはない．つまり，BHGB の挙動は集団全体の挙動に影響を及ぼす一方，BHGB 自体はそれ自身のペースで歌行動を行っていることが推測される．我々は BHGB のように群集全体の振る舞いを決める種を driver species と呼んでいる．また，BHGB から最も大きな情報の流れを受ける CHSP も，BTYW，NAWA の 2 種に影響する中心的な種である事を示唆している．この意味で，ある種の階層的な関係があることも推測される．その一方で，NAWA や BTYW は他種から影響を受けるのみであることもわかる．

最後に，情報の流れのある種間において，それが重複回避にどのように利用されているか検討するため，ある種が歌っている際にもう一つの種が歌い出した観測回数と，ランダムに歌い出した際の期待値を比較したところ，CHSP，BLPH，WETA の 3 種は BHGB の歌との重複が期待値を下回り，既に歌われている BHGB の歌との重複を回避する傾向があることがわかった．BHGB からこの 3 種への有効移動エントロピーが有意に大きいのは，既存 BHGB の歌の回避を反映したものであることが推測される．

いずれも単純な方法であるが，群集内の非対称な相互作用をおおまかに抽出することができた．本解析は単一の録音に基づく一事例であり，また各手法にも改善の余地が多くあるが，歌行動に基づく種間の複雑な相互作用の構造を明らかにするための第一歩と考えている．

3 多様な歌重複回避行動の共進化に関する構成論的モデル

このような種に応じた多様な重複回避の傾向はいくつか報告があり[13]，歌の長さと関係があることも指摘されている[11]．Ficken らは Least flycatcher と Red-eyed vireo の歌の重複度合いをランダムに歌った場合の期待値と比較し，観測データにおいて有意な重複回避が観察されたと報告している [11]．さらに，短い歌を歌う Least fly catcher は重複を積極的に避ける傾向がある一方，相対的に長い歌を歌う Red-eyed vireo は重複を避けない傾向があることを指摘し，その理由として，長い歌は重複しても一部であるため大きな影響はないが，短い歌の場合は歌全体が消されてしまうためにより深刻であることによる可能性を指摘している．

そこで，種特異的な歌の長さと回避傾向の関係を行動可塑性の適応的意義や進化の観点から理解するため，歌の

[3] 計算は各種の歌行動の有無を 1 秒毎に離散化した時系列に対して m=l=2 の条件で行った．
[4] z-score が 2.0 以上のものは青色，3.0 以上のものは赤色で表している．

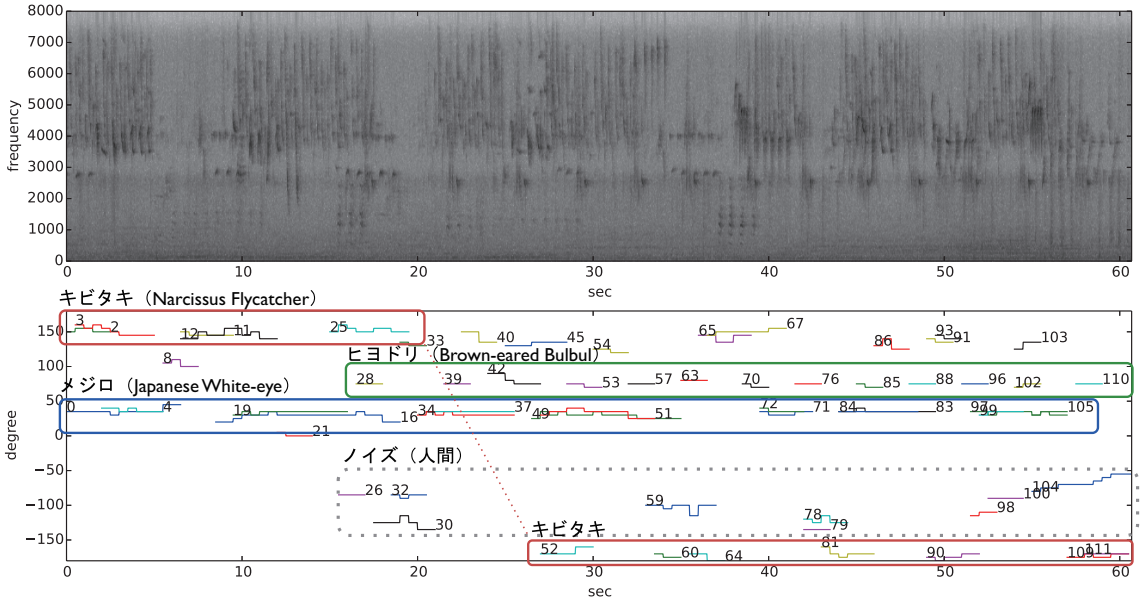


Figure 5: HARK を用いた野鳥の歌の音源定位の例.

長さが異なる複数種からなる集団での歌重複回避行動の共進化モデルを構築した（概要は Figure 4 参照）[8]．1）重複した時次回のタイミングをずらす場合，2）他種がすでに歌っている場合には歌い出しを遅らせる場合，3）直前と直後の他種の歌行動の中間のタイミングになるように次回の歌い出しを調整する場合（DESYNC-TDMA[16]）の 3 種類の重複回避行動をそれぞれ仮定した際，いずれの場合も重複時の回避確率（可塑性）が分化するように進化する傾向があることがわかった．具体的には，長い歌を歌う種は重複をより気にしない driver species へと進化する一方，短い歌を歌う種は積極的に重複を避けるように進化する傾向があった．また，重複回避の多様性が群集全体の情報伝達の効率化に貢献しうることも示唆された（詳細は文献[17]を参照されたい）．このような適応的観点からの知見を生態に関する仮説形成等に活かしたいと考えている．

4 HARK を用いた野鳥の歌の録音と分析に関する予備的検討

以上のような種間相互作用を明らかにするには，様々な環境条件において多数のデータ取得し，分析する必要がある．しかし，Figure 1 のような分析可能な時系列データを作成する際には，従来単一のマイクによる録音に対して手作業（耳作業？）で歌に対してタグ付けを行う場合が多く，コストがかかる．特に，重複回避行動を分析する際にはソナグラム上や聴き分けでの歌の重複の判別に注意が必要であり多大な労力を要する．また，同種が複数存在したり，歌にレパートリーがある場合や，複数の異なるフレーズから構成される場合[6]もあり，その解析にはさらに詳細なタグ付けが必要となる．

このような状況において，京都大学大学院情報学研究科奥乃博教授を中心に開発が進められているロボット聴覚システムである HARK(Honda Research Institute Japan Audition for Robots with Kyoto University)[18]によるマイクロフォンアレイを用いた自動音源定位や分離，認識等の技術を利用可能であれば大きなメリットが得られると考えられる．マイクロフォンアレイやセンサーネットワークを用いたシステムは野鳥に限らず音声コミュニケーションを行う種の生態分析に利用されている[1]が，独自開発であったり高価であったりする場合が多い．我々の研究グループも，センサーノードの開発を含む音源定位システム[2; 3; 4]や，同一種内においての歌に基づく個体の自動判別手法[5]などを開発している．一方，オープンソースで公開されている HARK は，Dev-audio 社の Microcone や Microsoft 社の Kinnect などの安価で入手しやすいデバイスを標準でサポートしており，かつ，音響工学等に関する詳細な知識を必要とせずに音源定位や分離等を含むシステムを柔軟に構築できる統一的枠組みが整備されている．HARK を用いた観測システムの有用性が示されれば，生態観測研究に大きく貢献すると期待できる．

そこで，HARK の音源定位・分離機能の野鳥の歌研究への適用可能性を検討するために，予備的調査を行った．2013 年 5 月の愛知県内の都市公園[5] において，Microcone とノート PC を用いて 7 チャネルの録音を数回試行した．Figure 5 上段は，そこから 1 分間の録音を取り出したソナグラムである．数種の特徴的な歌の構造が繰り返し出現していることがわかるが，いくつか重なり合っていて

[5] 2013 年 5 月 5 日の午前 8 時〜10 時頃，東三河ふるさと公園にて録音.

確認しにくい部分もある．下段は HARK を用いて音源定位を行った結果である[6]．図中の各線が定位された音源であり，その横の番号は定位された順番を表している[7]．

Figure 5 と分離された各音源を予備的に分析した結果，メジロ，キビタキ，ヒヨドリ，コガラなどの複数種の野鳥の歌行動（もしくは地鳴き）が確認できた．同図の矩形はそのうち周期的な歌行動が比較的明瞭に定位されたものを取り出したものである．メジロは 30 度方向で録音を通して長い歌を繰り返し歌っていたが，1 つの歌が 2 つの音源として並列に定位される傾向があった．キビタキは当初 150 度方向で繰り返し歌い，25 秒付近で-170 度方向に移動しているように見えるが，近い方向の他の音源が定位に影響した可能性もある．また，40 秒以降ではそれまでと異なる歌に切り替えたことがわかった．ヒヨドリは 16 秒付近から 70 度方向で短い歌をやや短い周期で鳴いていたが，同種の別個体の鳴き声（67 番，103 番付近）も観測された．一方，コガラの歌はその存在自体は元の録音から比較的明瞭にいくつか確認できるものの，メジロの長い歌の音源中に現れたり，150 度方向の音源に現れることがあり，どちらが正確な方向か判断しにくい状況であった．

以上の様に，今回の設定で得られた定位データを解析に用いるには補正等が必要だが，タグ付けの際にこのような定位情報があれば大変有益であるといえる．また，分離音源に基づく種の分別やクラス分けの自動化なども可能になればさらに有益であるといえる．複数のマイクロホンアレイのノードからなるネットワークを用いて地理的な構造も含んだデータを容易に取得できれば，より正確に歌行動の相互作用ダイナミクスを分析し興味深い知見を得ることが期待できると考えられる．今回は録音後のオフラインでの分析であったが，HARK のリアルタイム処理を活用すれば，近隣個体の歌行動に応じてリアルタイムに歌を再生することでインタラクティブなプレイバック実験も可能であることも示唆されたといえる．今後も調査を進め，歌に基づく野鳥間相互作用構造の分析につなげたいと考えている．

5 Conclusion

本稿では，野鳥の歌コミュニケーション理解への試みとして，野鳥の群集における歌行動の重複回避メカニズム理解のための実データの情報論的分析，構成論的モデルに基づく回避行動の共進化ダイナミクスの検討，および，HARK を用いた野鳥の歌行動の録音と音源定位の予備的調査について紹介した．いずれもまだ初期的段階であるが，このような情報論的・構成論的・生態的アプローチを組み合わせて野鳥の歌行動の生態とその適応的意義を明らかにしていきたいと考えている．

[6] HARK1.9.9 を使用．1K〜10KHz の周波数を用いて歌に対応した音源を多く定位可能なように MUSIC スペクトルと音源定位に関するパラメータ設定をいくつか行い解析した後，多くの場合ノイズであった持続時間 1 秒以下の音源を一意に取り除いた．より詳細な調整により大きな改善が見込まれると考えている．

[7] なお，ソフトウェア上では番号を押すと対応したデータが再生されるようになっている

Acknowledgements

野鳥の歌の録音分析システム試作にあたり，京都大学大学院情報学研究科奥乃博教授を初め HARK 開発チームに助言頂いた．また，録音した種の判別については川島賢治氏（日本野鳥の会，豊田市自然観察の森）に助言頂いた．ここに御礼申し上げる．

参考文献

[1] D. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, Krakauer A. H., C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, and A. N. G. Kirshel. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology*, 48:758–767, 2011.

[2] T. C. Collier, A. N. G. Kirschel, and C. E. Taylor. Acoustic localization of antbirds in a mexican rainforest using a wireless sensor network. *The Journal of the Acoustical Society of America*, 128:182–189, 2010.

[3] S. Cai, T. Collier, L. Girod, R. E. Hudson, K. Yao, C. E. Taylor, and M. Bao. Voxnet acoustic array for multiple bird source separation by beamforming using measured data. In *Proceedings of the 12th international conference on Information processing in sensor networks (IPSN'13)*, pages 355–356, 2013.

[4] Z. Harlow, T. Collier, V. Burkholder, and C. E. Taylor. Acoustic 3d localization of a tropical song bird. In *IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2013.

[5] J. G. Arriaga, G. Kossan, M. L. Cody, E. E. Vallejo, and C. E. Taylor. Acoustic sensor arrays for understanding bird communication. identifying Cassin's Vireos using SVMs and HMMs. In P. Liò, O. Miglino, G. Nicosia, S. Nolfi, and M. Pavone, editors, *Proceedings of the Twelfth European Conference on the Synthesis and Simulation of Living Systems (ECAL2013)*, pages 827–828. MIT Press, 2013.

[6] K. Sasahara, M. L. Cody, D. Cohen, and C. E. Taylor. Structural design principles of complex bird songs: A network-based approach. *PLoS ONE*, 7:e44436, 2012.

[7] R. Suzuki, S. Yamaguchi, M. L. Cody, C. E. Taylor, and T. Arita. iSoundScape: Adaptive walk on a fitness soundscape. In C. Di Chio, editor, *Applications of Evolutionary Computation, LNCS 6625 (Proc. of the 9th European Event on Evolutionary and Biologically Inspired Music, Sound, Art and Design (EvoMUSART2011))*, pages 404–413. Springer, 2011.

[8] R. Suzuki, C. E. Taylor, and M. L. Cody. Soundscape partitioning to increase communication efficiency in bird communities. *Artificial Life and Robotics*, 17(1):30–34, 2012.

[9] C. K. CatchPole and P. J. B. Slater. Bird song: Biological themes and variations, 2008.

[10] M. L. Cody and J. H. Brown. Song asynchrony in neighbouring bird species. *Nature*, 222:778–780, 1969.

[11] R. Ficken and M. Ficken. Temporal pattern shifts to avoid acoustic interference in singing birds. *Science*, 183(4126):762–763, 1974.

[12] R. Planqué and H. Slabbekoorn. Spectral overlap in songs and temporal avoidance in a peruvian bird assemblage. *Ethology*, 114:262–271, 2008.

[13] J. W. Popp, R. W. Ficken, and J. A. Reinartz. Short-term temporal avoidance of interspecific acoustic interference among forest birds. *Auk*, 102:744–748, 1985.

[14] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85:461–464, 2000.

[15] R. Marschinski and H. Kantz. Analysing the information flow between financial time series: An improved estimator for transfer entropy. *The European Physical Journal B*, 30:275–281, 2002.

[16] J. Degesys, I. Rose, A. Patel, and R. Nagpal. DESYNC: Self-organizing desynchronization and TDMA on wireless sensor networks. In *International Conference on Information Processing in Sensor Networks (IPSN)*, pages 11–20. IEEE Press, 2007.

[17] R. Suzuki, C. E. Taylor, and M. L. Cody. Soundscape partitioning to increase communication efficiency in bird communities. In *Proceedings of the 17th International Symposium on Artificial Life and Robotics*, pages 985–988. ALife Robotics Corporation Ltd., 2012.

[18] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system 'HARK' —open source software for listening to three simultaneous speakersh. *Advanced Robotics*, 24:739–761, 2010.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-06

複数のマイクロホンアレイを用いた理科室における音源アクティビティの分析

Analysis of sound source activity in science classes using multiple microphone arrays

○石井カルロス寿憲（ＡＴＲ知能ロボティクス研究所）
Jani Even（ＡＴＲ知能ロボティクス研究所）
塩見昌裕（ＡＴＲ知能ロボティクス研究所）
萩田紀博（ＡＴＲ知能ロボティクス研究所）

* Carlos Toshinori ISHI, Jani EVEN, Masahro SHIOMI, Norihiro HAGITA (Intelligent Robotics and Communication Labs., ATR)

carlos@atr.jp, even@atr.jp, m-shiomi@atr.jp, hagita@atr.jp

Abstract – We are developing a dialogue behavior recognition platform, which is able to detect who is talking, where and when, based on 3D sound direction estimation by multiple microphone arrays, and human tracking technologies. We installed the developed system in a science room of an elementary school, and collected data including real science classes during a period of one month. In the present paper, we present preliminary analysis results on the sound activities of the science room.

1　はじめに

我々は、3次元空間での音源方向と、人位置の推定情報を組み合わせることにより、誰が、いつ、どこでしゃべっているのかを推定する対話行動認識プラットフォームを開発している [1]。

このようなシステムを利用することにより、教室内や会議などのように、複数の人が時に席を移りながら会話や協調作業をする際のデータの観察が容易になることが期待できる。

マイクロホンアレイ処理による音源定位に関する研究はこれまで多くされてきたが[2-6]、マイクロホンアレイを単体で扱うことが多い。その中でも、3次元空間での音源定位に関するものは比較的少ないが、実環境で対象となる音源のアレイに対する仰角が固定できない場合は、方位角のみならず仰角も推定することが重要となる。また、音源とアレイとの距離に関しては、理論上は推定可能であるが、角度推定に比べて精度は低く、処理時間も膨大となってしまう。

また、教室のような広い空間の音源をカバーするためには、その空間の複数の箇所にマイクロホンを配置する必要がある。[5]のように、一つのキャプチャで同期させた96個のマイクロホンを空間内に配置する方法もあるが、コストパフォーマンスの問題も生じる。

上述の問題点を踏まえ、我々は複数のマイクロホンアレイを用いて空間的に情報を統合し、3次元空間で精度よくかつ効率よく音源定位を行う枠組みを提案した[7]。壁や天井での反射の利用も試みてきた[7]。レーザ距離センサも利用し、マイクロホンアレイと組み合わせた枠組みも検討してきた[8]。

本研究では、マイクロホンアレイ処理や人位置検出においてこれまで開発してきたシステムを、小学校の理科室に設置し、実際の理科の授業が行われたデータを収集した。本論文では、システムの紹介と、理科室で観測された音源のアクティビティについて、予備的な分析結果を報告する。

本論文は以下のように構成される。次ぐ2章では、開発したシステムの概要を説明する。3章では、小学校理科室でのデータ収集と分析結果について述べる。4章で考察と今後の課題を記す。

2　開発したシステムの概要

図１に理科室の様子を示す。理科室に机は全部で8つあるが、そのうち実際授業に使用されているのが前方の6つであるため、6つのマイクロホンアレイをこれらの机の上に設置した。それぞれの机に対するアレイの位置は、学校側と相談の上、生徒たちの視界の妨げとならないよう、かつ先生が頭をぶつけないように、2メートル程度の高さに、机の流し台の真上に、天井からマイクロホンアレイを吊るした（図１上部参照）。また、人位置検出に使用するセンサとして、Kinectを多数天井に設置した。

図１．データ収集を行った理科室の様子

図２に開発したシステムの概要図を示す。まず、複数のマイクロホンアレイにおいて、それぞれ3次元空間の音源方向推定（方位角および仰角の推定）を行う。多くの音源定位の研究では、方位角のみが推定されるが、教室のように人の数が多い場合、同じ方向に複数の音源が存在する確率が高くなり、仰角の推定も重要となる。3次元空間における方位角および仰角を求めるため、マイクロホンアレイとして、16個のシリコンマイクが直径30㎝の半球面上に配置するようなアレイフレームを作成した。図３にマイクロホンアレイのマイク位置情報を示す。

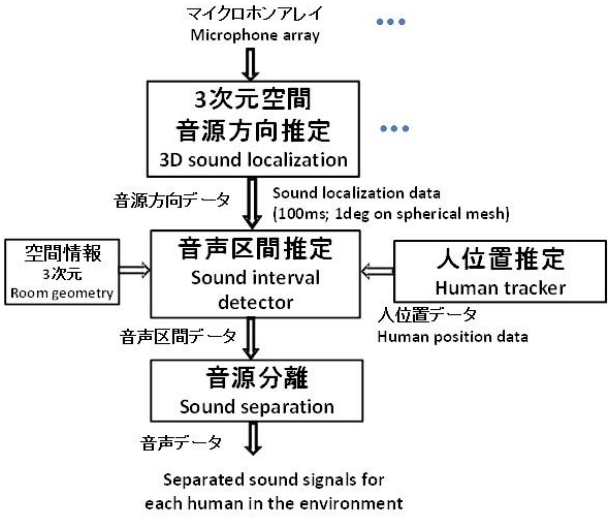


図２．開発したシステムの概要図

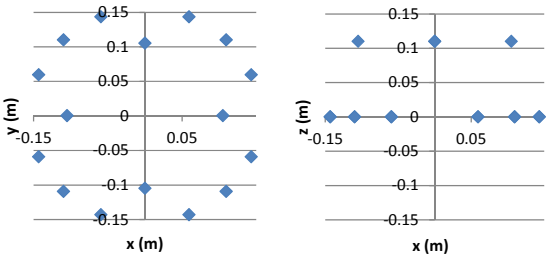


図３．マイクロホンアレイのマイク位置情報

音源方向推定部には、著者らが開発した実時間処理で3次元空間での音源方向を5度の空間的分解能および100msの時間分解能で推定するシステムを用いた[4]。音源方向推定は、空間的分解能が高いMUSIC（Multiple Signal Classification）法に基づいている（付録を参照）。周波数帯域は、アレイの形状を踏まえて、1000～5000Hzを使用している。アレイは2メートルの高さに設置しているため、方位角は0～360度で、仰角は0~-90度とした。実時間処理でMUSIC 法に基づいた3次元空間での音源方向推定を可能にするため、フレーム長を64点(4ms)としている。音源数の推定も難しいため、3に固定して、MUSICスペクトルで2.5dBの閾値を上回ったピークのみを探索している。

人位置検出部には、天井に設置した多数のKinectセンサによる3次元の人位置推定を用いている[9]。レーザ距離センサによる2次元の人位置推定も一つの選択肢であったが[8]、理科室で対象となる生徒の数が多く、センサも天井に設置した方が望ましいため、Kinectセンサによる手法を採用した。

音声区間推定部では、音源方向と人位置情報を基に、その人が発話しているか否かを判断する。部屋の空間情報とアレイの位置情報を基に、それぞれのアレイから得られた音源方向と、人位置推定部から得られる人の位置情報を重ね合わせる。検出された音源方向が、検出された人の口元の位置と重なった場合、その人が発話している確率が高いとみなす。本研究で用いた3次元の人位置検出は、空間内の2次元位置と身長を推定することが可能であるが、身長の推定は比較的精度がよくないため、口元の位置を、子供が座っている場合の80cmから大人が立っている場合の170cmに制限した。

人位置は33~66msごとに推定され、音源方向は100msごとに推定されるため、100msの時間分解能で音声区間が検出できる。

最後に、検出されたそれぞれの音源区間に対し、音源に最も近いマイクロホンアレイを用いて、検出された方向にビームを当て、音源分離を行う[8]。

3　データ収集および分析結果

3.1 データ収集

およそ１ヵ月に渡り（2013 年 2 月）、開発したシステムを用いて理科教室の授業時間を含むデータ収集を行った。各クラスの生徒の数はおよそ 30 名で、先生はクラス担当と理科担当の 2 名である。本論文では、そのうちの 1 日の授業における予備的な分析結果を示す。

図４に 6 つのアレイにより理科室で測定された音源方向推定結果の例を示す。点線は 1 メートル間隔で表示している。それぞれのアレイから出る直線は検出された方向を示し、線が出ていない丸は検出された人位置を表している。左図は、教室の前方で先

生が説明している場面で、右図は、実験中、2 列目と 3 列目の左側の机の生徒が同時に声を発している瞬間を示している。音源方向の線の色は高さ情報を表している。緑は 0 ~ 0.5 m、水色は 0.5 ~ 1.0 m、青は 1.0 ~ 1.5 m、ピンクは 1.5 ~ 2.0 m に対応する。複数のアレイから推定された方向が特定の位置で重なっていることが確認できる。左図ではピンク色で交わり、先生の口元の高さが 1.5m 以上であることに対応している。右図では、いずれも水色と青の境界周辺で線が交わっていることが分かるが、子供が椅子に座った時の口元の高さが 1m 弱であることに対応している。これらの例より、それぞれのアレイによる音源方向推定は、方位角のみならず、仰角も精度よく推定できていることが確認できる。

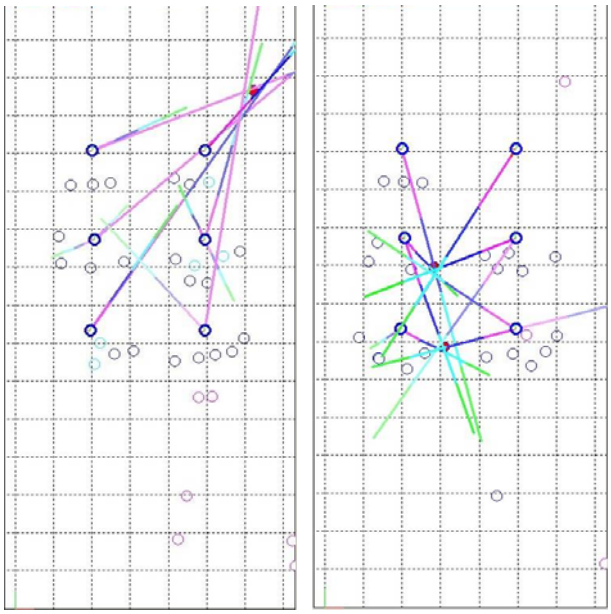


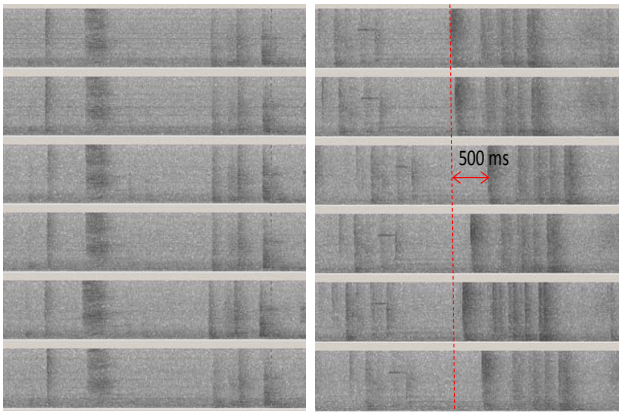
図４．6 つのマイクロホンアレイにより、理科室で測定された音源方向推定結果の例

人位置検出においては、位置検出の精度はそれなりに出ていたが、追跡に失敗することが多く、特定の人と音声発話を対応付けるまでは至っていない。特に生徒達の距離が近くなると追跡が難しく、一旦検出がされず数秒後に再度検出されて別の ID が割り当てられるようなケースが多かった。人位置追跡においては、現在研究開発が進められており、本論文では、アレイデータのみから得られる教室内の音源アクティビティについて分析結果を示す。

3.2 多キャプチャのデータの同期における注意事項

オフライン処理に関する問題点として、多チャンネルオーディオキャプチャデバイスのクロックが異なるため、長時間録音すると、徐々にキャプチャ間で時間ずれが生じることを観測した。

図５にキャプチャ間の時間ずれの例を示す。



(a) 8:50　　(b) 14:20

図 5. 異なった時刻における 6 つのマイクロホンアレイのスペクトログラム（0~6kHz）の例：キャプチャデバイス間のクロックの違いによる時間ずれ

午前 8 時 50 分頃にシステムを起動した際には 6 つのキャプチャのスペクトログラムで突発的な雑音による縦線が揃っていることが分かるが、午後 2 時 20 分頃にシステムを終了した際には、キャプチャ間で最大 500 ms 程度の時間ずれが生じていることが観測された。音源方向推定は 100 ms の分解能であることを踏まえると、この時間ずれは無視できない。オンライン処理では、それぞれのキャプチャからデータが届いた時刻を基に同期を行えば、ネットワーク遅延のみで多キャプチャのデータ同期には比較的影響は小さいが、オフライン処理の場合は、上述のキャプチャ間のクロックの違いにより、時間補正を行う必要がある。

3.3 理科教室の音源アクティビティの分析結果

図６に、６つのアレイにおける音源方向推定結果の例を示す。先生と生徒達が実験についてインタラクションを行っている際の 20 秒間の区間を表示している。各パネルの縦軸は方位角を示し（上半分は 180 ~ 0 度、下半分は 0 ~ -180 度）、色が仰角の違いを表している（赤が-90 ~ -67.5 度、ピンクが-67.5 ~ -45 度、青が-45 ~ -22.5 度、水色が-22.5 ~ 0 度；-90 度が真下方向、0 度が水平方向を差す）。丸は検出された音源方向を表す（時間間隔は 0.1 秒である）。

検出された音源方向において、各パネルの下半分で、仰角を示す色がピンクか青の横線は、各机の周りに座っている生徒達の音源アクティビティに対応している。各パネルの上半分の水色の横線は、教室の前方で先生が発話している区間、または机の前方の音源アクティビティに対応している。この場面では、すべての机の周りで、数名の生徒達が発言していることが分かる。

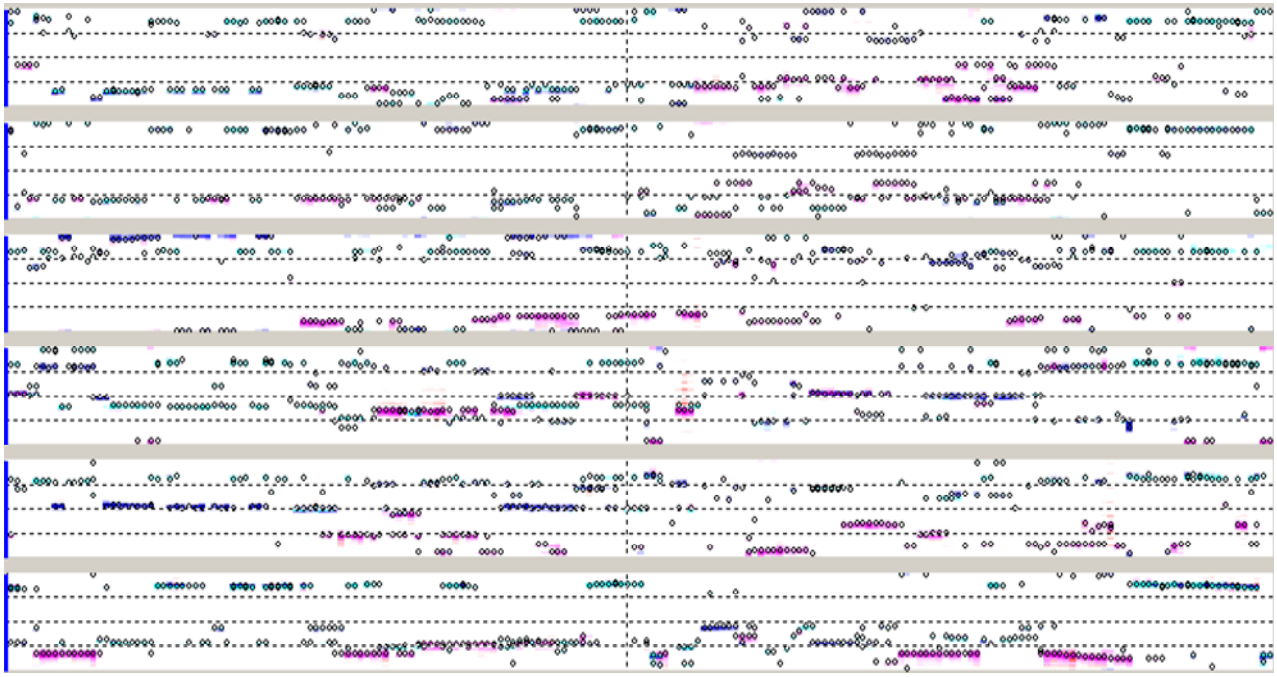


図６　６つのアレイにおける音源方向推定結果の例：先生と生徒達が実験についてインタラクションを行っている場面（20 秒間）

図４に示した例は、ある瞬間（0.1 秒以内）の音源アクティビティを空間的に表示したものであり、図６に示した例は、20 秒間における音源アクティビティの変化を示したものである。しかし、膨大なデータが蓄積された際に、もう少し長いスパンでデータを表示することも重要であると考えられる。

そこで、各机の周りの音源アクティビティのおおまかな流れを観測するため、5 分刻みに特定方向の範囲内に発生している音源アクティビティを集計（定量化）することとした。

音源アクティビティの集計には、0.1 秒ごとに算出される音源方向推定結果を用いて、仰角を-25 度～-85 度の角度領域において、5 分間（300 秒）の区間に対し、対象の方位角の範囲内（10 度間隔）に音源が検出された回数を 0.1 秒で掛ける。また、0.1 秒以内の突発的な音によるもの（足音や机に物を置いたときの音など）は、孤立した点を削除することにより、音源アクティビティの集計から除外している。

仰角においては、0 度が水平方向で-90 度が真下の方向を差すが、-25 度に制限することにより、隣の机の音源アクティビティの影響を避けるようにしている。また、-85 度の制限は、多チャンネルキャプチャの同位相の雑音による誤検出を避けるためであり、アレイの真下方向に位置する流し台周辺の音源アクティビティを観測しないこととなる。

図７に各机のマイクアレイで計測された 1 日分の収録に対する音源アクティビティの時系列ヒストグラムを示す。横軸の時間分解能を 5 分刻みとし、縦軸は方位角で分解能を 10 度刻みとしている。それぞれの時刻と方位角における音源アクティビティの集計秒数を 15 秒刻みで色別に表示している。

アレイの位置および向きにより、方位角が 0～180 度（各パネルの上半分）は、教壇側の音源アクティビティを反映し、-180～0 度（各パネルの下半分）は生徒達が座っている机の周りの音源アクティビティを反映している。

図７には、午前中４クラス（8:50～9:35、9:40～10:25、10:35～11:20、11:25～12:10）、お昼休みを挟んで午後の１クラス（13:05～13:50）を含む音源アクティビティが表示されている。

まず、8:50 までの授業前のアクティビティはすべてのアレイで低いことが分かる。左上のアレイでは、80 度周辺に強いアクティビティを持つ音源が観測されているが、これは教室の左前の角にヒーターが作動し、その定常雑音が観測されたものである。

授業中、教室前方の両アレイで、正の角度（0~180）で 15 秒以上のアクティビティが発している区間が観測できるが、これは先生が教壇周辺で説明をしている時間帯となる。

また授業時間内に、全アレイにおいて、負の角度（-180~0）の領域で 15 秒以上のアクティビティが発している区間が複数観測できる。これは理科の実験中、机の周りの生徒達のアクティビティを反映している。机とクラスによって、アクティビティが高い方向が異なることが分かる。

クラスとクラスの間の休憩時間およびお昼休み時間では、音源アクティビティが低くなっていることが観測できる。またお昼休み時間には右前のアレイで 130 度周辺の方向に強いアクティビティが観測されている。これは校内に流れていた音楽が教室の前方の右側のドアから漏れてきていたことを反映している。

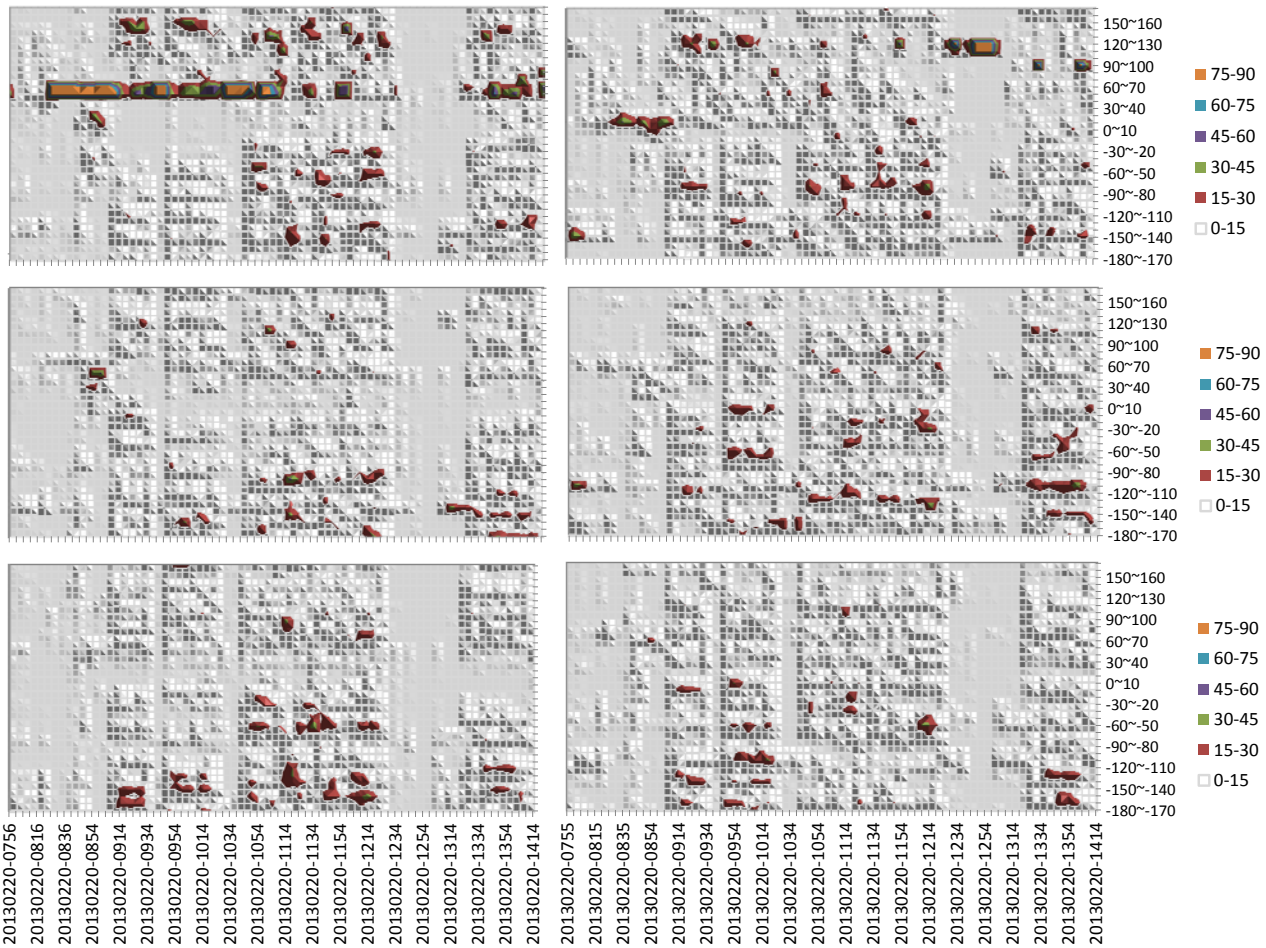


図７．６つのアレイ（上図は教室の前方から１列目の机、中央図は２列目の机、下図は３列目の机）による音源アクティビティの時系列ヒストグラム（横軸：日付-時間（YYYYMMDD-HHMM の形式）を５分刻みで；縦軸：方位角を 10 度刻みで；色別で５分以内の音源アクティビティの時間を 15 秒刻みで表示）

4　考察

本論文では、小学校の理科室の６つの机に設置したマイクロホンアレイによる音源アクティビティの分析を行った。

図７のようなおおまかな音源アクティビティの表示より、理科室内のおおまかな状況が把握可能であり、特定の時間帯におけるより詳細な音源アクティビティの探索が容易となる。また図６のような表示で詳細な音源アクティビティの区間が可能となり、図４のように空間的にどこで音が鳴ったのかが表示できる。似たような音や似たような声では、空間情報がその識別に重要である。

おおまかな音源アクティビティの分析より、クラスと机によって、音源アクティビティが変化することが観測された。例えば、左前の机のように特に目立った音源アクティビティがない机も観測されたが、音声アクティビティの高い生徒をこの机に席替えして、議論を活発化させるなど、クラス活動の助けとして利用できることも考えられる。あるいは、ロボットが先生のお手伝いとして教育現場への活用が可能になれば、アクティビティの低いグループを音環境知能システムが感知し、ロボットが積極的にそのグループに近づいて支援するような用途も考えられる。

現時点では、音源の方向のみに基づき、音声以外の音もアクティビティとして集計されている可能性もあり、机の周りのおおまかな分析に留まっている。しかし、これらの方向と人位置検出が結びつける段階まで研究開発が進めば、先生および生徒達の音声アクティビティが測定可能となる。これは今後の課題となる。

付録：MUSIC 法

M個のマイク入力のフーリエ変換$X_m(k,t)$は、式(1)のようにモデル化される。

$$\mathbf{x}(k,t) = [X_1(k,t),..., X_M(k,t)]^{\mathrm{T}} = \mathbf{A}_k\, \mathrm{s}(k,t) + \mathbf{n}(k,t) \quad (1)$$

ベクトル$\mathrm{s}(k,t)$はN個の音源のスペクトル$S_n(k,t)$から成る：$\mathrm{s}(k,t) = [S_1(k,t), \ldots, S_N(k,t)]^{\mathrm{T}}$ 。kとt はそれぞれ周波数と時間フレームのインデックスを示す。ベクトル$\mathbf{n}(k,t)$は背景雑音を示す。行列 $\mathbf{A}_k$ は変換関数行列であり、(m,n)要素はn番目の音源からm番目のマ

イクロホンへの直接パスの変換関数である。$\mathbf{A}_k$ のn列目のベクトルをn番目の音源の位置ベクトル（steering vector）と呼ぶ。

まず、式(2)で定義される空間相関行列$\mathbf{R}_k$を求め、式(3)に示す$\mathbf{R}_k$の固有値分解により、固有値の対角行列$\mathbf{\Lambda}_k$および固有ベクトルから成る$\mathbf{E}_k$が求められる。

$$\mathbf{R}_k = E[\mathbf{x}(k,t)\mathbf{x}^H(k,t)] \qquad (2)$$

$$\mathbf{R}_k = \mathbf{E}_k\mathbf{\Lambda}_k\mathbf{E}_k^{-1} \qquad (3)$$

固有ベクトルは $\mathbf{E}_k=[\mathbf{E}_k^{s} \mid \mathbf{E}_k^{n}]$ のように分割出来、$\mathbf{E}_k^{s}$と$\mathbf{E}_k^{n}$はそれぞれ支配的なN個の固有値に対応する固有ベクトルと、それ以外の固有ベクトルである。

MUSIC空間スペクトルは式(4)と(5)で求める。rは距離、θとφはそれぞれ方位角と仰角を示す。式(5)は、スキャンされる点 (r,θ,φ) における正規化した位置ベクトルである。

$$P(r,\theta,\varphi,k) = \frac{1}{|\ \tilde{\mathbf{a}}_k^H(r,\theta,\varphi)\mathbf{E}_k^n\ |^2} \qquad (4)$$

$$\tilde{\mathbf{a}}_k(r,\theta,\varphi) = \frac{\mathbf{a}_k(r,\theta,\varphi)}{\left\|\mathbf{a}_k(r,\theta,\varphi)\right\|} \qquad (5)$$

空間スペクトル（本稿ではMUSIC応答と呼ぶ）は、MUSIC空間スペクトルを式(6)のように平均化したものである。

$$\overline{P}(r,\theta,\varphi) = \frac{1}{K}\sum_{k=k_L}^{k_H} P(r,\theta,\varphi,k) \qquad (6)$$

k_Lとk_Hは、周波数帯域の下位と上位の境界のインデックスであり、$K = k_H - k_L + 1$。音源の方位は、MUSIC応答のN個のピークから求められる。

謝辞

本研究は、MEXT 科研費 21118003 及び 21118008 の助成を受けたものである。実験にご協力いただいた京都府精華町立東光小学校の皆様、および実験に参加いただいた児童・保護者の皆様にお礼申し上げる。

参考文献

1) 宮下敬宏, J. Even, P. Heracleous, 石井カルロス, 塩見昌裕, 萩田紀博.「対話行動認識プラットフォームを利用したオーバーラップする発話での話者同定」 日本ロボット学会第30回記念学術講演会講演論文集, RSJ2012, 4M1-4, 2012
2) Y. Sasaki, S. Kagami, H. Mizoguchi, T. Enomoto "A predefined command recognition system using a ceiling microphone array in noisy housing environments," in *Proc. of IROS 2008*, Nice, France, 2008, pp. 2178–2184.
3) K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. of IROS 2009*, St. Louis, USA, 2009, pp. 664–669.
4) C. T. Ishi, O. Chatot, H. Ishiguro, N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. of the 2009 IEEE/RSJ Intl. Conf. on Intelligent Robots and System*, St. Louis, USA, 2009, pp. 2027–2032.
5) H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, Y. Hasegawa, "Real-time sound source orientation estimation using a 96 channel microphone array," in *Proc. of IROS 2009*, St. Louis, USA, pp. 676-683.
6) R. Chakraborty, C. Nadeu, T. Butko, "Detection and positioning of overlapped sounds in a room environment," in *Proc. of Interspeech 2012*, Portland, USA, 2012.
7) C. Ishi, J. Even, N. Hagita, "Using multiple microphone arrays and reflections for 3D localization of sound sources," in *Proc. of IROS 2013*, Tokyo, Japan, 2013
8) J. Even, C. T. Ishi, P. Heracleous, T. Miyashita, N. Hagita: "Combining laser range finders and local steered response power for audio monitoring," Proc. IROS 2012: 986-991, 2012.
9) H. Kidokoro, T. Kanda, D. Brscic, and M. Shiomi, "Will I bother here? - A robot anticipating its influence on pedestrian walking comfort," Proc. HRI2013, 2013

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-07

ガウス回帰に基づく両耳間レベル差の補間

Interpolation of interaural level difference based on Gaussian regression

木元 大輔，尾堂　航，公文　誠
Daisuke KIMOTO，Wataru ODO，Makoto KUMON
熊本大学
Kumamoto University
d.kimoto@ick.mech.kumamoto-u.ac.jp

Abstract

著者らはバイノーラル聴覚ロボットでの規範データを用いた音源定位手法として，音響特徴量の不確かさを考慮した音源定位手法を提案している．この際，規範データ取得に工数がかかるため，これを軽減する方法として疎な収録データを補間して規範データを擬似的に取得する方法が考えられる．このような方法として本報告では補間された点での不確かさを適当なモデルの下で推定するガウス回帰に基づいた補間方法に着目した．この方法より得られる音響特徴量とその不確かさを利用して，不確かさを考慮した音源定位手法を用いた場合，一定条件の下で，定位性能が向上することが確認できた．

1　はじめに

音源定位は，音の情報を利用して周辺環境を認識する上で基本的な聴覚機能であり，ロボットにおいても，ロボット周辺の環境認識を行う方法の 1 つとして有用である．これはロボット聴覚として盛んに研究されている[奥乃, 2010]．その中でも，人間や動物は 2 つの耳で音源位置の情報を得ていることから，2 つのマイクロホンを持つロボットを用いたバイノーラル聴覚ロボットにおける研究が行われている[奥乃, 2002]．バイノーラル聴覚ロボットを用いた音源定位手法の 1 つとしては，事前に学習した音響特徴量を規範データとし，それとマッチングを行うことで，仰伏角及び方位角方向を推定する方法が提案されている[章, 2008]．

人間や動物には，耳に耳介と呼ばれる音の反射・集音を果たす器官が存在している．この耳介の影響により，音の到来方向に応じて音響特性が変化することが知られている[Shaw, 1968]．このことから著者らのグループは，到来方向により音響特性が変化することで音の到来方向が推定しやすくなると考え，マイクロホン近辺に動物の耳介の形状に類似した反射板を取り付けた装置[野田, 2012]を提案してきた．しかしながら，頭部形状による影響は複雑であり，正確にこれをモデル化することは困難で，常に一定の不確かさを考慮する必要がある．著者らはこのことを踏まえ，音響特徴量の不確かさを考慮した音源定位法を提案した[木元, 2013]．この方法を用いることで従来の不確かさを考慮していない音源定位手法に比べ，定位性能が向上することを確認している．

規範データを使用する音源定位手法では，環境ごとに規範データを作成しなければならず，多くの規範データを観測より得ることは困難である．この問題を解決する方法として伝達特性を線形補間する方法が提案されている[中村, 2012]．しかし，この方法では補間より得られる音響特徴量の不確かさは考慮しておらず著者らが提案している不確かさを考慮した音源定位手法を使用することができない．そこで，本研究では音響特徴量の補間方法としてベイズ回帰に基づく方法を提案し，補間より作成した規範データの不確かさも同時に推定し，その検証を行う．

2　バイノーラル聴覚ロボットでの音源定位に用いる特徴量

ロボットに搭載されているマイクロホンに受聴される音信号について考える．環境やロボット自身の影響により，ロボットに搭載されている 2 つのマイクロホンに収録される音は原信号 $s_o(\omega)$ とは異なったものとなる．今，ロボットを基準とした音源位置を $\boldsymbol{x}$ とすると，ある周波数 ω の音源から 2 つのマイクロホンへの伝達関数は $H_l(\omega, \boldsymbol{x})$，$H_r(\omega, \boldsymbol{x})$ と表わすことができる．これより，ロボットに搭載されている 2 つのマイクロホンが受聴する音信号 $s_l(\omega, \boldsymbol{x})$，$s_r(\omega, \boldsymbol{x})$ は Figure 1 に示すような

$$s_l(\omega, \boldsymbol{x}) = H_l(\omega, \boldsymbol{x}) s_o(\omega)$$
$$s_r(\omega, \boldsymbol{x}) = H_r(\omega, \boldsymbol{x}) s_o(\omega) \quad (1)$$

の関係がある．

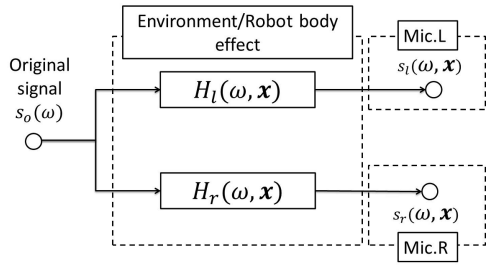


Figure 1: 左右のマイクロホンに受聴される音信号

両耳間レベル差 (ILD) は 2 つのマイクロホン間での音の大きさの違いに相当する．ロボットに搭載されているマイクロホンに受聴される音信号が式 (1) で与えられるとすると，周波数 ω における ILD の値を $Z(\omega, \boldsymbol{x})$ と表すと，

$$Z(\omega, \boldsymbol{x}) \equiv 20 \log |s_l(\omega, \boldsymbol{x})| - 20 \log |s_r(\omega, \boldsymbol{x})|$$
$$= 20 \log |H_l(\omega, \boldsymbol{x}))| - 20 \log |H_r(\omega, \boldsymbol{x})| \quad (2)$$

で表わすことができる．原信号 $s_o(\omega)$ の影響が消去され ILD に影響がないことがわかる．ILD が $\boldsymbol{x}$ ごとに異なっていれば，予め測定した規範データと観測量を比べることで，原信号 $s_o(\omega)$ によらず音源定位に利用できる．

3 両耳間レベル差における不確かさ

式 (2) の ILD の実際の観測量には不確かさが含まれる．著者らの提案する不確かさを考慮した音源定位手法[木元, 2013]ではこの不確かさを適当なガウス分布に従うと考えてきたが，ここではこの仮説の妥当性について検証する．このため本研究ではコルモゴロフースミルノフ検定 (KS 検定) を用いて，観測から得られた ILD がガウス分布に従うと言えるのかを確認する．

まず，帰無仮説を標本が確率密度関数 $F(z)$ から発生するとする．標本が $z_1, z_2, \cdots z_n$ で与えられたとすると，この標本の経験分布は

$$F_n(z) = \frac{1}{n} \sum_{i=1}^{n} R_i(z)$$
$$R_i = \begin{cases} 1 (z_i \leq z) \\ 0 (z_i > z) \end{cases} \quad (3)$$

となる．これより KS 検定統計量は

$$D = \sup_z |F_n(z) - F(z)| \quad (4)$$

で与えられる．ここで，ガウス分布の場合 $F(z)$ は

$$F(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \quad (5)$$

で与えられる．

KS 検定統計量 D の有意確率は

$$\Pr(D\sqrt{n} > \lambda) = 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2j^2\lambda^2} \quad (6)$$

で与えられる．これより，有意水準 5% で与えられた場合には λ　1.36 となり，$D\sqrt{n}$ が 1.36 以上の場合，帰無仮説を棄却し，標本 $z_1, z_2, \cdots z_n$ は確率密度関数 $F(z)$ と一致しないという結論となる．

実際に KS 検定を行った．まず，本研究では Figure 2 に示すマイクロホン近傍に耳介を取り付けたバイノーラル聴覚ロボットを使用している．検定に使用する音は，Figure 3 に示すような環境で収録を行い，Figure 4 に示すような位置にスピーカを設置し収録を行った．音源位置は $x = -0.5$m, $y = 2.0$m を 1, $x = -0.4$m, $y = 2.0$m を 2, ..., $x = 0.5$m, $y = 1.0$m を 121 と番号を割り当てている．対象音としては，白色雑音を使用し 32000Hz でサンプリングを行い，FFT 長 1024 点で処理を行った．また，ILD に周波数方向のフィルターをかけ，平滑化を行った結果を用いている．

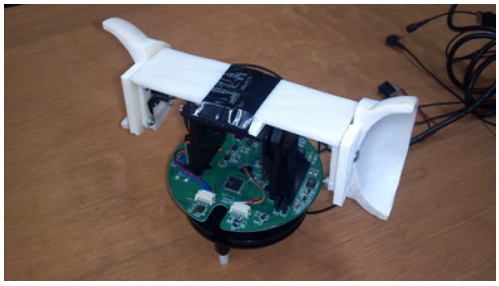


Figure 2: バイノーラル聴覚ロボット

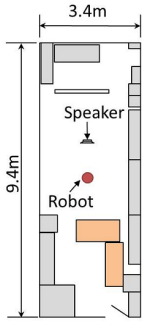


Figure 3: 収録環境

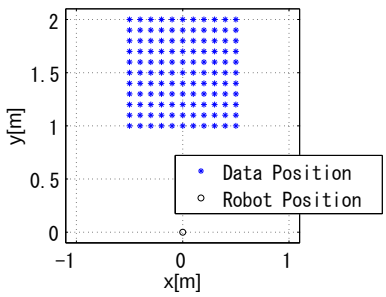


Figure 4: 収録位置

検定を行った結果を Figure 5 に示す．横軸が音源位置を表す番号，縦軸が周波数，色が検定結果を表しており，

黒色となっている部分が帰無仮説を棄却し，ガウス分布ではないとされた部分となる．

本来，帰無仮説が棄却されなかったことから，帰無仮説を採択することはできないが，一点のみでなく 98.26% の音源位置及び周波数帯域でガウス分布ではないとされていないことから，本研究では，ILD はガウス分布に従うとした．

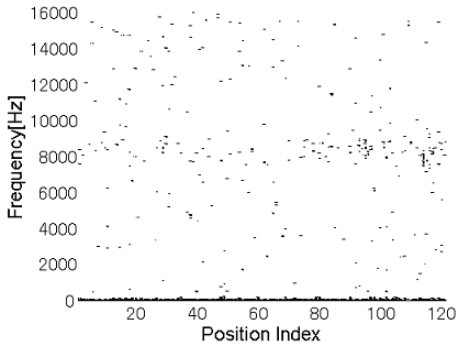


Figure 5: 検定結果

4 両耳間レベル差 (ILD) の補間

この節は，参考文献[ビショップ, 2009]に基づき ILD の補間方法について説明する．

ある周波数での観測される ILD z_n を

$$z_n = Z_n + w_n \tag{7}$$

とする．ここで，$Z_n = Z(\omega, \boldsymbol{x}_n)$ であり位置 $\boldsymbol{x}_n$ で観測される理想的な ILD を表し，$\boldsymbol{x}_n$ は n 番目の観測位置，w_n は n 番目の観測値に含まれるノイズで，独立同分布であると考える．ここで，ノイズはガウス分布に従い

$$p(z_n|Z_n) = \mathcal{N}(z_n|Z_n, \beta^{-1}) \tag{8}$$

であるものとする．また β はノイズの精度を表す超パラメータである．ノイズは各データに対して独立に決まるため，$\boldsymbol{Z}_{1:N} = (Z_1, ..., Z_N)^T$ が与えられた下での ILD $\boldsymbol{z}_{1:N} = (z_1, ..., z_N)^T$ の同時分布は以下の等方的なガウス分布に従う．

$$p(\boldsymbol{z}_{1:N}|\boldsymbol{Z}_{1:N}) = \mathcal{N}(\boldsymbol{z}_{1:N}|\boldsymbol{Z}_{1:N}, \beta^{-1}\boldsymbol{I}_N) \tag{9}$$

ここで，$\boldsymbol{I}_N$ は $N \times N$ の単位行列とする．ガウス過程のモデルとしてカーネル関数を式 (10) と考えるとき

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right) \tag{10}$$

周辺分布 $p(\boldsymbol{Z}_{1:N})$ は，平均が $\boldsymbol{0}$ で共分散が $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ を i, j 要素に持つグラム行列 $\boldsymbol{K}$ で与えられるガウス分布となる．

$$p(\boldsymbol{Z}_{1:N}) = \mathcal{N}(\boldsymbol{Z}_{1:N}|\boldsymbol{0}, \boldsymbol{K}) \tag{11}$$

位置 $\boldsymbol{x}_1, ...\boldsymbol{x}_N$ で条件づけられたときの周辺分布 $p(\boldsymbol{z}_{1:N})$ を求めるためには，$\boldsymbol{Z}_{1:N}$ についての積分が必要であるが，それは以下のように求まる．

$$\begin{aligned} p(\boldsymbol{z}_{1:N}) &= \int p(\boldsymbol{z}_{1:N}|\boldsymbol{Z}_{1:N})p(\boldsymbol{Z}_{1:N})d\boldsymbol{Z}_{1:N} \\ &= \mathcal{N}(\boldsymbol{z}_{1:N}|\boldsymbol{0}, \boldsymbol{C}_{1:N}) \end{aligned} \tag{12}$$

ここで，共分散行列 $\boldsymbol{C}_{1:N}$ は要素

$$C_{1:N}(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \beta^{-1}\delta_{ij} \tag{13}$$

を持つ．

訓練集合として，位置 $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ と対応する $\hat{\boldsymbol{z}}_{1:N} = (\hat{z}_1, ..., \hat{z}_N)^T$ が与えられているときに，新しい位置 $\boldsymbol{x}_{N+1}$ に対する ILD z_{N+1} を予測したいものとする．そのために，予測分布 $p(z_{N+1}|\hat{\boldsymbol{z}}_{1:N})$ を求める必要がある．

条件付き分布 $p(z_{N+1}|\boldsymbol{z}_{1:N})$ を求めるためには，同時分布 $p(\boldsymbol{z}_{1:N+1})$ を書き下す必要がある．ここで，$\boldsymbol{z}_{1:N+1}$ はベクトル $(z_1, ..., z_N, z_{N+1})^T$ を表し，式 (12) から同時分布は

$$p(\boldsymbol{z}_{1:N+1}) = \mathcal{N}(\boldsymbol{z}_{1:N+1}|\boldsymbol{0}, \boldsymbol{C}_{1:N+1}) \tag{14}$$

で与えられる．ここで，$\boldsymbol{C}_{1:N+1}$ は，$(N+1) \times (N+1)$ の共分散行列であり，その要素は式 (13) で与えられる．この同時分布はガウス分布なので，条件付きガウス分布が得られる．これを行うために，次のように共分散行列の分割を行う．

$$\boldsymbol{C}_{1:N+1} = \begin{pmatrix} \boldsymbol{C}_{1:N} & \boldsymbol{k} \\ \boldsymbol{k}^T & c \end{pmatrix} \tag{15}$$

ここで，$\boldsymbol{C}_{1:N}$ は要素が式 (13)$(n, m = 1, ..., N$ に対する) であるような $N \times N$ の共分散行列，$\boldsymbol{k}$は要素 $k(\boldsymbol{x}_n, \boldsymbol{x}_{N+1})(n = 1, ..., N)$ を持つベクトルであるとする．また，スカラー $c = k(\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+1}) + \beta^{-1}$ とする．これらを用いると，条件付き分布 $p(z_{N+1}|\hat{\boldsymbol{z}}_{1:N})$ は，次に示すような平均と共分散を持つようなガウス分布になることが知られている．

$$\begin{aligned} \mu(\boldsymbol{x}_{N+1}) &= \boldsymbol{k}^T\boldsymbol{C}_{1:N}^{-1}\hat{\boldsymbol{z}}_{1:N} \\ \sigma^2(\boldsymbol{x}_{N+1}) &= c - \boldsymbol{k}^T\boldsymbol{C}_{1:N}^{-1}\boldsymbol{k} \end{aligned} \tag{16}$$

式 (16) を用いることで，位置 $\boldsymbol{x}_{N+1}$ の ILD $\tilde{z}(\omega, \boldsymbol{x}_{N+1})$ とその分散 $\tilde{\sigma}^2(\omega, \boldsymbol{x}_{N+1})$ の補間が可能である．

5 不確かさを考慮した音源定位法

不確かさを考慮した音源定位では，予め各音源位置での ILD の平均，分散を取得しておく必要がある．

ある位置 $\boldsymbol{x}$ より得られる周波数 ω の ILD $\tilde{z}(\omega, \boldsymbol{x})$ と分散 $\tilde{\sigma}^2(\omega, \boldsymbol{x})$ が式 (16) より得られる．今，観測により ILD

$z(\omega)$ が得られたとすると，この時，音源が位置 $\boldsymbol{x}$ にある尤度 $l(\omega, \boldsymbol{x})$ を

$$l(\omega, \boldsymbol{x}) = \exp\left[-\frac{\{z(\omega) - \tilde{z}(\omega, \boldsymbol{x})\}^2}{\tilde{\sigma}^2(\omega, \boldsymbol{x})}\right] \tag{17}$$

とする．本研究では，1 つ以上の周波数点において音源が存在すると考えられれば高い尤度を与えるとの考えから，式 (17) の尤度の否定に相当する音源が位置 $\boldsymbol{x}$ にはない尤度 $\overline{l(\omega, \boldsymbol{x})}$ として

$$\overline{l(\omega, \boldsymbol{x})} = 1 - l(\omega, \boldsymbol{x}) \tag{18}$$

を考える．これを用いて，音源定位で使用する周波数帯域全体にわたって音源が位置 $\boldsymbol{x}$ にはない結合尤度 $\overline{l(\boldsymbol{x})}$ を考え，

$$\ln \quad \overline{l(\boldsymbol{x})} = \frac{1}{M}\sum_{i=1}^{M} \ln \overline{l(\omega_i, \boldsymbol{x})} \tag{19}$$

とする．ここで，M は音源定位に使用する周波数点数である．これより，最終的に位置 $\boldsymbol{x}$ に音源がある尤度 $L(\boldsymbol{x})$ を

$$L(\boldsymbol{x}) = \eta\left[1 - \exp\left\{\ln \overline{l(\boldsymbol{x})}\right\}\right] \tag{20}$$

と定めることとする．ここで η は正規化項である．

式 (20) より得られる各音源位置の尤度に閾値 ϵ を考え，$\epsilon \leq L(\boldsymbol{x})$ となった位置 $\boldsymbol{x}$ を音源位置と見做す．

6 検証

6.1 ILD の補間の検証

上述した方法を用いて，実際に ILD の補間を行った，検証用の音は KS 検定で使用した音データと同一のデータを使用した．訓練集合として Figure 4 の位置から 0.2m 四方の位置で取り出した Figure 6 に示す位置のデータを使用した．また本検証では，超パラメータである θ_0，θ_1 は適当な値を与え，β は，EM 法[ビショップ, 2009]を用いて推定を行った．

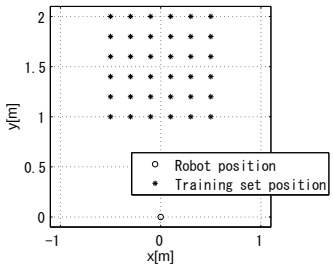


Figure 6: 訓練集合位置

ガウス回帰による補間 (GR 補間) の妥当性について検証を行うために，中村らの方法[中村, 2012]に基づいて ILD を線形補間によって求めたものとの比較を行う．線形補間は，補間したい ILD $\tilde{z}(\omega, \boldsymbol{x}_{N+1})$ が Figure 7 のような位置 $\boldsymbol{x}_{N+1}$ である場合，それを囲む 4 点の ILD を用いて

$$\begin{aligned}\tilde{z}(\omega, \boldsymbol{x}_{N+1}) = (1-\zeta)(1-\xi)z(\omega, \boldsymbol{x}_1) + \zeta(1-\xi)z(\omega, \boldsymbol{x}_2) \\ +(1-\zeta)\xi z(\omega, \boldsymbol{x}_3) + \zeta\xi z(\omega, \boldsymbol{x}_4) \\ (0 \leq \xi, \zeta \leq 1)\end{aligned} \tag{21}$$

のように行う．

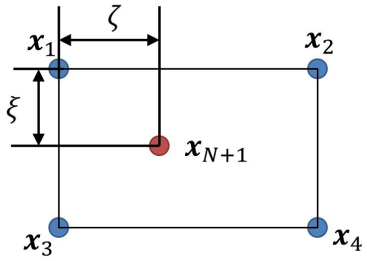


Figure 7: 線形補間概要

実際に補間を行った結果を Figure 8，9 に示す．Figure 8，9 はそれぞれ 500Hz，5000Hz の各位置での ILD を示している．

まず，500Hz での補間を見ると，GR 補間，線形補間ともに実際の ILD の分布と同様な分布が得られていることから，補間が可能であるということがわかる．5000Hz の補間結果を見ると，GR 補間，線形補間は同様な ILD の分布となっているが，実際の分布と比較すると，位置によって細かく変化している ILD の補間は必ずしも完全ではないが，おおよその傾向は再現できていることがわかる．

GR 補間と線形補間を比較するために，各位置で補間した ILD $\tilde{\boldsymbol{z}}(\boldsymbol{x}_i)$ と実際の ILD $\boldsymbol{z}(\boldsymbol{x}_i)$ を式 (22) のように内積を行った．

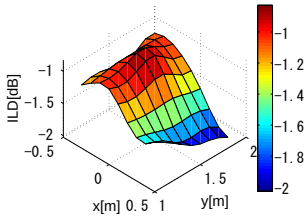
$$\mathrm{IP}_i = \frac{< \tilde{\boldsymbol{z}}(\boldsymbol{x}_i), \boldsymbol{z}(\boldsymbol{x}_i) >}{|\tilde{\boldsymbol{z}}(\boldsymbol{x}_i)||\boldsymbol{z}(\boldsymbol{x}_i)|} \tag{22}$$

その結果，各位置で得られる内積結果の平均が GR 補間で 0.9203，線形補間で 0.9301 となり，どちらの手法を用いても補間の性能にほとんど差がないことがわかった．

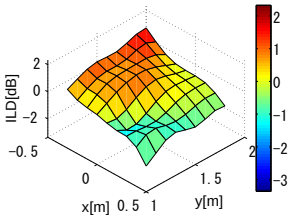
6.2 音源定位結果

補間より得られた ILD を規範データとし実際に音源定位を行った．音源定位には，GR 補間より得られた結果を用い，音源定位は不確かさを考慮した方法 (提案手法) と考慮していない方法で行った．音源定位の対象音としては白色雑音を使用した．また，音源定位には，500〜5000Hz の帯域を使用した．Figure 10 に代表的な定位結果を示す．$*$ が正解の位置を示しており，色で尤度を表す．

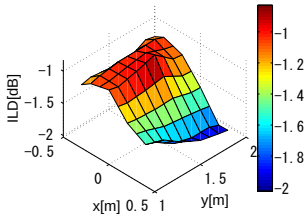
Figure 10(a)，10(b) また 10(c)，10(d)，10(e)，10(f) はそれぞれ提案手法の方がうまく定位が出来ている点，双方の手法で定位が出来ている点，双方の手法で定位が出



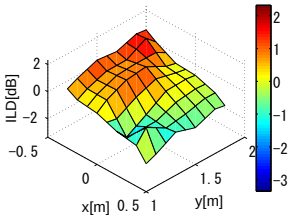
(a) GR 補間



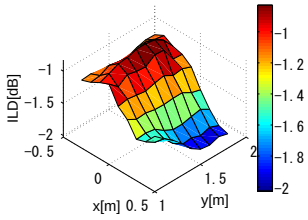
(a) GR 補間



(b) 線形補間

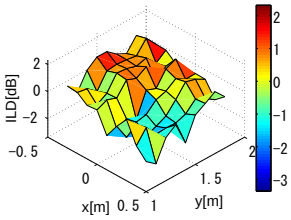


(b) 線形補間



(c) 実際の ILD

Figure 8: ILD の補間結果 (500Hz)



(c) 実際の ILD

Figure 9: ILD の補間結果 (5000Hz)

来ていない点という組合せとなっている．まず，すべての結果を見ると不確かさを考慮した提案手法の方が全体の尤度が高くなっており，これは不確かさを考慮したことによる影響だと考えられる．Figure 10(a)，10(b) を見ると，不確かさを考慮していない手法に比べ，不確かさを考慮した手法の方が正解位置のピークが際立っている形となっている．Figure 10(c)，10(d) を見ると，どちらの手法を用いても正解位置の尤度が高くなっており，音源定位が出来ている結果となっている．Figure 10(e)，10(f) を見るとどちらの手法を用いても正解位置よりも x の位置が-側の尤度が高くなっており，正確な音源定位が出来ていない．補間より得られた ILD と実際の ILD の内積を求めると，Figure 10(a)，10(b) の位置での補間では 0.950，Figure 10(d)，10(e) の位置での補間では 0.925，Figure 10(e)，10(f) の位置での補間では 0.898 となっていることから，Figure 10(e)，10(f) の位置で音源定位が出来なかった原因として ILD の補間自体がうまく出来ていなかったためだと考えられる．

6.3 音源定位性能の評価

定位結果の評価方法として，ROC 曲線[James, 1989] における検出 / 誤り率に基づいた指標を用いる．

ROC 曲線では，尤度に適当な閾値 ϵ を設け，その下で二値判別を行った時の False positive の割合 (FP)，True positive の割合 (TP) を考えるので，音源位置 $\boldsymbol{x}$ での実験データに対して，曲線上の点 $(FP, TP)^T$ は

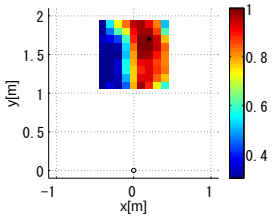
$$(FP, TP)^T = \mathrm{ROC}(\epsilon, \boldsymbol{x}) \tag{23}$$

と表される．$(FP, TP)^T = (0, 1)^T$ が理想的な定位を実現している状態に対応していることから ROC 曲線の値が $(FP, TP)^T = (0, 1)^T$ に近いほど定位性能が良いと考える．

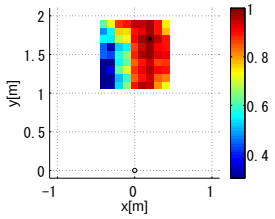
閾値 ϵ を 0.3 から 0.9 まで変化させ ROC 曲線を求めた結果を Figure 11 に示す．Figure 11 の ROC 曲線は，全音源位置で求めた True positive の割合と False positive の割合の平均を使用して求めている．

Figure 11 を見ると，不確かさを考慮した音源定位の方が ROC 曲線の値が $(FP, TP)^T = (0, 1)^T$ に近く，また $\epsilon = 0.9$ の点を比較すると提案手法の方が良好な性能を示していることがわかる．

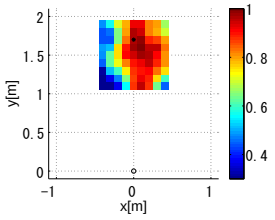
次に，True positive の割合，False positive の割合が最も悪い場合の結果を用いて ROC 曲線を求めた結果を Figure 12 に示す．この結果を見ると，不確かさを考慮した音源定位の方が考慮していないものに比べ $(FP, TP)^T = (0, 1)^T$ に近いことから，不確かさを考慮することで定位性能の低下が抑えられることがわかる．このことから，不確かさを考慮することで定位性能の改善が行えることがわかった．



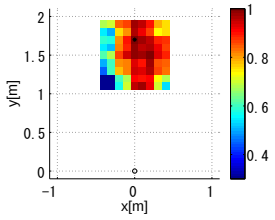
(a) 不確かさ考慮なし
(正解位置 $x = 0.2$m,$y = 1.7$m)



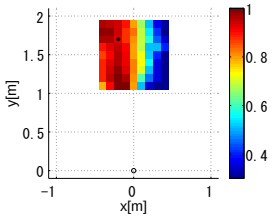
(b) 不確かさ考慮
(正解位置 $x = 0.2$m,$y = 1.7$m)



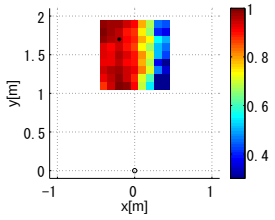
(c) 不確かさ考慮なし
(正解位置 $x = 0.0$m,$y = 1.7$m)



(d) 不確かさ考慮
(正解位置 $x = 0.0$m,$y = 1.7$m)



(e) 不確かさ考慮なし
(正解位置 $x = -0.2$m,$y = 1.7$m)



(f) 不確かさ考慮
(正解位置 $x = -0.2$m,$y = 1.7$m)

Figure 10: 音源定位結果

7　まとめ

音響特徴量の補間の手法として，ガウス回帰に基づく手法を使用し，補間点の不確かさを考慮した．規範データとして，補間より得られた音響特徴量を使用し，対象音を白色雑音とし音源定位を行った場合，音源定位性能が向上することを確認することができた．

音楽，音声での音源定位は今後の課題である．

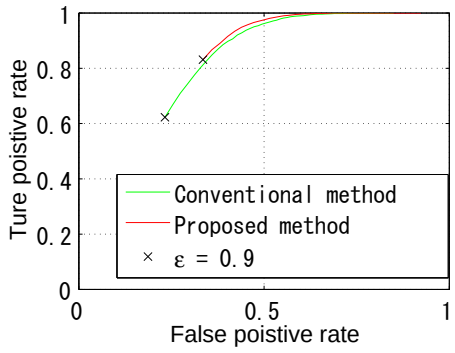


Figure 11: ROC 曲線

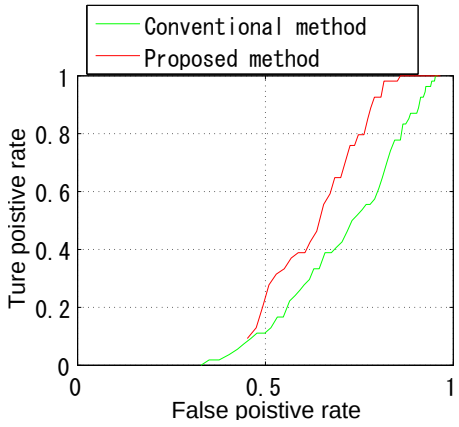


Figure 12: ROC 曲線 (最悪値)

参考文献

[奥乃, 2010] 奥乃　博: ロボット聴覚の現状と展望, 日本ロボット学会誌, vol.28, no.1, pp.2-5, 2010.

[奥乃, 2002] 奥乃　博, 中臺　一博: ロボットの耳は二つで十分か (< 小特集 > なぜ耳は二つあるか?), 日本音響学会誌, vol.58, no.3, pp.205-210, 2002.

[章, 2008] 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.

[Shaw, 1968] Shaw E.A.G., Teranishi R.: Sound pressure generated in an external-ear replica and real human ears by a nearby point source, J.Acoust.Soc.Am., vol.44, pp.240-249, 1968.

[野田, 2012] 野田　佳孝, 公文　誠: 2 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), pp.1643-1646, 2012.

[木元, 2013] 木元　大輔, 尾堂　航, 公文　誠: 観測データの不確かさを考慮したバイノーラル聴覚ロボットでの音源定位手法, 第 31 回日本ロボット学会学術講演会, RSJ2013AC3D3-05, 2013.

[中村, 2012] 中村　圭介, 中臺　一博：ハイブリッド伝達関数補間法とそのロボット聴覚システムへの応用, 日本ロボット学会第 30 回記念学術講演会, 3D1-5, 2012.

[ビショップ, 2009] C.M. ビショップ: パターン認識と機械学習 (上, 下) ベイズ理論による統計的予測, シュプリンガー・ジャパン, 2009

[James, 1989] James A.Hanley: Receiver operating characteristic(ROC) methodology:The state of the art, Crit.Rev.Diagn.Imaging, vol.29, Issue3, pp.307-335, 1989.

Combining Steered Response Power with 3D LIDAR scans for building sound maps

Jani Even, Yoichi Morales, Jonas Furrer, Carlos Toshinori Ishi, Norihiro Hagita

***Abstract*— This paper presents a framework for building 3D map of sounds. The environment is scanned by using a mobile platform equipped with a microphone array and a 3D LIDAR. A steered response power algorithm gives an angular distribution of the sound power at the mobile platform's position. This angular distribution is combined with the distances estimated by the 3D LIDAR in order to generate the spatial distribution of the sound's power. The fusion of the successive measurements obtained while the platform explores the environment results in the creation of the 3D sound map.**

I. INTRODUCTION

In acoustical signal processing, knowing the locations from which sounds are emitted is a very important task referred to as sound source localization (see [5]). Steered response power (SRP) algorithms are among the most effective methods that have been proposed [5], [4]. In particular, the SRP with PHase Transform (SRP-PHAT) [4] is well suited for robotic applications [1].

When the microphone array used for acquiring the audio data fed to the SRP algorithm is mounted on a mobile robot, the operational range of sound localization is extended as the robot can explore the environment. A natural framework for using the robot's mobility for sound source localization is to use a conventional sound source localization algorithm at different locations and combine the results from all these different locations [14], [8], [10], [11], [7].

In this paper, we present a framework for building a 3D map of sound using an autonomous mobile robot equipped with a microphone array and a 3D LIDAR (see Fig.1). The resulting 3D maps, referred to as *sound maps* in the remainder, can be exploited for sound source localization. The proposed method is a multi-modal approach that combines the bearing and power estimates from the SRP algorithm with the range estimates given by the 3D LIDAR. The 3D map is a 3D grid of voxels (3D cubes) that fill the space. Each of the voxel contains the information about the presence of an object at its location but it also contains the probability that this object emits sound. In order to build a precise map by fusing audio and LIDAR data, the platform has to localize itself in the environment. It is possible to proceed to local maxima search on the 3D grid in order to find the locations of the sound sources in the environment.

This research was funded by the Ministry of Internal Affairs and Communications of Japan under the Strategic Information and Communications R&D Promotion Programme (SCOPE).

The authors are with ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan. even at atr.jp

II. MAP BUILDING

To precisely localize itself in the environment, the mobile robot requires a map describing the environment, referred to as the *geometric map*.

The geometric map is built in advance using the 3D Toolkit library framework [3], [12]. To build the geometric maps, a mobile platform is driven through the environment. During this drive, the wheel encoders provide odometry data and the 3D Lidar provides scan data. Then the scans are aligned by correcting the trajectory of the platform using iterative closest point based simultaneous location and mapping (SLAM) [2]. Rather than using the aligned point cloud, an octree representation of the environment is created [6]. The geometric map refers to the voxels at the lowest level that are occupied (the edge length of the voxels composing the geometric map is 0.05 m). In Fig.2, a view of an indoors environment and the corresponding view in the associated geometric map are juxtaposed. Note that the voxels that compose the octree are clearly visible.

A. ROBOT LOCALIZATION

In this paper, it is assumed that the ground is flat and that the platform's pitch and roll are negligible. Consequently, the pose of the platform is composed of its 2D location $\{x_r(t), y_r(t)\}$ and its orientation $\theta_r(t)$. The altitude is assumed constant $z_r(t) = z_0$ and the pitch and roll null $\{\phi_r(t) = 0, \gamma_r(t) = 0\}$.

Since the localization is reduced to a 2D problem, laser range finders (LRFs) scanning in the horizontal plane at a height h_{LRF} are used to localize the mobile platform in a 2D map. The 2D map is created by taking an horizontal slice of the geometric map at the height $\{h_{\text{LRF}} - \epsilon, h_{\text{LRF}} + \epsilon\}$ and flattening it. Then the referential in the 2D map coincide perfectly with the one in the geometric map. Fig.3 gives the naming conventions for the pose $\{x_r(t), y_r(t), \theta_r(t)\}$ in the referential of the 2D map. The green arrows shows the

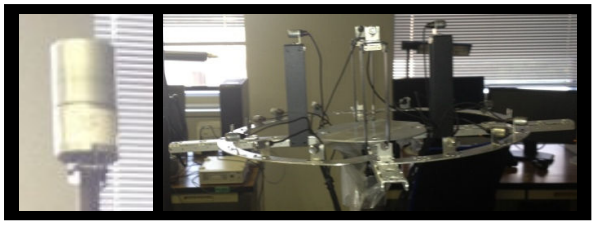


Fig. 1. 3D LIDAR (left) and microphone array (right).

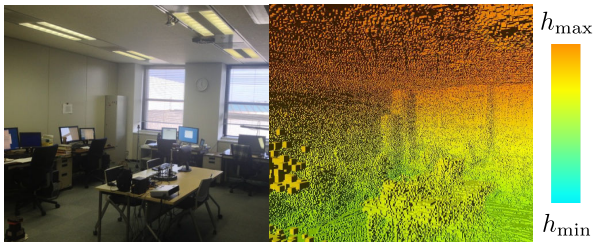


Fig. 2. Photo of the indoor environment and the corresponding view in the geometric map.

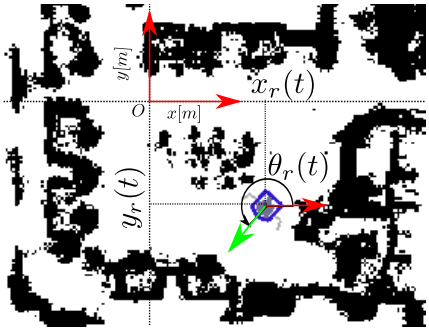


Fig. 3. Mobile platform localization in the $2D$ map created from the geometric map.

orientation of the mobile platform. This map correspond to the environment depicted in Fig.(2).

The localization algorithm is a particle filter (see [13] and references herein). In the prediction step, the particles are propagated accordingly to the odometry data. In the correction step, the likelihood of the particle is computed by using the ray tracing approach to match the LRFs scan to the 2D map. Resampling is performed when the number of effective particles is too low.

The number of particles is 200 and correction is performed when the platform moved by 0.1 m or rotated by 5 degrees.

III. STEERED RESPONSE POWER

In this paper, we use an SRP-PHAT algorithm to process the signals from the microphone array mounted on the mobile platform. The audio processing is done in the frequency domain. The frequency domain signals are denoted by $X_n(f,t)$ where n is the microphone index, f the frequency bin index and t the frame index. They are obtained by applying a short time Fourier transform (STFT) to the audio signals. The analysis window is W points long and the shift of the window is $W/2$.

First the PHAT transform is applied to the frequency components

$$V_n(f,t) \quad = \quad \frac{X_n(f,t)}{|X_n(f,t)|}. \qquad (1)$$

Then the power of the received sound is estimated for a set of candidate directions $\{\theta_i,\phi_i\}_{i\in[1,I]}$. The green dots in Fig.4 represent a set of candidate directions.

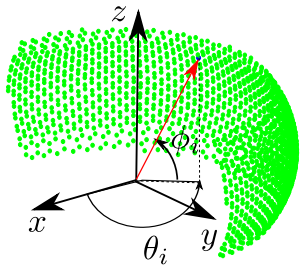


Fig. 4. Set of candidate directions (green dots) and conventions for the angles $\{\theta,\phi\}$ in the array referential.

For each of the candidate directions, the frequency domain processing is decomposed in 3 stages. First the response is steered in the candidate direction $\{\theta_i,\phi_i\}$ by applying a delay and sum spatial filter

$$Y(\theta_i,\phi_i,f,t) \quad = \quad H(\theta_i,\phi_i,f)\begin{bmatrix} V_1(f,t) \\ \vdots \\ V_N(f,t) \end{bmatrix}, \qquad (2)$$

with

$$H(\theta_i,\phi_i,f) = \frac{1}{N}\left[e^{-j\tau_1(\theta_i,\phi_i,f)},\cdots,e^{-j\tau_N(\theta_i,\phi_i,f)}\right], \quad (3)$$

where $\tau_n(\theta_i,\phi_i,f)$ is the phase delay at the microphone n in the frequency bin f for a signal coming from the direction $\{\theta_i,\phi_i\}$. Assuming that the sound sources are in the far field, the filter is entirely characterized by the angles $\{\theta_i,\phi_i\}$ and the microphone positions.

Then the power of the beamformer output is estimated by a K frame averaging

$$S(\theta_i,\phi_i,f,k) \quad = \quad \frac{1}{K}\sum_{t=0}^{K-1}|Y(\theta_i,\phi_i,f,k-t)|^2. \quad (4)$$

Note the introduction of the index k to show that the power has a different rate (the period is $KW/2$ samples).

Finally, the steered response power in the direction $\{\theta_i,\phi_i\}$ is obtained by selecting a limited band of frequencies

$$S(\theta_i,\phi_i,k) \quad = \quad \sum_{f=f_{\min}}^{f_{\max}} S(\theta_i,\phi_i,f,k). \qquad (5)$$

In the remainder, the term *audio scan* refers to the set of candidate directions $\{\theta_i,\phi_i\}_{i\in[1,I]}$ and their associated power $S(\theta_i,\phi_i,k)$ computed at a given frame k. The kth audio scan is denoted by $\mathcal{S}(k) = \{S(\theta_1,\phi_1,k),\cdots,S(\theta_I,\phi_I,k)\}$.

An audio scan is represented as a colored portion of a sphere in Fig.5 (left). The color is function of the power for each of the candidate directions. This audio scan clearly exhibits an area of higher power on the top left side indicating the presence of a sound source.

IV. AUDIO LIKELIHOOD

In order to fuse the sound source localization results of different audio scans together, the power $S(\theta_i,\phi_i,k)$ is

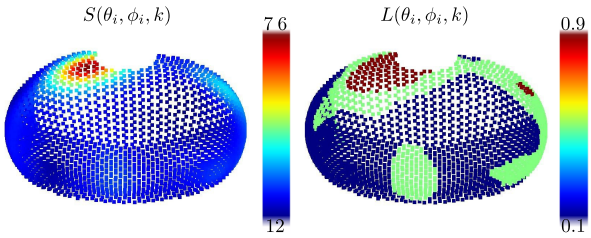


Fig. 5. Transformation by the thresholding function of the power (left) into a likelihood (right).

first transformed in a likelihood $L(\theta_i, \phi_i, k)$. This likelihood $L(\theta_i, \phi_i, k)$ expresses the belief of having a sound source in the candidate direction $\{\theta_i, \phi_i\}$.

From the sound source localization literature and the idea behind the SRP approach, it is expected that a large power should correspond to a strong belief. For example in [9], [7] a scale version of the power was used as likelihood. In audio source tracking, creating a likelihood by scaling the power is also common [15].

In this paper, rather than using a scaled power, a nonlinear function is applied in order to create the likelihood. The selected nonlinear function is a double thresholding function

$$F(x) = \begin{cases} p_{\text{min}}, & \text{if } x < T_1 \\ p_{\text{max}}, & \text{if } x > T_2 \\ p_{\text{med}}, & \text{else} \end{cases} \quad (6)$$

Fig.5 shows the transformation of an audio scan with the nonlinear function (the parameters are set to $T_1 = 20$, $T_2 = 30$, $p_{\text{min}} = 0.1$, $p_{\text{med}} = 0.5$, $p_{\text{max}} = 0.9$).

Consequently while the mobile platform is navigating the environment for each audio scan $\mathcal{S}(k) = \{S(\theta_1, \phi_1, k), \cdots, S(\theta_I, \phi_I, k)\}$ a *likelihood scan* $L(k) = \{L(\theta_1, \phi_1, k), \cdots, L(\theta_I, \phi_I, k)\}$ is create by applying the nonlinear function. Each of these likelihood scans contains the likelihood of having a sound source for one of the candidate directions $\{\theta_i, \phi_i\}$.

V. AUDIO MAP BUILDING

To understand the creation of the sound map, let us first discuss about the structure used to store the audio information. The sound map is an octree representation [6]. At the finest level of decomposition, the edge length of the voxels is 0.05 m and the voxels centered at the position $\{x, y, z\}$ is denoted by c_{xyz}. The voxels of the sound map have some fields to store the audio information:

- $\mathcal{L}(c_{xyz})$ denotes the log-odds of having a sound source within the voxel c_{xyz},
- $\mathcal{M}(c_{xyz})$ counts the number of times the voxel c_{xyz} was updated during sound map creation,
- $\mathcal{U}(c_{xyz})$ contains the last time the voxel c_{xyz} was updated.

Initially, all the voxels at the lowest level in the sound map are considered not occupied.

The candidate directions $\{\theta_i, \phi_i\}$ are defined in the referential centered at the microphone array depicted in Fig.4. This referential is rigidly attached to the mobile platform. The axis directions coincide with the platform's ones but the array origin is at the position $O_a = (x_a, y_a, z_a)_r$ (the subscript r denotes coordinate in the robot's frame).

A scan of the 3D LIDAR is composed of Q points $M_j = (x_j, y_j, z_j)_r$ in the robot's frame. The range in the direction $\{\theta_i, \phi_i\}$ is obtained by finding the point M_j the closest to that direction. For this purpose, let us define the audio direction

$$\vec{v}_a(i) = \begin{bmatrix} \cos(\theta_i)\cos(\phi_i) \\ \sin(\theta_i)\cos(\phi_i) \\ \sin(\phi_i) \end{bmatrix},$$

and the LIDAR direction

$$\vec{v}_L(j) = \frac{\overrightarrow{M_j O_a}}{|\overrightarrow{M_j O_a}|}.$$

Then the index j_i of the closest point M_j to the direction i is selected by finding

$$j_i = \text{argmin}_j \; 1 - \vec{v}_L(j) \cdot \vec{v}_a(i).$$

The point is considered valid if $1 - \vec{v}_L(j) \cdot \vec{v}_a(i) < \epsilon$, where ϵ is a small threshold, and the range associated to $\{\theta_i, \phi_i\}$ is $\rho_i = |\overrightarrow{M_j O_a}|$.

Then to relate the likelihood $L(\theta_i, \phi_i, k)$ to a geometric structure in the environment, the candidate direction has to be combined with the estimated pose of the platform. For the likelihood scan $L(k)$, the pose $\{x_r(t), y_r(t), \theta_r(t)\}$ of the platform with t the closest to k is considered.

This combination is illustrated in Fig.6. For simplicity, a top view is presented and the elevation angle ϕ_i is omitted. The circles represents the points M_j of the LIDAR scan.

For each of the candidate direction, a ray is casted from the array origin O_a in the referential of the sound map. Namely a ray of length ρ_i is casted from the point $(x_a(t), y_a(t), z_a(t))_w$ in the direction $\{\theta_r(t) + \theta_i, \phi_i\}$. Note that the coordinate of the array origin is a function of t in the world frame (denoted by subscript w) as the robot moves. Thus the end point falls in a voxel c_{xyz} of the sound map. Then the likelihood $L(\theta_i, \phi_i, k)$ is used to update the log-odds of having a sound source in this voxel.

The rationale behind the use of ray casting is to trace back the sound until its sources as in [7]. Contrary to [7], in this paper, the range of the sound source is given by the 3D LIDAR and not estimated from the position in the geometric map.

In practice, the ray casting is limited to a maximum range R_{max} as sound intensity decreases rapidly with the distance.

The audio related fields of the voxel are updated as follows

$$\begin{aligned} \mathcal{L}(c_{xyz}) &= \mathcal{L}(c_{xyz}) + \log \frac{L(\theta_i, \phi_i, k)}{1 - L(\theta_i, \phi_i, k)} \\ \mathcal{M}(c_{xyz}) &= \mathcal{M}(c_{xyz}) + 1 \\ \mathcal{U}(c_{xyz}) &= t_k, \end{aligned}$$

where t_k is the time corresponding to the frame k. At initialization $\mathcal{L}(c_{xyz}) = 0$, $\mathcal{M}(c_{xyz}) = 0$ and $\mathcal{U}(c_{xyz})$ is undetermined. The choice $\mathcal{L}(c_{xyz}) = 0$ means that a voxel has equal chance to emit or not sound.

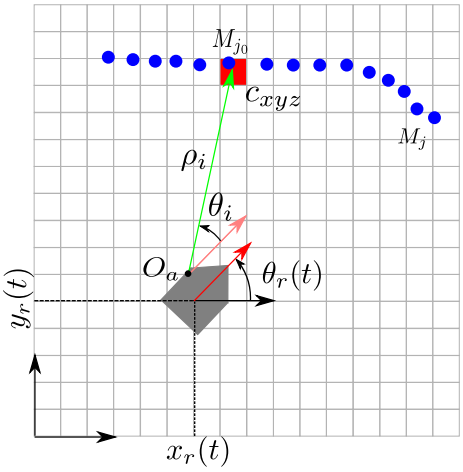


Fig. 6. Ray casting from the mobile platform pose $\{x_r(t), y_r(t), \theta_r(t)\}$ in the direction $\{\theta_i, \phi_i\}$ with a range ρ_i that falls in the voxel c_{xyz} (in red).

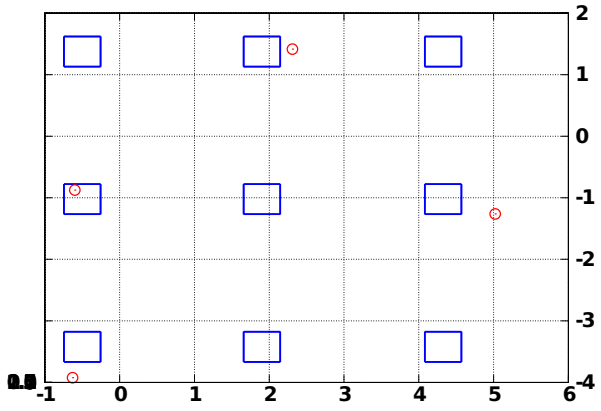


Fig. 7. Top view of the grills placement in blue and the estimated positions in red.

The log-odds $\mathcal{L}(c_{xyz})$ is no longer updated when it goes out of the interval $[\mathcal{L}_{\min}, \mathcal{L}_{\max}]$. Meaning that the odds of having a sound source at the voxel c_{xyz} is considered high or low enough to stop updating it.

The voxels having a count $\mathcal{M}(c_{xyz}) > \epsilon_C$ are considered occupied.

VI. EXPERIMENTAL RESULTS

This part reports the results of the experiments conducted to detect sound sources by using the sound map framework.

The experimental setting corresponds to the indoors environment depicted in Fig.2. At the time of the experiments, the sound sources in this environment are the grills of the air conditioning system. The grills are in the ceiling of the room and have a square shape (0.5 m edge). Fig.7 shows a top view of the room with the grills in blue.

To build the sound map, the mobile platform was driven three time around the table in the center of the room in a clockwise manner (see the 2D map in Fig.3). The parameters of the methods are set to $W = 400$, $K = 10$, $f_{\min} = 1000$ Hz, $f_{\min} = 3000$ Hz, $\mathcal{L}_{\min} = -40$ and $\mathcal{L}_{\max} = 40$. The

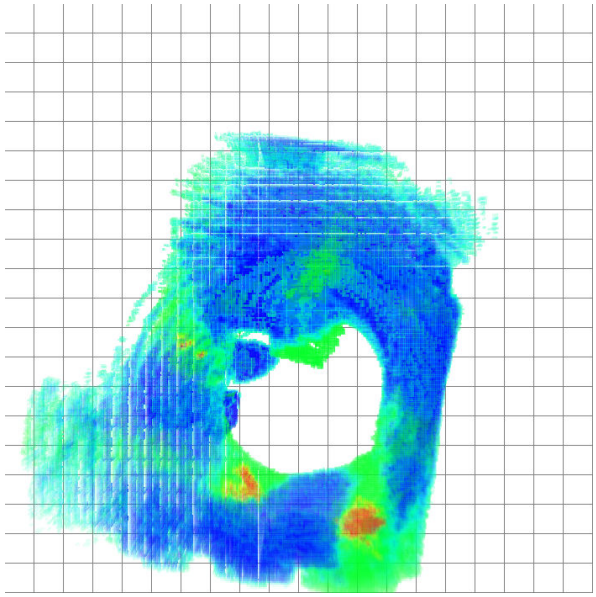


Fig. 8. Top view of the 3D sound map, the color represents the probability of sound source presence (ϵ_C=0).

parameters of the nonlinear function are set to $T_1 = 20$, $T_2 = 30$, $p_{\min} = 0.1$, $p_{\text{med}} = 0.5$, $p_{\max} = 0.999$. The angles of the candidate locations for the SRP algorithm are limited to $\theta \in [-45, 15]$ and $\phi \in [075]$. The maximum range is set to $R_{\max} = 6$ m.

Fig.**??** shows part of the sound map creation while the robot is moving. Fig.8 shows the top view of the sound map generated. Areas of high log-odds are visible around the location of the grills. The localization of the sound sources is estimated by clustering the voxels with positive log-odds (probability of having a sound source larger than 0.5). The clustering method is a kmeans method seeded with the positions of the local maxima of the log-odds. Each obtained cluster is assigned to the closest grill, then that grill is marked as detected and the distance to this grill is computed. In fig.7, the red circles indicate the positions of the detected sound sources (note that only a few sources are detected).

The sound source detection is evaluated in term of localization error E for these detected sources. The average error is 0.49 m with a standard deviation of 0.22 m. The errors are relative to the centers of the grills that have a 0.5 m edge. As a comparison, the maps presented in [8], [7] exhibit localization errors in the 0.2∼0.3 m range for the 2D case.

The undetected grills are the ones that were not for a long time in the aperture of the SRP while the platform made three loops around the table in the center of the room. The threshold ϵ_C for the count of the number of time a voxel was updated affects the number of occupied voxels. By plotting only the voxels c_{xyz} of the sound map such that $\mathcal{M}(c_{xyz}) > \epsilon_C$, it is possible to refine the map as illustrated in Fig. 10. The delimitation of the map and the areas of higher likelihood appear more clearly when the voxels with

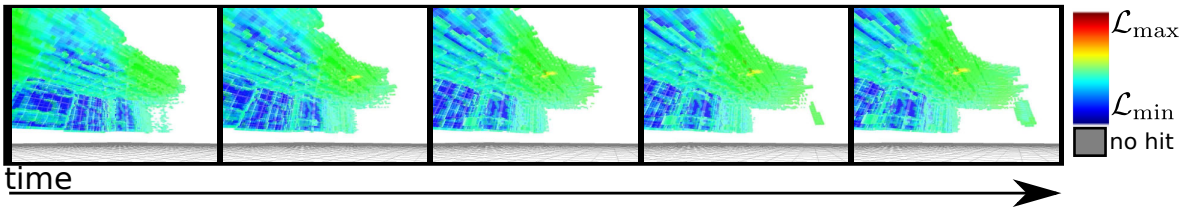


Fig. 9. Sound map creation.

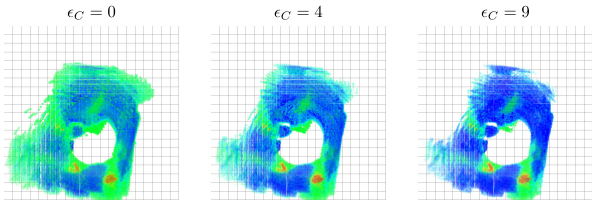


Fig. 10. Top view of the probabilistic 3D sound map for different count threshold ϵ_C.

few update are filtered out.

VII. CONCLUSIONS

This paper introduced a framework for creating a 3D description of the environment that contains the probability that a structure to emit sound. It is a multi-modal approach that combines 3D SRP with 3D LIDAR scans. Experimental results in an indoors environment showed that using the proposed approach it is possible to detect air conditioning grills and associate them with geometric features in the environment. The future work is to experiment in more diverse environments in order to determine the best parameter settings for different situations.

REFERENCES

[1] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition on mobile robots," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, 2009, pp. 2033–2038.

[2] P. Besl and H. McKay, "A method for registration of 3-d shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[3] D. Borrmann, J. Elseberg, K. Lingemann, A. Nüchter, and J. Hertzberg, "The Efficient Extension of Globally Consistent Scan Matching to 6 DoF," in *Proceedings of the 4th International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT '08)*, Atlanta, USA, June 2008, pp. 29–36.

[4] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *IEEE Conference on Acoustics, Speech, and Signal Processing, ICASSP 1997*, 1997, pp. 375–378.

[5] H. DiBiase, J. nad Silverman and M. Brandstein, *Microphone arrays : Signal Processing Techniques and Applications*. Springer-Verlag, 2007.

[6] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013, software available at http://octomap.github.com. [Online]. Available: http://octomap.github.com

[7] N. Kallakuri, J. Even, Y. Morales, C. Ishi, and N. Hagita, "Probabilistic approach for building auditory maps with a mobile microphone array," in *Proceedings of 2013 IEEE International Conference on Robotics and Automation, ICRA 2013*, 2013, pp. –.

[8] E. Martinson and A. C. Schultz, "Auditory evidence grids." in *IROS*. IEEE, 2006, pp. 1139–1144.

[9] ——, "Robotic discovery of the auditory scene," in *ICRA*, 2007, pp. 435–440.

[10] K. Nakadai, H. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition hark and its evalation," in *IEEE-RAS International Conference on Humanoid Robots*, 2008, pp. 561–566.

[11] Y. Sasaki, S. Thompson, M. Kaneyoshi, and S. Kagami, "Map-generation and identification of multiple sound sources from robot in motion," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2010*, 2010, pp. 437–443.

[12] slam6d, "Slam6d - simultaneous localization and mapping with 6 dof," Retrieved December May, 20 2011 from http://www.openslam.org/slam6d.html, 2011.

[13] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[14] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3, sept.-2 oct. 2004, pp. 2123 – 2128 vol.3.

[15] D. B. Ward, E. A. Lehmann, and R. C. Williamsin, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 826–836, Nov. 2003.

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-B302-09

ロボット聴覚ソフトウェア HARK を用いたクイズの同時回答を識別するロボット司会者の設計と実装

Design and implementation of emcee robot of the quiz distinguishing simultaneous answer using Robot Audition Software HARK

西牟田 勇哉 † **Izaya Nishimuta**
平山 直樹 † **Naoki Hirayama**
大塚 琢馬 † **Takuma Otsuka**
杉山 治 † **Osamu Sugiyama**
糸山 克寿 † **Katsutoshi Itoyama**
奥乃 博 † **Hiroshi G. Okuno**

† **京都大学 大学院情報学研究科** Graduate School of Informatics, Kyoto University

{nisimuta, hirayama, ohtsuka, sugiyama, itoyama, okuno}@kuis.kyoto-u.ac.jp

Abstract

実環境で複数の人とコミュニケーションを行うロボットの開発では，話者を一人に絞るカクテルパーティ効果ではなく，雑環境音下で複数人が同時に話しかける状況でも対話が可能な機能が必要である．本稿ではその第一歩として,「早言い」クイズのロボット司会者の設計と実装について報告する．本ロボット司会者は，ロボット聴覚ソフトウェア HARK を用いて発話者の位置を同定し，その発話を分離，音声認識することでクイズの同時回答を識別する．音声認識を頑健にするために，言語モデルの切り替えにより誤認識を抑制し，音韻タイプライタを用いた雑音棄却によって環境音の影響を抑制し，TV 番組「アタック 25」と類似したクイズを対象とした対話システムを開発した．また,「早言い」者の同定精度について評価を行い，60msec の発話のタイミングのずれでは，正しく発話者を同定できることを確認した．

1 はじめに

近年の音声認識技術の発展は著しく，その応用として音声対話システムが数多く登場している[Young *et al.*, 2013]．音声対話システムは Apple 社の Siri，NTT ドコモ社のしゃべってコンシェルといった携帯デバイスにおけるソフトウェアに始まり，PaPeRo [藤田善弘, 2003]や PALRO [富士ソフト株式会社, 2010] といったコミュニケーションロボットにも応用され，人とロボットのインタラクションに貢献している．既存のコミュニケーションロボットは，主に 1 対 1 で直接的にインタラクションを行っていた．一方で，実環境におけるインタラクションでは，多人数を相手にすることが想定される．そのため，参加人数を拡大し，多人数でインタラクションを行うことが可能なロボットが期待されている．ここで，従来のロボットが多人数インタラクションを行う研究では，同時発話の聞き分けや対話参加者の識別を行っていなかった．

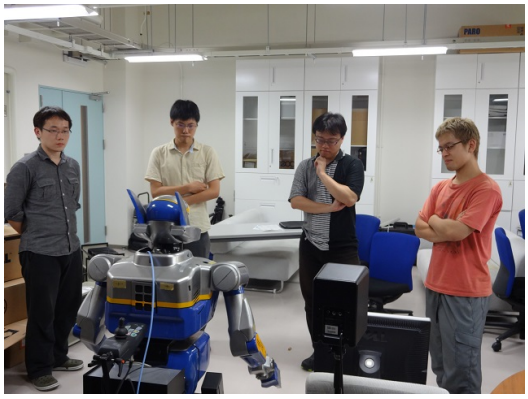


Figure 1: 実装した多人数インタラクション “HATTACK25” の様子．4 人の人がロボット司会者の質問を聞いている．

本研究では，ロボット聴覚ソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto Univercity) [Nakadai *et al.*, 2010]を用いることで前述の同時発話の聞き分け，対話参加者の識別を行うクイズ司会者を設計し実装する．そのために多人数が対戦形式で行うクイズゲーム「パネルクイズ アタック 25」(朝日放送) をケーススタディに採用した音声ベースのクイズゲーム “HATTACK25” (HARK を用いた ATTACK25) を設定した．また，実環境ではロボットの自己雑音，環境音が存在するため，音声認識精度が劣化する．そこで，言語モデルの切り替えによる誤認識の抑制や音韻タイプライタを用いた雑音の棄却によって，実環境に頑健な音声認識を行うことができるようロボットを実装した (Figure 1).

本稿の構成は次の通りである．2 章で関連研究を紹介する，3 章で本ロボットのタスクと課題を定義する．4 章でシステムを設計し，5 章で実装した HATTACK25 の実行例を示す．6 章で話者同定の性能評価の結果を示し，7 章でまとめとする．

2 関連研究

実環境において多人数で行う人・ロボットインタラクションの研究として，Matsusaka ら [Matsusaka *et al.*, 2003]，藤江ら [藤江真也 *et al.*, 2012]の研究が挙げられる．Matusaka らの研究では二人の人とロボットの一問一答形式の質疑応答を実環境で行うことを試みているが，発話の音声認識はロボット聴覚ではなく各対話参加者にマイクロフォンをもたせることで行なっている．藤江らの研究では人同士のクイズコミュニケーションにロボットを介在させることで，そのコミュニケーションを活性化させることを試みている．しかしこの研究では，多人数インタラクションの重要な要素である対話参加者の識別は行なっていない．また，どちらの研究も複数話者が同時に発話することは考慮されていない．

同時発話を処理するロボットの例として，1 章で述べたロボット聴覚ソフトウェア HARK を用いた口じゃんけんの審判を務めるロボット [Nakadai *et al.*, 2008]がある．この研究では同時発話の聞き分けの枠組みを述べているが，インタラクションへの応用は行なっていない．

対話システム構築の手法は，[河原達也 and 荒木雅弘, 2006]，[MacTear, 2004]で紹介されている．しかしこれらは音声認識に重点をおいており，音源の定位・分離結果を用いたシステムについては考えられていない．本研究では，同時発話処理によって得た情報を音声認識結果に加えてインタラクションに利用することで，従来研究では実現出来なかったインタラクションを取り扱う．

3 同時回答を識別するロボット司会者

多人数を相手にインタラクションを行う場合，ロボットはそれぞれの発話が自分に必要な音，聞きたい音だけを取捨選択する必要がある．例えば，ロボットが対話状況を管理する必要があるような役割を持つ場合，複数人が同時に発話したとき，だれに発話権を与えるかの決定をしなければならない．また，発話権を持っている人以外の発話を受理してしまうことがないようにしなければならない．本研究では，発話権の管理が重要となる例として多人数クイズの司会者を取り上げる．

ロボット司会者を実装するための課題は次の通りである．

- 対戦形式であるため対話者の識別が必要
- 音声の早言い (音声ベースであるため，早押しではなく早言いとなる) の合図を適切に処理するために発話混合音の分離が必要
- 回答権を持たない人の発話の棄却が必要

この章では本研究で実装したクイズゲーム”HATTACK25”について述べる．なお，以下ではクイズゲームに参加する人間を「プレイヤー」，司会者を務めるロボットを「ロボット」と表記する．

3.1 概要

本研究ではクイズゲームのケーススタディとして，日本の代表的なクイズ番組である「パネルクイズ アタック 25」(朝日放送) を採用した．アタック 25 は日本で最長寿のクイズ番組である．このクイズ番組の司会進行を参考にすることで，ロボットがクイズ司会者を務めるために必要な課題，要素技術について分析した．

本研究ではアタック 25 をモデルとして音声ベースで再現した，HATTACK25 を実装した．HATTACK25 は基本的にアタック 25 と同じであるが，次のように音声ベースへの変更を施した．

- 問題は読み上げによる一問一答のクイズのみを取り扱う．映像，音楽を用いたクイズは用いない．
- 問題の読み上げはロボット司会者が行う．
- 回答の合図は発話によって行い，早押しボタンは用いない．
- ロボット司会者が問題を読み上げている最中でも回答の合図を行なってもよい．(バージイン発話を許容する)

ゲームは 4 人でプレイする．ディスプレイ上に 1 から 25 の数字が格子状に並んだパネルがあり，プレイヤーはクイズによってこのパネルを取り合う．最終的にパネルを最も獲得したプレイヤーが勝利となる．ゲームは Figure 2 のフローチャートに従って行われ，基本的に出題，回答，パネル選択が繰り返し行われる．またゲーム開始に先立って，ロボットがプレイヤーを識別するために必要な位置情報を取得するための初期化を行う．

3.2 ロボットのタスクと課題

上記の HATTACK25 の司会者をロボットで構築するにあたり，ロボットがなすべきタスクと，そこで発生する課題について明らかにする．HATTACK25 におけるロボットのタスクは，(1) 複数のプレイヤーの回答合図を処理し適切な回答者を決定する，(2) 発話とプレイヤーを対応付ける，(3) クイズの正解・不正解を判定，選択されたパネルを受け付ける，の 3 つである．

それぞれのタスクを達成するためには，(1)，(2) については同時発話の聞き分けやどのプレイヤーが発話した

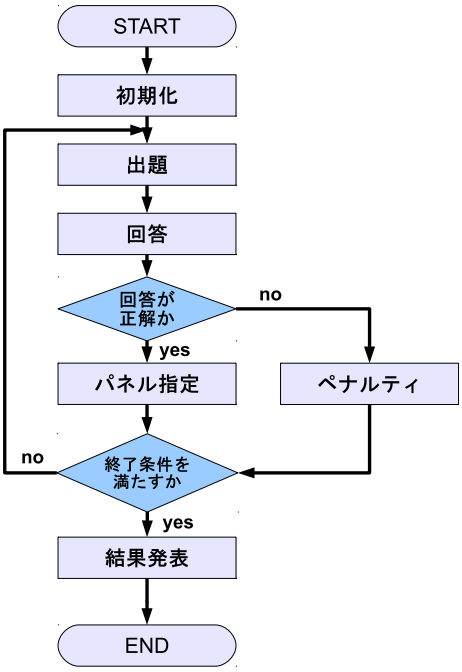


Figure 2: HATTACK25 フローチャート

のかの識別が，(3) については実環境における高い音声認識精度が課題となる．

4 章でシステムの構成と前述の各課題の解決方法について述べる．

4 システムの設計

3 章のロボットのタスク，解決すべき課題の分析結果に基づいて，HATTACK25 の司会を務めるロボットを設計し実装した．はじめに，ロボットの構成をハードウェア，ソフトウェアの両面から詳細に述べる．

4.1 ハードウェア構成

本研究ではロボットを HRP-2 [Kaneko *et al.*, 2004]を用いて実装した．HRP-2 は人の上半身を模したヒューマノイドロボットであり，頭部には 8ch のマイクロフォンアレイを搭載している．外部には合成音声を出力するためのスピーカが接続されており，パネルを表示するためのディスプレイが設置されている．

4.2 ソフトウェア構成

Figure 3 に本研究で設計したシステムの構成を示す．プレイヤーはマイクロフォンアレイを通してロボットへ音声を入力する．そして入力された音響を HARK を用いて定位・分離する．HARK で分離された音声の認識には大語彙連続音声認識システム Julius[1] を用いている．HARK によって得られた定位結果，Julius によって得られた認識結果は状況に応じてゲームの管理に用いられ，必要に応じてパネルディスプレイを変化，合成音声の出力を行う．

[1]http://julius.sourceforge.jp/

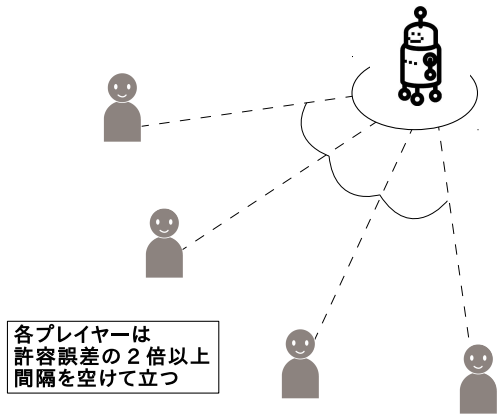


Figure 4: プレイヤーとロボットの位置関係

Figure 3 における Game Management Module とはゲームの管理モジュールの集合であり，この部分を変化させることで様々なインタラクションに応用が可能である．

4.3 課題と解決手法

ロボットを実装する上での主な課題は，3 章で述べたプレイヤーの識別と実環境の中での高い音声認識精度の 2 点である．本研究では HARK による音源定位・分離を用いてプレイヤーの識別を行い，音声認識精度の向上のために雑音・環境音の棄却，誤認識を抑制するための手法を実装した．以下にそれぞれの詳細について述べる．

4.3.1 プレイヤーの識別

HATTACK25 では，回答の合図を発話によって行うため，ロボットは同時に行われる合図の混合音を聞き分ける必要がある．また，どの発話がどのプレイヤーによるものなのかを識別する必要がある．本研究ではこのプレイヤーの識別を，HARK を用いた話者位置同定によって実現した．その手法を以下に示す．

初期化

まず，ゲームを開始する前に位置同定を行うために必要な初期化を行う．プレイヤーはロボットの前方に Figure 4 のように間隔を空けて立つ．続いてロボットの位置確認に対して返事をし，その返事の定位結果をプレイヤーの位置情報として登録する．

HARK を用いた話者位置同定

話者の位置同定は次のように行う．

1. 先に説明した初期化による各プレイヤーの登録位置を θ_i $(1 \leq i \leq 4)$ とする．
2. 発話の定位結果 ϕ が θ_i と式 1 の関係を満たすとき，プレイヤー i が発話したものとみなす．なお式 1 における ε は許容する定位結果の誤差を示す．HATTACK25 では，HARK の定位分解能が 5° 間隔であることと，各プレイヤーの許容誤差範囲が被らない限界を考慮して，$\varepsilon = 15°$ と設定した．

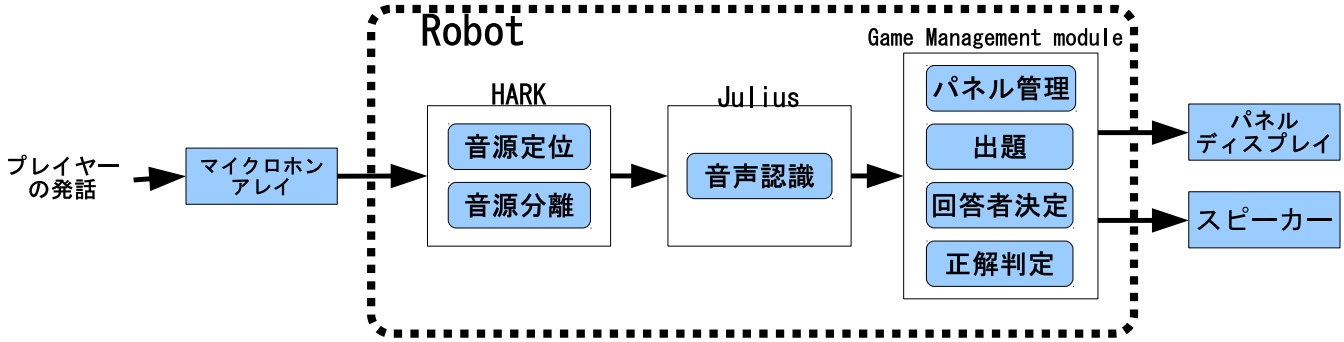


Figure 3: システム構成

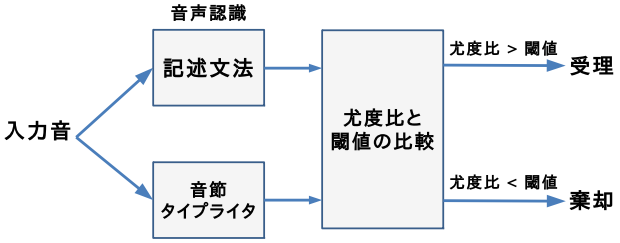


Figure 5: 雑音棄却方法の概念図

$$|\phi - \theta_i| \leq \varepsilon \tag{1}$$

4.3.2 雑音・環境音棄却

実環境でロボットが自身に搭載されたマイクロフォンで音声を認識する場合，周りの環境音やロボットのモータ音などの自己雑音，笑い声・独り言などを何らかの単語として認識し，誤作動を引き起こさないようそれらを棄却する必要がある．本システムでは音韻タイプライタ[伊藤克亘, 1992]を利用することで，そのような雑音や環境音，本来の目的ではない発話の認識結果を棄却する．音韻タイプライタとは，音韻の構造のみを反映した文法であり，あらゆる入力の音響に対してその認識結果の候補仮説の尤度の上限を求める．その音韻タイプライタと目的の文法を並行させて認識を行い，その際の音韻タイプライタに対する目的の文法の尤度比が一定の閾値より小さいとき発話を雑音とみなし棄却する．Figure 5 に音韻タイプライタを用いた雑音棄却方法の概要を示す．

4.3.3 誤認識の抑制

ロボットとのインタラクションにおいて，誤認識は誤動作を引き起こす原因となる．よって本システムでは誤認識を抑制するために，音声認識の際に言語モデルの切り替えを行った．今回の音声認識は，自分で記述した文法モデルを言語モデルとして使用している．ゲームの進行状況によって求められる発話は異なる．そのため，必要な情報のみを記した記述文法を複数用意し，状況に応じて切り替えながら用いることで想定外の発話が認識されないようにしている．例えば，HATTACK25 は基本的に，回答者の決定，問題への回答，パネルの選択が繰り返し行われるが，回答者を受け付けたり，パネルを選択する際に問題の回答がされることはなく，問題回答時に合図がなされたり，パネルが選択されることもない．そのため HATTACK25 では回答者決定における合図のみを受け付ける，問題ごとの回答候補を認識する，パネルの番号を受け付けるといった 3 つのモデルを用意し，切り替えながら音声認識を行っている．

5 実行例

本クイズゲームを実際にプレイした際に，プレイヤーとロボットの間で行われたインタラクション例を示す．これはロボットの出題からプレイヤーが回答し，パネルを選択するまでの一連の流れを示す．以下では Robot，Player がロボット，プレイヤーの発話，*はシステム内部の処理を示す．

インタラクション例

Robot: 次の問題，4 人です.
Robot: ブラジルの首都はどこでしょう.
* 言語モデル: 「はい」モデルへ変更
Player: はい.
Robot: 赤.
* 言語モデル: 「問題」モデルへ変更
Player: ブラジリア
Robot: 正解，ブラジリアだ.
Robot: さぁ，赤の方，何番.
* 言語モデル: 「番号」モデルへ変更
Player: 15 番.
Robot: 15, 14, 13 と赤に変わった.
* パネルディスプレイ: 15, 14, 13 番を赤に更新

6 性能確認

本研究で提案するロボット対話システムの動作確認を行った．動作確認では，4 人の被験者がいることを想定した実験環境を作り，その環境でシステムが設計通りに動くことを確認する．様々な確認項目のなかで今回は，同時発話が

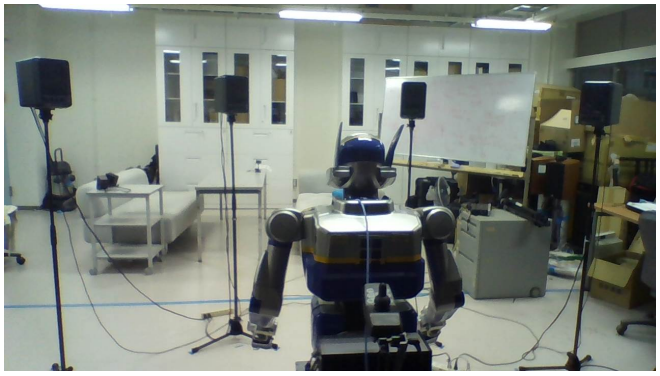


Figure 6: 実験環境

Table 1: 発話内容

発話スピーカ数	4 台中 2 台 (6 通り)
ディレイ	20-200 ms (10 通り)
ディレイを加えるスピーカ	いずれか (2 通り)
繰り返し回数	5 回
総発話回数	600 発話

行われたときの最速発話者の検出と，その位置同定精度の検証を行った．

6.1 環境設定

本実験では人の代わりにスピーカーを使用し，以下の設定に従い実験環境を Figure 6 のように構築した．プレイヤーの間隔は，人の両眼視野が 120° であることから，その視野内にスピーカが配置されるように 40 度間隔で設置した．今回の多人数インタラクションにおける司会者とプレイヤーの関係は Hall の対人距離の定義 [Hall, 1966, pp. 113–125] において，社会的距離に相当すると考えられる．そのためスピーカーはロボット頭部のマイクロフォンアレイの中心から 1.5m の位置に設置した．スピーカの高さは，人間の口の高さに近づけるために地上から 1.5m とした．また実験に先立ち，合図である「はい」という音声を研究室の学生 (いずれも 20 代男性)4 名に発話してもらい，スピーカから再生する音声を録音した．Figure 6 のロボット後方には多数の計算サーバ，ファイルサーバが稼働し，定常的にノイズが発生している．その実測値はロボットのマイクロフォンアレイ周辺において，A 特性音圧レベルの測定平均で求めたところ 61.2 [dB] であった．

6.2 実験内容

Table 1 の発話内容に従ってスピーカーを選択し，ディレイを与えて発話させる．複数の発話情報について，発話音源の最初のフレームの時刻 (Figure 7 では丸で囲った部分の時刻に相当) を比較し，最も早かった発話情報から最速発話者を決定する．そして，その発話の定位結果から同定されたプレイヤーと正解のプレイヤーを比較する．それ

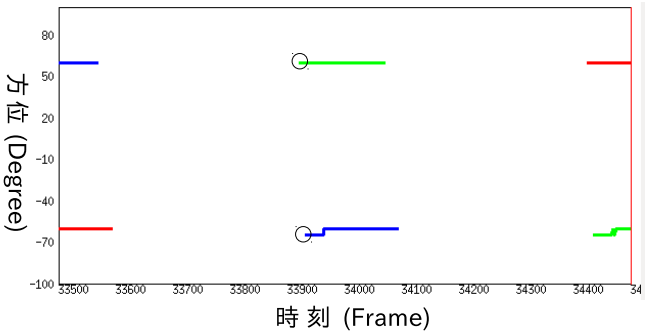


Figure 7: 定位結果

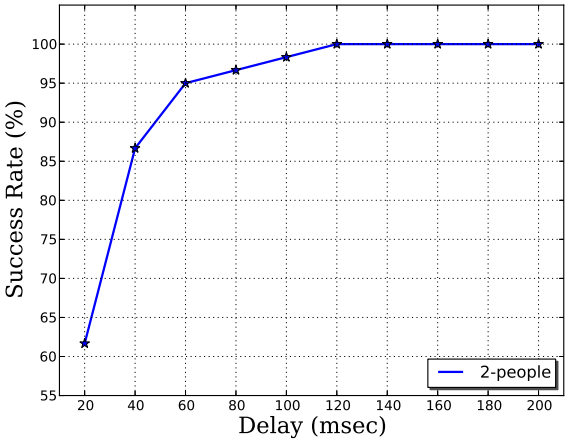


Figure 8: 話者位置同定成功率

によって得られた話者位置同定の成功回数 $N_{success}$ と総発話回数 N_{all} から，式 (2) によって話者位置の同定精度が求められる．

$$N_{SR} = \frac{N_{success}}{N_{all}} \tag{2}$$

6.3 実験結果・考察

Figure 8 に，ディレイと同定成功率の関係を示す．同定成功率はディレイが大きくなるほど 100 %に近づき，60 msec で 90 % 以上，140 msec で 100 %の値を得た．この同定成功率は，2 話者の同定精度という点では十分であると考える．ただし，今回は 4 話者がそれぞれ 1 音声ずつ録音した 4 音声から 2 音声を選び出力する限られた条件での実験であり，結果が話者に依存している可能性もありうる．また，発話人数が 2 人であり，スピーカの間隔や範囲誤差を十分にとった理想的な条件で行なっていた．そのため，発話人数や間隔，誤差おいてより難しい環境を設定した場合，その同定成功率は低下する．よって今後は 3 話者以上が同時に発話した場合や，スピーカの間隔，範囲誤差を変更した場合の実験を行い，今回の実験結果と比較することで現状のシステムの問題発見と解決のために役立てたいと考えている．

7 まとめ

本稿では，同時回答を識別して多人数で対戦を行うクイズゲーム “HATTACK25” の司会を行うロボットを設計し実装した．同時発話の聞き分けやプレイヤーの識別はロボット聴覚ソフトウェア HARK の音源定位，分離結果を用いることで実現し，実環境における音声認識の精度向上のために，言語モデルの切り替えによる誤認識の抑制と音韻タイプライタを用いた雑音棄却を行った．

今後の課題として，性能評価の充実や音声認識の精度向上のために実装した技術の有効性を示すための実験を行うこと，同時発話の聞き分けについての情報をインタラクション部分に組み込むことが挙げられる．今回提案した聞き分けを用いることで，例えば，複数のプレイヤーが同時に反応した，あるプレイヤーが別のプレイヤーにわずかに遅れて反応したといったインタラクションも可能になるのではと考える．

謝辞

本研究は科研費 基盤研究 (S) No.24220006 の補助を受けた．

参考文献

[Hall, 1966] Edward Twitchell Hall. *The hidden dimension*. Doubleday, 1966.

[Kaneko *et al.*, 2004] Kenji. Kaneko, Fumio. Kanehiro, Shuuji. Kajita, Hiroshita. Hirukawa, Toshikazu. , Kawasaki, Masaru. Hirata, Kazuhiro. Akachi, and Takakatsu. Isozumi. Humanoid robot hrp-2. In *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, volume 2, pages 1083–1090, 2004.

[MacTear, 2004] Michael F MacTear. *Spoken Dialogue Technology - Toward the Conversational User Interface*. Springer, 2004.

[Matsusaka *et al.*, 2003] Yosuke Matsusaka, TOJO Tsuyoshi, and Tetsunori Kobayashi. Conversation robot participating in group conversation. *IEICE transactions on information and systems*, 86(1):26–36, 2003.

[Nakadai *et al.*, 2008] Kazuhiro Nakadai, Shunichi Yamamoto, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. A robot referee for rock-paper-scissors sound games. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 3469–3474, 2008.

[Nakadai *et al.*, 2010] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. 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.

[Young *et al.*, 2013] Steve Young, Milica Gašić, Blaise Thomthon, and Jason D Williams. Pomdpbased statistical spoken dialog systems: A review. In *Proceedings of the IEEE*, pages 1160–1179, 2013.

[伊藤克亘, 1992] 田中穂積 伊藤克亘, 速水悟. 音声対話システムにおける未知語の扱い. **人工知能学会研究会資料**, SIGSLUD-9201:1–9, 1992.

[河原達也 and 荒木雅弘, 2006] 河原達也 and 荒木雅弘. **知の科学 音声対話システム**. オーム社, 2006.

[藤江真也 *et al.*, 2012] 藤江真也, 松山洋一, 谷山輝, and 小林哲則. 人同士のコミュニケーションに参加し活性化する会話ロボット (対話生成,< 特集 > 人とエージェントのインタラクション論文). **電子情報通信学会論文誌**. *A,* **基礎・境界**, (1):37–45, 2012.

[藤田善弘, 2003] 藤田善弘. 人工知能の現在と将来 パーソナルロボット PaPeRo の開発. **計測と制御**, 42(6), 2003.

[富士ソフト株式会社, 2010] 富士ソフト株式会社. 小型ヒューマノイド・ロボット PALRO, 2010. `http://www.fsi.co.jp/company/news/100201.html`.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-10

ホースの伸び縮みによるマイク位置の変化を許容するマイクロホンアレイを用いたホース型ロボットの姿勢推定

Posture Estimation of Hose-shaped Robot Using Microphone Array Localization

坂東宜昭 [1] Yoshiaki Bando
大塚琢馬 [1] Takuma Otsuka
糸山克寿 [1] Katutoshi Itoyama
中村圭佑 [2] Keisuke Nakamura
昆陽雅司 [3] Masashi Konyo
田所諭 [3] Satoshi Tadokoro
中臺一博 [2,4] Kazuhiro Nakadai
奥乃博 [1] Hiroshi G. Okuno

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

Abstract

レスキューロボットの一つであるホース型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という利点があるものの，柔軟な本体の制御，姿勢推定が難しいという課題がある．本論文ではホース型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，姿勢推定を行う．ここで，隣り合うマイクロホンと小型スピーカ間の距離は一定であると仮定すると，ホースの湾曲や伸縮により精度が低下することがある．本論文では，マイクロホンの位置と小型スピーカへの距離を同時推定する問題に取り組み，Unscented Kalman Filter を用いたオンライン推定法を開発した．モックアップロボットを用いた実録音データで姿勢推定を評価し，マイクロホンと小型スピーカ間距離を一定とした場合と比較して姿勢推定が 84%抑制されることを確認した．

1 はじめに

災害現場でのレスキューロボットによる，人では捜索が危険な場所や困難な場所の探索が期待されている[Akin et al., 2013]．例えば，汚染物質や倒壊の危険が存在する建築物内の捜索にはクローラ型のロボット[Nagatani et al., 2011][Birk and Pathak, 2006]や，地上からの捜索が困難であれば無人飛行機型のロボット[Onosato et al., 2006]が適用されるなど，状況に応じて様々な形態のレスキューロボットが開発されている．レスキューロボットの設計開発の指針として Robin Murphy は，アメリカでの 5 つの災害における 9 つのレスキューロボットの適応事例から

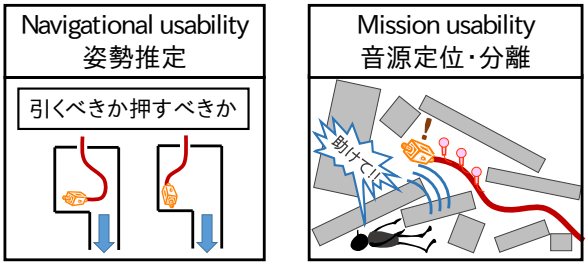


Figure 1: ホース型ロボットにマイクロホンアレイと小型スピーカを装着し，姿勢推定と音源定位・分離機能を同時に実現する．

navigational usability と mission usability が不可欠であると指摘している [Murphy, 2010]．前者は，ロボットの航行能力についての有用性を指し，推進機構や，操縦に必要な情報収集能力の欠如が問題点である．後者は各種センサデータとリモートオペレータへの情報提供についての有用性を指し，データの統合方法，センサシステムの欠陥，オペレータへの提示方法が問題点である．

レスキューロボットの一つであるホース型ロボットは，細長く，ロボット体表と環境との接地面での摩擦を利用して推進できるため，リモートオペレータによる狭い隙間の探索が可能である．リモートオペレータは，先端のカメラから情報を収集し，手元のホースの抜き差しと先端のアクチュエータを用いて目的の方向へロボットを推進させる．例えば，タイミングベルトと小型車輪を用いた Active-Hose [Kitagawa et al., 2003]や，繊毛の振動を用いた Active Scope Camera (ASC) [Namari et al., 2012] などが報告されている．

ホース型ロボットの navigational usability と mission usability を向上させるためには以下の 2 つの機能 (Fig. 1) の実現が不可欠である．

1)navigational usability: 姿勢推定 ホース型ロボット

が瓦礫の隙間などに進入すると，リモートオペレータはその姿勢を視認できない．狭い隙間ではホースがたわむことがあり，押すのか引くのかわからず，ロボットの進入に支障をきたすことがある．

2)mission usability: 音源定位・分離 従来のホース型ロボットには被災者発見と位置推定のため，先端にビデオカメラと単一マイクロホンが搭載されている．瓦礫などの隙間は暗く遮蔽物が多いため，ビデオカメラのみで被災者を発見することは困難である．音は暗闇でも伝わり，また遮蔽物を回りこむため，被災者の音声を定位・分離できれば，被災者の位置や健康状態の把握が可能となり mission usability の向上に寄与する．特に音源定位にはマイクロホンの位置が必要であり，ホース型ロボットではマイクロホンの位置関係が変化するため，音源定位のためにもロボットの姿勢推定が不可欠である．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，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。

Navigational usability としてホース型ロボットの姿勢推定法が開発されてきたが，累積誤差の問題があった [Ishikura et al., 2012]．Ishikura らは，ロボットの姿勢を動的柔軟モデルとして表現し，加速度センサとジャイロセンサにより姿勢を推定した．また，Ishikura らは 3.0[m] の ASC を用いて，推定開始後 35[s] 時点で先端位置の誤差が 0.2[m] 程度となる姿勢推定法を実現した．この手法のセンサ情報は加速度と角速度であり，現在の姿勢は過去の姿勢との差分として得る．そのため累積誤差が蓄積し，ロボットの運用時間の増加すれば姿勢推定の誤差が増加する問題がある．

Navigational usability と mission usability としてマイクロホンアレイを用いた姿勢推定法を開発する．ホース型ロボットにマイクロホンアレイと小型スピーカを搭載する．小型スピーカから試験音を発し，マイクロホン間の到達時間差を手がかりに，ロボットの姿勢を推定する．到達時間差には，現在のマイクロホン位置と音源位置に関する情報が含まれるので累積誤差を修正できる．また，小型スピーカは被災者の呼びかけに，マイクロホンアレイは音源定位・分離に使用できる．本論文では，内界センサによる姿勢推定の欠点を補うために，3.0[m] 以上のホース型ロボットで先端位置の誤差が 0.2[m] 以下となる姿勢推定法の開発を目指す．

本論文の構成は以下のとおりである．第 2 章では音による姿勢推定法開発のために，従来法のマイクロホンアレイ位置推定法問題点を述べ，本論文の立場を明らかにする．第 3 章では音を用いたホース型ロボットの姿勢推定法について述べる．第 4 章では，モックアップロボットを用いた実録音データによる実験から本手法が従来法より誤差が強く抑制されることを確認する．第 5 章でまとめる．

2 マイクロホンアレイ位置推定の関連研究

音の到達時間差を用いた姿勢推定法の開発のため，音を用いたマイクロホン位置推定に関する従来法を概観する．

Ono ら[Ono et al., 2009]は，音源からの直接音到達時間の差から，音源位置と各マイクロホンの位置，録音開始時刻を推定する Blind Alignment 問題を定義し，補助関数法による解法を示した．この手法はオフライン処理を想定しており，ロボットの姿勢推定のような逐次的にマイクロホン位置を推定する問題には不向きである．

Miura ら[Miura et al., 2011]は，Simultaneous Localization and Mapping (SLAM) に基づくオンラインマイクロホン位置推定法として，ロボット周囲を旋回しながら拍手する人のように，既知の移動モデルに従う音源を 1 つ仮定し，音源とマイクロホンの位置と録音開始時刻を同時に推定した．この手法では状態空間モデルを用いて観測に含まれる誤差を考慮している．しかし，瓦礫内で移動する音源を使うことはできず，静止音源 1 つだけでは推定できない．

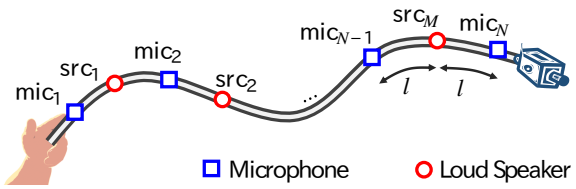
我々はこれまで，複数の小型スピーカをホース上に装着し，ホースの形状制約を用いた状態空間モデルを用いて，スピーカから再生する試験音の到達時間差を手がかりとしたオンライン姿勢推定法を開発してきた[Bando et al., 2013]．しかし，この手法は隣り合うマイクロホンと小型スピーカ間の距離は一定であると仮定し，ホースの湾曲や伸縮により精度が低下することがあった．本論文では，マイクロホンと小型スピーカ間距離を状態変数に追加し，姿勢と同時にオンライン推定する．

3 音によるホース型ロボットの姿勢推定

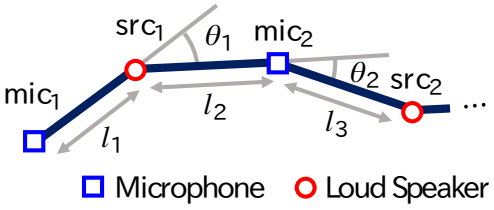
Figure 2(a) に提案法でのマイクロホンと小型スピーカの配置を示す．マイクロホンと小型スピーカは交互にホース上で等間隔 l だけ離して配置する．各マイクロホンと小型スピーカをそれぞれ，手元から順に $\mathrm{mic}_1, \mathrm{src}_1, \cdots, \mathrm{src}_N, \mathrm{mic}_M$ とする．ただし，M, N はそれぞれマイクロホンと小型スピーカの個数を表し，$N = M - 1$ である．本論文ではロボットの姿勢は二次元平面上での曲線で表現されるとし，$\mathrm{mic}_i, \mathrm{src}_j$ の各座標は，$\boldsymbol{x}_{\mathrm{mic}_i}, \boldsymbol{x}_{\mathrm{src}_j} \in \mathbb{R}^2$ とする．Table 1 に本論文で使用する記号の意味を列挙する

以下に本論文が扱う問題設定を述べる．

入力 試験音 $H(\omega)$ を録音した M チャネル同期音響信号 $\{S_{\mathrm{mic}_1}(\omega), \cdots, S_{\mathrm{mic}_M}(\omega)\}$
出力 ロボット姿勢 $\boldsymbol{x}_{\mathrm{mic}_i}, \boldsymbol{x}_{\mathrm{src}_j}$
仮定 (1) 推定中ロボットは静止．
(2) 再生スピーカの番号 j は既知．
(3)$\boldsymbol{x}_{\mathrm{mic}_1}$ と $\boldsymbol{x}_{\mathrm{src}_1}$ は既知．



(a) マイクロホンとスピーカの配置



(b) 姿勢モデル

Figure 2: ホース型ロボットのマイクロホンと小型スピーカの配置，および，姿勢モデル．

Algorithm 1 マイク位置の逐次推定

for $j = 1 \to N$ **do**
　試験音 $H(\omega)$ を小型スピーカ src_j から再生
　M チャネルマイクロホンアレイで録音
　試験音の到達時間差 $\tau^{\mathrm{src}_j}_{\mathrm{mic}_{i_1}\to\mathrm{mic}_{i_2}}$ を計算
　$\tau^{\mathrm{src}_j}_{\mathrm{mic}_{i_1}\to\mathrm{mic}_{i_2}}$ を UKF に入力
　事後確率最大となるように姿勢 $\boldsymbol{\xi}_k$ を更新
end for

入力は到達時間差を得るために使用される．ここで，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源信号である．出力は，マイクロホンと小型スピーカの二次元座標であり，仮定 (1) は問題の簡単化のために設定する．仮定 (2) により，複数の小型スピーカを同じ小型スピーカと混同しない．仮定 (3) により，推定姿勢の回転と平行移動が制限される．

3.1　手法概要：マイクロホン位置のオンライン推定

提案法では，ホース型ロボットの姿勢を音の到達時間差を手がかりとし，Unscented Kalman Filter (UKF) [Julier et al., 1997]を用いて推定する．本オンライン推定法では Algorithm 1 に示すように，1) 各小型スピーカから順に試験音を再生し，2) 試験音の各マイクへの到達時間差を推定し，3) 得られた到達時間差から 3.3 節で述べる姿勢モデルの事後確率が最大となるように姿勢を更新する．以降では試験音の各マイクへの到達時間差推定法，姿勢を表す状態空間モデルについて述べる．

3.2　到達時間差推定

ホース型ロボットの試験音の到達時間差推定の課題は以下の 3 つである．

1. **外部雑音の対処：** 実環境では試験音以外に常に雑音が存在する．
2. **残響・反射の対処：** ホース型ロボットが進入する狭い空間は試験音の残響や反射が発生する．
3. **直接音が取れない場合の対処：** ロボットが湾曲した通路を進入したとき，マイクロホンと小型スピーカを結ぶ直線上に障害物が存在すると直接音の到達時間差が得られない．

Table 1: 記号の定義

記号	意味
M	マイクロホンの数
N	小型スピーカの数 $(N = M - 1)$
C	音速
l	ホース上のマイクロホンと小型スピーカの間隔
ω	周波数
k	観測回数
mic_i	i 番目のマイクロホン $(1 \le i \le M)$
src_j	j 番目の小型スピーカ $(1 \le j \le N)$
$\boldsymbol{x}_{\mathrm{mic}_i}$	mic_i の座標 $(\mathbb{R}^2)$
$\boldsymbol{x}_{\mathrm{src}_j}$	src_j の座標 $(\mathbb{R}^2)$
$\boldsymbol{\xi}_k$	k 回目の観測時の姿勢 $(\mathbb{R}^{2(M+N)-3})$
$\theta_{m,k}$	各頂点の角度 $(\theta_m \in \mathbb{R}, 1 \le m \le N+M-2)$
$l_{n,k}$	各頂点間距離 $(l_n \in \mathbb{R}, 1 \le n \le M+N-1)$
$\boldsymbol{y}^j_k$	j 番目の小型スピーカ再生時の観測 $(\mathbb{R}^{(N^2+N)/2+M-1})$
$H(\omega)$	試験音の音響信号
S_{mic_i}	mic_i で観測した音響信号

特に 1，2) はどのような室内でも起こりうるため，先に解決する必要がある．提案法では，外部雑音と残響・反射への対応のために，それぞれ以下の方法で解決する．

1．外部雑音の対処： 試験音に信号対雑音比が高い Time Streched Pulse (TSP) [Suzuki et al., 1995]を試験音として用いる．自己相関が小さく，エネルギーの大きい試験音を利用すると，到達時間差計測の上でノイズの影響を受けづらい．Miura らのマイク位置推定 [Miura et al., 2011] では，拍手の到達時間差を計測していたが，拍手などのインパルス音は自己相関が小さい．しかし，小型スピーカではインパルス音を大きなエネルギーで再生することが困難である．TSP はインパルスのエネルギーを時間で分散させた信号で，小型スピーカでもエネルギーを確保できる．長さ L の TSP は以下で定義される

$$H(\omega) = \begin{cases} \exp(j2\pi\omega^2/L^2), & 0 \le \omega \le L/2 \\ H(L-\omega), & L/2 \le \omega \le L \end{cases}.$$

2．残響と反射の対処： 残響に頑健な到達時間差推定法である GCC-PHAT [Zhang et al., 2008]を使用し，直接音のピークを抽出する．各チャネルごとに TSP と相関係数を計算し，反射音のピークは直接音より後になるので，閾値以上となるピークの内最初に出現するピークを直接音の

ピークとして抽出する．$\mathrm{mic}_{i_1}, \mathrm{mic}_{i_2}$ 間の試験音 $H(\omega)$ の到達時間差 $\tau^{\mathrm{src}_j}_{\mathrm{mic}_{i_1 \to i_2}}$ は録音信号 $S_{\mathrm{mic}_{i_1}}(\omega), S_{\mathrm{mic}_{i_2}}(\omega)$ から次のように計算する．まず，$S_{\mathrm{mic}_{i_1}}(\omega)$ および，$S_{\mathrm{mic}_{i_2}}(\omega)$ と $H(\omega)$ 間の時間ずれ τ における相関係数を計算する．

$$R_{\mathrm{mic}_{i_1}}(\tau) = \int \frac{G_{\mathrm{mic}_{i_1}}(\omega)}{|G_{\mathrm{mic}_{i_1}}(\omega)|} e^{j2\pi\omega\tau} d\omega$$

$$R_{\mathrm{mic}_{i_2}}(\tau) = \int \frac{G_{\mathrm{mic}_{i_2}}(\omega)}{|G_{\mathrm{mic}_{i_2}}(\omega)|} e^{j2\pi\omega\tau} d\omega$$

ここで，$G_{\mathrm{mic}_{i_1}}(\omega), G_{\mathrm{mic}_{i_2}}(\omega)$ はそれぞれ，$H(\omega)$ と $S_{\mathrm{mic}_{i_1}}, S_{\mathrm{mic}_{i_1}}$ とのクロススペクトルである．次に，閾値を超える相関係数のうち最初のピークとなる τ を選択し，$\mathrm{mic}_{i_1}, \mathrm{mic}_{i_2}$ 間のピークの差を計算して到達時間差を得る．

3.3 姿勢の定式化

Figure 2(b) にホースの姿勢モデルを示す．ロボットの姿勢は，マイクロホンと小型スピーカを頂点とする区分線形曲線により近似する．ホースの姿勢を表す状態変数 $\boldsymbol{\xi}_k$ は，各頂点の角度 $\theta_{m,k}$ $(1 \leq m \leq M+N-2)$ と隣り合うマイクロホンと小型スピーカ間の距離 $l_{n,k}$ $(1 \leq n \leq M+N-1)$ からなる $2(M+N)-3$ 次元ベクトルである

$$\boldsymbol{\xi}_k = \begin{bmatrix} [\theta_{1,k}, \theta_{2,k}, \cdots, \theta_{N+M-2,k}]^T \\ [l_{1,k}, l_{2,k}, \cdots, l_{N+M-1,k}]^T \end{bmatrix}.$$

各マイクロホンと小型スピーカの座標は手元側のマイクロホンとスピーカの座標 $\boldsymbol{x}_{\mathrm{mic}_1}, \boldsymbol{x}_{\mathrm{src}_1}$ を用いて再帰的に計算される

$$\boldsymbol{x}_{\mathrm{mic}_{i,k}} = \boldsymbol{x}^*_{2\times i,k}, \quad \boldsymbol{x}_{\mathrm{src}_{j,k}} = \boldsymbol{x}^*_{2\times j+1,k}$$
$$\boldsymbol{x}^*_{i,k} = \boldsymbol{x}^*_{i-1,k} + l_{i,k} \times [cos(\theta^*_{i,k}), sin(\theta^*_{i,k})]$$
$$\theta^*_{i,k} = \sum_{m=1}^{i-1} \theta_{m,k}, \quad \boldsymbol{x}^*_{2,k} = \boldsymbol{x}_{\mathrm{src}_{1,k}}, \quad \boldsymbol{x}^*_{1,k} = \boldsymbol{x}_{\mathrm{mic}_{1,k}}.$$

状態遷移モデル　状態遷移はランダムウォークで表現する

$$p(\boldsymbol{\xi}_k|\boldsymbol{\xi}_{k-1}) \sim \mathcal{N}(\boldsymbol{\xi}_{k-1}, \mathbf{Q})$$
$$\mathbf{Q} = \begin{bmatrix} \sigma^2_{q\theta}\mathbf{I}_{M+N-2} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{ql}\mathbf{I}_{M+N-1} \end{bmatrix}.$$

ただし，$\sigma_{q\theta}$，σ_{ql} はランダムウォークの標準偏差を表す．

観測モデル　観測変数 $\boldsymbol{y}^j_t \in \mathbb{R}^{(N^2+N)/2+M-1}$ は (1)j 番目の小型スピーカが再生時の到達時間差と (2) 隣接するマイクロホンと小型スピーカ間の距離 l_n $(1 \leq M+N-1)$ とする．本モデルでは l_n が可変となるので，各頂点は 2 次元平面上で自由な位置を取りうる．観測として l_n を与え l_n の存在範囲に制限を加えることで，推定値の発散を抑制する．姿勢推定時に正解の l_n は得られないので，ホース上の配置間隔 l を観測値として推定する．

$$p(\boldsymbol{y}^j_k|\boldsymbol{\xi}_k) = \mathcal{N}(g(\boldsymbol{\xi}_k), \mathbf{R}_k)$$

$$g(\boldsymbol{\xi}_k) = \begin{bmatrix} \left[\tau^{\mathrm{src}_j}_{\mathrm{mic}_{i_1 \to i_2}} = \frac{D^{i_2,j}_k - D^{i_1,j}_k}{C} \middle| \begin{matrix} i_1 = 1, \cdots, N \\ i_2 = i_1+1, \cdots, N \end{matrix} \right]^T \\ [l_n | n = 1, \cdots, M+N-2]^T \end{bmatrix}$$

$$D^{i,j}_k = |\boldsymbol{x}_{\mathrm{mic}_i} - \boldsymbol{x}_{\mathrm{src}_j}|$$

$$\mathbf{R}_k = \begin{bmatrix} \sigma^2_{r_\tau}\mathbf{I}_{(N^2-N)/2)} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(\sigma^2_{r_{l_1},k}, \cdots, \sigma^2_{r_{l_{M+N-1}},k}) \end{bmatrix}$$

ここで，C は音速を表し，$\boldsymbol{v}_k$ は観測誤差を表す確率変数である．$\boldsymbol{v}_k$ のうちマイクロホン・小型スピーカ間距離 l_n の分散 σ_{l_n} は姿勢が大きく曲がりくねるほど大きな分散をもつと考えられるので以下のように与える

$$\sigma_{r_{l_n},k} = a \times \frac{|\theta_{n,k}| + |\theta_{n+1,k}|}{2} + b.$$

4 姿勢推定精度の評価実験

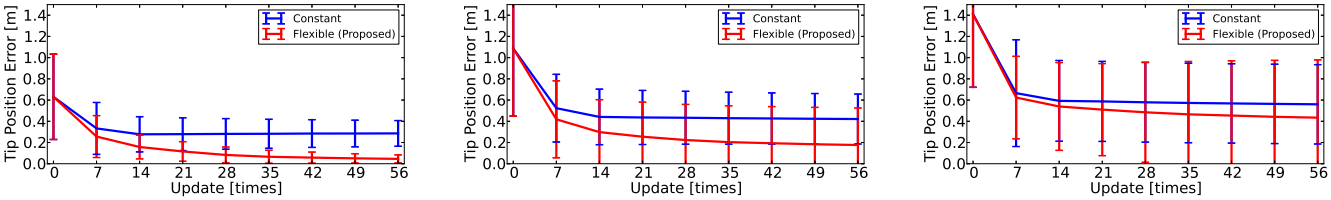
実験ではモックアップロボットによる実録音を用いて提案手法と，従来手法としてマイクロホン・小型スピーカ間距離が一定の手法[Bando et al., 2013]との比較を行う．

4.1 モックアップロボットの仕様

ホースにマイクロホンと小型スピーカだけを装着した推進機能のないモックアップを構築した．本体のホースはポリプロピレン製のコルゲートチューブ (内径 15[mm]) である．表面には $l = 0.25$[m] 間隔で，小型スピーカとマイクロホンを交互に配置した．小型スピーカには，直径 20[mm] の磁気小型スピーカを用いた (Fig. 4(a))．マイクロホンには，デジタル MEMS マイク (12[mm]×12[mm], Fig. 4(b)) を使用した．マイクロホンの数 M は $M = 8$, 両端のマイクロホン間のホースの長さは 3.5[m] である．

4.2 実験設定

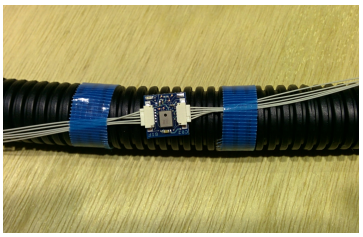
Figure 4(c) のように，実験は高さ約 1[m] の壁を設置した実験室に，モックアップを配置して行った．実験室の残響時間 (RT_{60}) は 800[ms] である．正解の姿勢はモーションキャプチャシステムである NaturalPoint, Inc 製の OptiTrack を用いてマイクロホンと小型スピーカの位置を計測して作成した．試験音には TSP 信号 (1.0[s]) を用い，各小型スピーカから 8 回ずつ再生して録音した．A/D 変換器は，株式会社 システムインフロンティア製の RASP-ZX を用い，録音は HARK を用いて 16kHz，24bit でサンプリングを行った．カルマンフィルタの初期値として，θ_m には正解データに正規分布に従う誤差 (標準偏差: σ_{ini}) から与え，l_n には 0.25[m] を与えた．ランダムウォークのパラメータ $\sigma_{q\theta}$,σ_{ql} はそれぞれ 0.001[rad]，1.0[mm] を，観測誤差のパラメータ σ_τ, a, b は 0.4[ms], 0.283, 0.001[deg] を与えた．これらの値は実験的に定めた．σ_{ini} は 15, 30, 45 [deg] の三種とし，それぞれ 64 回の異なる初期値による推定結果について先端位置の誤差を評価した．



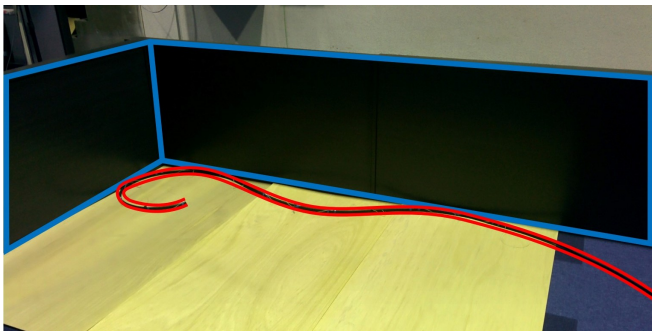
(a) $\sigma_{ini} = 15$ [deg] (b) $\sigma_{ini} = 30$ [deg] (c) $\sigma_{ini} = 45$ [deg]

Figure 3: 姿勢推定結果の先端位置の誤差．初期値に加える誤差の標準偏差を 15[deg] から 45[deg] に変動させた．

(a) 小型スピーカ



(b) MEMS マイクロホン



(c) 実験の様子．モックアップを赤で，木製の壁を青で囲んだ

Figure 4: 実験時の写真

4.3 評価実験結果

Figure 3 に，各更新ごとの先端位置の誤差を示す．誤差の平均を折れ線で，標準偏差をエラーバーで示した．従来手法 (青) に対し提案法 (赤) の方が誤差が，初期値誤差の標準偏差 σini = 15[deg] のとき 84%, 30[deg] のとき 58%, 45[deg] のとき 22%抑制された．特に従来手法では提案法に比べて 14 回目以降の更新から誤差の減少が少ない．これは，従来手法がロボットの湾曲やホース上のマイクロホンと小型スピーカの配置間隔の誤差を許容できなかったためと考えられ，提案法が有効に作用していることが分かる．

4.4 考察

内界センサを用いた手法で，Ishikura らの実験では 3[m] のホース型ロボットに対し先端位置の誤差は 0.2[m] 程度であった [Ishikura et al., 2012]．初期値誤差の標準偏差 $\sigma_{ini} = 15$[deg] において提案法の先端位置誤差の平均は 56 回目の更新で 0.1[m] 以下で，内界センサを用いた手法より高精度な推定を実現した．また，初期値の誤差が抑制されており，累積誤差の問題が解決できた．

一方，初期値誤差の標準偏差 σ_{ini} の増大と共に推定値の誤差が増大している．本論文で示したマイクロホンと小型スピーカの配置では $\mathrm{mic}_1, \mathrm{src}_1$ を軸として鏡対称な姿勢はすべて同じ到達時間差を観測する．このため，対称な姿勢の区別ができず提案法の状態空間は多峰性となるので，適切な初期値を与える必要があると考えられる．この対称性問題は，(1) ホース上のマイクロホンの配置を 2 列以上にして左右の弁別を可能にすることや，(2) 内界センサの情報と統合し内界センサで左右の弁別を，音によって累積誤差の軽減を行うことで，頑健性が向上すると考えられる．

5 まとめ

本論文では，ホース型ロボットの navigational usability と mission usability の向上のために，音によるホース型ロボットの姿勢推定法を述べた．改良前の手法ではホースの湾曲や伸縮により精度が低下する問題があった．また，従来の内界センサを用いた手法には，累積誤差の問題があった．提案法では，隣接するマイクロホンと小型スピーカ間の距離を推定することにより，内界センサによる手法が達成した精度以上である「3.0[m] 以上の長さのロボットで先端位置の誤差が 0.2[m] 以下」となる精度を目指した．評価実験では，$\mathrm{RT}_{60} = 800$[ms] の残響のある環境下での 3.5[m] のモックアップホース型ロボットを用いて提案法を評価し，初期値誤差の標準偏差が $\sigma_{\mathrm{ini}} = 15$[deg] のとき，先端誤差が 4.6[cm] 程度となる精度を実現した．これは，改良前の手法にくらべ 84%誤差を抑圧しており，目標である内界センサによる手法より良い精度を得られた．

今後の課題としては，navigational usability においては，内界センサの導入，マイクロホンの配置の検討，および，気温変化に伴う音速のゆらぎや，直接音が取得できない場合への対処など頑健性の向上が挙げられ，mission usability においては，音源定位・分離技術を利用したリモートオペレータへのマルチモーダル情報提示システムの開発などが挙げられる．

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

[Akin et al., 2013] H. Akin et al. Robocup rescue robot and simulation leagues. *AI Magazine*, 2013.

[Bando et al., 2013] Y. Bando et al. Posture estimation of hose-shaped robot using microphone array localization. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3446–3451, 2013.

[Birk and Pathak, 2006] Andreas Birk, Kausthub Pathak, Soeren Schwertfeger, and Winai Chonnaparamutt. The iub rugbot: an intelligent, rugged mobile robot for search and rescue operations. In *2006 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, volume 10, 2006.

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

[Julier et al., 1997] S. Julier et al. New extension of the kalman filter to nonlinear systems. In *AeroSense'97*, pages 182–193, 1997.

[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 *2011 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. In *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2130–2137, 2011.

[Murphy, 2010] Robin R Murphy. Navigational and mission usability in rescue robots. *Journal of the Robotics Society of Japan*, 28(2):142–146, 2010.

[Nagatani et al., 2011] K. Nagatani et al. Redesign of rescue mobile robot quince. In *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics (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 *2012 IEEE/RSJ IROS*, pages 3679–3686, 2012.

[Ono et al., 2009] N. Ono et al. Blind alignment of asynchronously recorded signals for distributed microphone array. In *WASPAA'09.*, pages 161–164, 2009.

[Onosato et al., 2006] M. Onosato et al. Aerial robots for quick information gathering in usar. In *SICE-ICASE, 2006*, pages 3435–3438, 2006.

[Suzuki et al., 1995] Y. Suzuki et al. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *The Journal of the Acoustical Society of America*, 97:1119, 1995.

[Zhang et al., 2008] C. Zhang et al. Why does phat work well in lownoise, reverberative environments? In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008.*, pages 2565–2568, 2008.