人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B102

ＡＩチャレンジ研究会 (第３４回)

Proceedings of the 34th Meeting of Special Interest Group on AI Challenges

CONTENTS

日　時　2011 年 12 月 15 日　　場　所　慶應義塾大学 日吉キャンパス 来往舎　シンポジウムスペース
Keio University, Kanagawa, Dec. 15, 2011

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

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-B102-2

聴覚実験における音響テレプレゼンスロボットの有用性
Acoustical telepresence robot for auditory psychophysics

○戸嶋 巌樹*
近藤 洋史*
Daniel Pressnitzer[†‡]
柏野 牧夫*[§]
* NTT コミュニケーション科学基礎研究所,
[†]UMR 8158, CNRS and Université Paris Descartes
[‡]Ecole normale supérieure
[§]東京工業大学大学院 総合理工学研究科

Iwaki TOSHIMA*, Hirohito M. KONDO*, Daniel Pressnitzer (Univ Paris)[†‡], and Makio KASHINO*[§]

* NTT Communication Science Laboratories,
[†]UMR 8158, CNRS and Université Paris Descartes
[‡]Ecole normale supérieure
[§]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

{toshima.iwaki, kondo.hirohito, kashino.makio}@lab.ntt.co.jp, Daniel.Pressnitzer@ens.fr

Abstract—"Embodiment" is one of the most important key words of studies of robots' learning, human-robot interaction, and understanding of environment for robots. In the area of robot-audition and human auditory perception, this is also important. We use an acoustical telepresence robot: *TeleHead* that has human-like dummy head and is synchronized with user's head movement in realtime. We can control "embodiment" of the TeleHead such as head shape and head movement. Then, we tried several psychophysical experiments; sound localization, delay discrimination, and streaming segregation. Then, we conclude that TeleHead can be used for psycho-physical experiments.

1. はじめに

急速に進歩するロボット技術を用いて，人間の性質を解明しようとする研究が多くなされている．そもそも人間の発達や知能を議論するにおいて，外界という未知なるものを如何にロボット（即ちシミュレートされた人間）に理解させるかということは大問題である．そこで，身体を通じた外界の理解に代表される，いわゆる「身体性」が提唱されて久しい[1]．これは，1990 年代の Brooks らが強く主張し，ロボット研究で議論されてきた，知能は身体を必要とする，という議論の発展であると捉えられる[2]．このように，古くから知能そのもの，あるいはその獲得プロセスをシミュレーションするためには，物理的存在のある身体，すなわちロボットの身体が必要であると考えられてきた．近年，そのロボットの身体性能が人間に近付いてきたことにより，ロボット技術を用いた人間の性質解明への挑戦が活発になってきたと言える．これは，言語が，その入れ物である脳と共に進化してきた [3]こととのアナロジーを持って受け止める事ができると考える．環境を理解し，知覚し，環境に働きかけるためには，自らの持つ物理特性をはじめとする諸性質を踏まえ，その理解と行動の可能な範囲において，環境を理解し，環

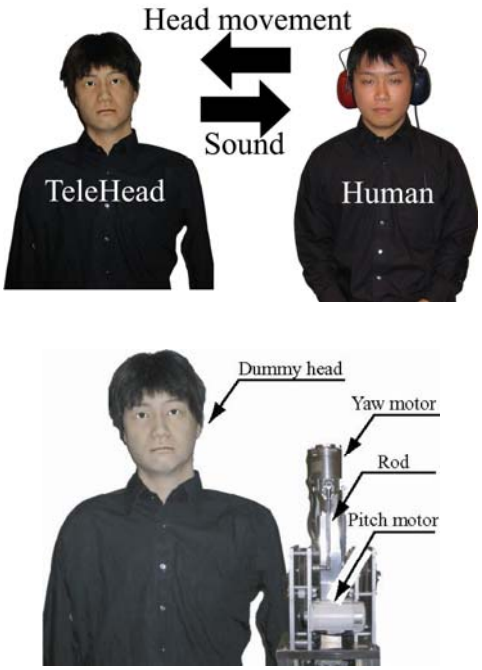


Fig. 1 Outline of TeleHead system and inside of the TeleHead.

境への働きかけを行うのが効率的である．

ところで，本論文で研究対象とする聴覚においても，身体性の役割は重要である．人間は個々に異なる形状の頭部を持つ．従って，同一の環境，同一の音源からの音であったとしても，鼓膜に伝わる音響信号は異なる．また，頭部運動は音響信号の測定点，即ち耳介の位置を変化させるため，音環境理解において，重要な要素となる．この場合，聞き手としての人間は，自らの頭部形状の影響，頭部運動の影響を事前に学習し，音環境を理解していると考えられる[4].

前述のロボットを用いた人間のシミュレーションという意味においては，人間と頭部形状の寸分違わない頭部を持つロボットを製作し，人間の動きに追従動作し，さらにその影響を確かめるために，それらの性能を自由に変化させることが可能なロボットがあれば，それらの影響を自由に調べることが出来る．そこで，現状で可能な限り，上記の性能を満たすような，音響テレプレゼンスロボット，テレヘッドを作成した[5]．テレヘッドを用いることで，頭部形状や頭部運動の効果を任意に操作し，人間の性質解明に役立てたいと考えている． 本稿では，テレヘッドによって知覚実験を行うにあたり，必要と考えられる基礎的性質，音像定位感や遅延の影響を評価し，それらを踏まえた上で，様々な要素が関係するストリーミング課題での活用の是非を検討した．

2. 音響テレプレゼンスロボット「テレヘッド」

我々は使用者と同形状の頭部を持ち，使用者の頭部運動に追従動作するロボット，テレヘッドを作成した[5]．概要を図１に示す．使用者は，テレヘッドの外耳道入り口付近に設置されたマイクロホンにより集音した音を，ヘッドホンを通じて聴く．頭部形状が使用者の頭部と一致していることにより，使用者の頭部が音環境に与える影響を計算することなく，頭部形状の影響を加味することが出来る．また，ヘッドホンに付けられた頭部運動センサが使用者の頭部運動を測定し，この情報に基づいて，ロボットの頭部が制御される．これにより，使用者の頭部運動の効果が遠隔の音環境で実現され，使用者に提示される．以上により，使用者は，遠隔において，自らの頭部形状と頭部運動の影響が加味された音を聴くことが出来る．

この時，実機を使うデメリットとして，騒音と遅延が考えられるが，騒音は 24dB SPL 程度であり，静かな部屋でも，何らかの刺激音を聞いている場合はほとんど知覚されない[6]．また，遅延は 80ms 程度（delay time，動き出しのタイミング dead time としては 10ms 程度）は，後述の検討により，臨場感のある音を聴くという使用目的に対して，知覚上は問題ないことを確認している．

3. 音像知覚への影響

3.1 実験概要

音が聞こえる方向を正しく知覚すること，即ち音像定位感は聴覚の実験をする上で重要である．テレ

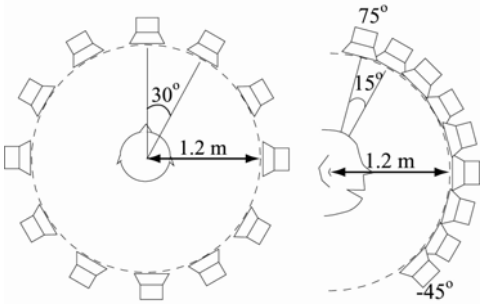


Fig. 2 Setup of sound localization experiments.

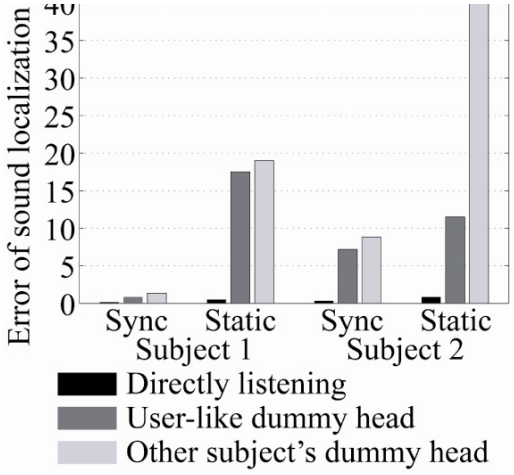


Fig. 3 Sound-localization error in azimuth plane.

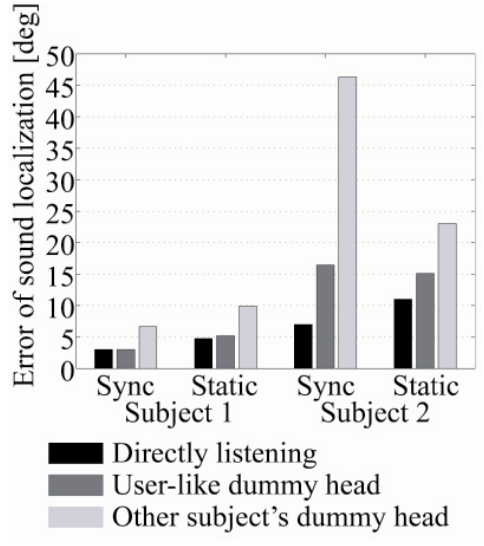


Fig. 4 Sound-localization error in vertical plane.

ヘッドを用いることで，どの程度音像定位感が得られるか実験的に明らかにした[6].

テレヘッドもしくは使用者自身の周りに図 2 に示すようにスピーカを配置した．左は水平面条件，右は正中面条件でのスピーカ配置を示している．テレヘッドを使用する場合，使用者は別室で受聴し，音の聞こえる方向を回答した．テレヘッドを使用しない場合は，スピーカ配置の中心に使用者が座り，直接受聴することとなる．この場合の回答は HMD を用いて提示された画面に対して，手に持ったマウスを用いて回答した．いずれの場合もスピーカは見えない．また，テレヘッドの特長である，頭部形状を一致させたことの効果，頭部運動を遠隔で実現したことの効果をそれぞれ検証する意味で，頭部形状を一致させた場合とさせない場合，また，頭部運動を行った場合と行わない場合についても実験した．

3.2 実験結果

結果を図 3, 4 に示す．図 3 は水平面，図 4 は正中面の定位誤差を表している．いずれも左のバーが直接受聴条件であり，中央が頭部形状一致条件，右が頭部形状不一致条件である．また，各被験者について，左のブロックが頭部運動有り条件，右のブロックが頭部運動無し条件である．頭部形状を制作することの困難さが問題となり，被験者数が 2 名と少ないが，この範囲においては，いずれの条件でも頭部形状が一致し，頭部運動が再現されている場合において，高い音像定位結果が得られている．これは頭部形状の再現，頭部運動の再現共に，定位感の向上に寄与していることを示している．また，テレヘッド使用で，頭部形状一致かつ頭部運動有りの条件で，少なくとも被験者 1 は直接受聴と同等の音像定位精度となっている．これは，テレヘッドシステムが，十分に高精度で機能している事を示しており，音像定位に関して，知覚実験が可能な装置であることを示唆する結果である．被験者 2 については，頭部形状一致，頭部運動有り条件でも，直接受聴と同等までは定位精度が向上しなかったが，元々被験者自身の定位精度が高くない事と，頭部形状の再現精度が被験者 1 のロボットに比べて若干劣ったことが分かっている．個人性の問題もあるため，安易な結論は述べられないが，少なくとも直接受聴並の音像定位精度が得られる結果があることは，テレヘッドの様なシステムが音像定位精度に使用できる可能性を示唆する結果と言える．

4. 遅延知覚

ロボットを用いる際に考慮すべき重大な問題に遅延がある．実機を用いる場合は，計測の時間（多くの場合無駄時間：dead time になる）に加え，機械のイナーシャによる動作遅延が不可避である．テレヘッドでは，系全体としての遅延が，60deg 程度のステップ状の頭部運動に対して 80ms 程度ある[6]. さて，この 80ms が許容範囲かどうかという事について，以下の通りに実験した[7].

4.1 実験方法

テレヘッドを利用し，正面から音を出して，意識的に頭部運動をする．この際，トータルの遅延時間の異なる 2 回の試行を行い，被験者はどちらの試行の方が，遅延時間が長かったかを回答した．被験者は 4 名，少なくとも耳介形状は一致するように頭部を制作した. テレヘッド自体の遅延が 80ms 程度あることから，遅延時間の短い条件は 80,100,120ms の 3 通りとし，これらと，これらよりもさらに長い遅延について，弁別できるかどうかを調査した.

4.2 実験結果

結果を図 5 に示す． Weber 則によれば，比較の基準となる短い遅延刺激の遅延時間が長くなれば，追加遅延時間の閾値が高くなることが予想された．しかし，結果は被験者によって閾値自体はことなるものの, そのような傾向は見られず, むしろ 40ms, 20ms などで，被験者毎に一定の閾値となった．これは，タスクへの回答が遅延時間自信ではなく，何か他の手がかりを利用しており，その結果，一定になったことを示唆しているのではないかと考えた．その手

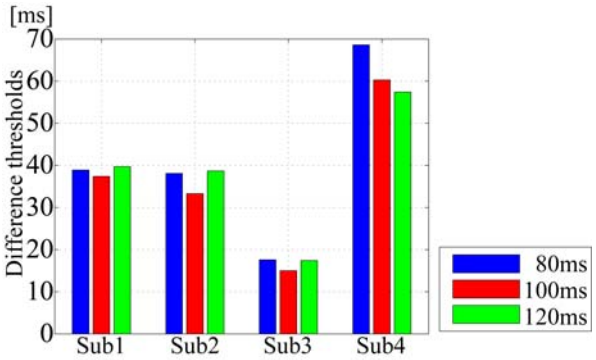


Fig. 5 Delay thresholds

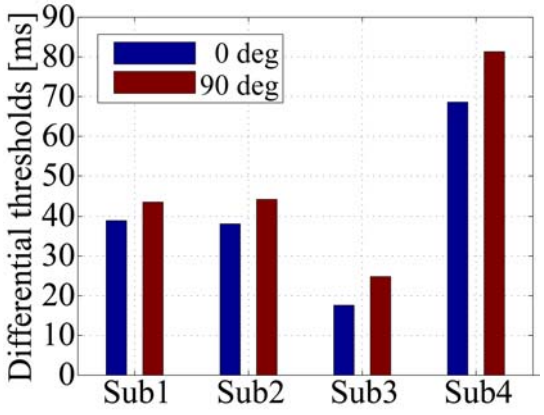


Fig. 6 Delay thresholds depend on sound direction

がかりを空間的な手がかり，つまり音像定位感の差と仮定して，いくつかの追加実験を行った．仮説は，頭部運動の遅れを定位感のずれとして知覚しているということである．仮説が正しければ，頭部形状を変化させ，定位感を低下させた場合，閾値は高くなる．また，音源の方向を定位精度の高い正面に配置した場合と，定位精度の低い横方向に配置した場合では，正面の方が，閾値が低くなる．さらに，広帯域雑音に比べて，音声では定位精度が低下すると考えられるため，閾値が高くなるはずである．

実際に正面と横方向に音源を配置した場合の結果について，図 6 に示す．いずれの被験者においても，正面方向に音源を設置した場合に閾値が下がり，遅延に対して敏感になった．他の 2 つの条件，頭部形状の変化と刺激音の変化についても，定位感によって遅延を弁別していることを支持する結果を得ており，空間知覚が遅延の知覚に重要であることを示唆する結果を得た．なお，その際の閾値が頭部運動の最高速度と遅延時間の積，即ち遅延に影響を受けた音像のずれ角度は約 10 度〜17 度となるあたりに遅延の閾値があることが実験的に分かっている．

ところで，人間の首の動きは瞬間的でも高々 300deg/sec 程度であり，テレヘッドの無駄時間 10ms の間に移動する距離は 3deg 程度である．また，遅延時間 80ms の間，300deg/sec の運動を継続できたと仮定しても，80ms 後に発生する音像の空間的誤差は 12deg（300deg/sec × 80ms × 1/2）であり，これは広帯域雑音を正面から提示したとしても，知覚限界付近である．音声を用いた場合はさらに知覚されにくくなることも確認済みであり，テレヘッドは単体で通信を伴わずに使用する場合には，許容される遅延レベルであることが分かる．

さらに実用的な側面では，頭部形状の一般化が考えられるが，その場合，この運動遅延の影響はさらに小さくなることも分かっている[8]．また，同様のシステムを２台遠隔でつないでテレプレゼンスの質を評価するような検証も行われている．それによれば，TCP/IP で東北大学と富山県立大学を結んだ場合，テレヘッドシステムの頭部運動は数 630ms，UDP で約 140ms 遅れ，遅延時間の揺らぎによって，両耳間時間差が定位方向で 0.5deg に相当する程度，誤差を持つことが分かっている．このようなより実用に即した場面での遅延の影響の評価等は今後の課題として残っていると言える[9].

5. 複数の要因が関係する知覚実験へのテレヘッドの使用:ストリーミング課題

5.1 実験の背景・方法

ここまで頭部形状や頭部運動にまつわる，音像定位や，遅延の感覚について議論してきたが，様々な要因が関係する知覚現象も多くある．本稿では，その一例として，音脈分凝について，議論する． ストリーミング課題と呼ばれる知覚課題である．図 7 に示すような音の時系列刺激が提示された場合，聞き手には多くの解釈が存在する．例えば，ABA や AAB 等でセットとなるリズミックな音列が提示されているという解釈（1 ストリーム）．あるいは，音の繰り返しの速さと高さの異なる AAA と BBB という 2 つの音列が同時に提示されているという解釈（2 ストリーム）である．一般的には，聞こえはじめは 1 ストリームであるが，連続して聞いているとやがて 2 ストリームになると言われている．しかし，この現象は，音の空間配置や頭部運動，あるいは注意（聞き手が 1 ストリームと 2 ストリームのどちらとして知覚したいと考えているか）などが複雑に関係するため，その機序の解明は容易ではない．

ここで，テレヘッドを用いれば，空間的知覚，頭部運動，頭部運動の意識の影響を切り分けることが出来る．頭部運動にあわせてテレヘッドを動かす通常の条件に加え，頭部運動にかかわらずテレヘッドを動かさないことで，頭部運動の音響的効果を打ち消す条件，さらに，頭部運動していないのに，以前の使用者の頭部運動をテレヘッドが再生することにより，頭部運動の音響的効果のみを与え，自己運動

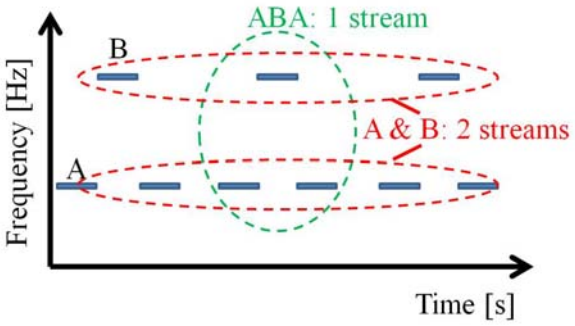


Fig. 7 Streaming segregation.

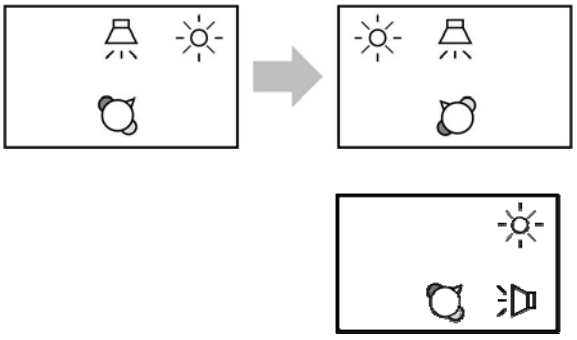


Fig. 8 Head movement of the experiments. Upper panel shows the normal situation. Under panel shows the condition that subject does not move the head and TeleHead moves his head. This makes sound direction change.

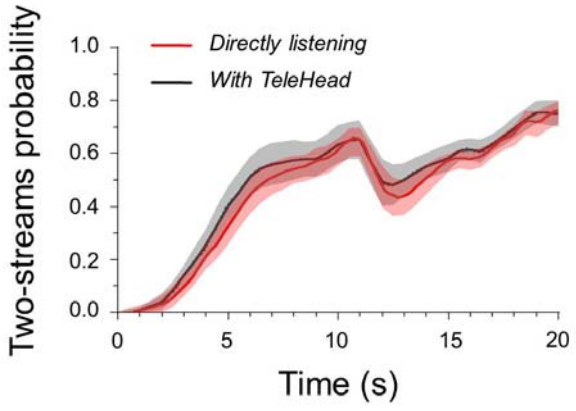


Fig. 9 Results of streaming task.

の効果を与えない条件，などがすぐに実施可能である．具体的な実験の例を図 8 に示す．始め被験者は右の LED を注視し，正面のスピーカからストリーミングの音響刺激を聴いている．あるタイミングで左の LED が点灯し，被験者は素早く頭部運動を行う．この時に，頭部運動の影響を受けて，ストリーミングがどのように変わったかを回答す．一方で，図 8 下に示す条件では，同じタイミングで被験者へは頭部運動の指示を出さず，TeleHead のみが直前の被験者の頭部運動と同じ運動を行う．この場合，被験者もしくはテレヘッドの運動後の，被験者と音源の相対的方向は一致する．従って，被験者への音響信号は一致しているが，自らの意識を伴って頭部運動したかどうかが異なる．この 2 つの実験の結果を比較することで，自発的頭部運動の効果について論じることが出来る．

既存のバーチャルリアリティシステムでも，たたみ込み演算による加減運動で音の運動や頭部運動を再現できる．しかしこれは，パラパラ漫画が動画に見えるのと同じ事であるが，この様な複雑な知覚現象においては，連続運動と知覚されることが，本当に連続的に変化する音響信号の情報と等価かどうか疑問が残る．

もちろん，一方で，ロボットを使用する弊害も多く存在する．遅延や，頭部形状の再現精度の問題も大きな物である．そこで，常に，人間が直接受聴する場合とどの程度異なるのか，議論する必要がある．そこで，使用者にストリーミング課題を課し，時間と共に聞こえがどう変化するか，ロボットを介す場合と介さない場合を比較した．

結果を図 9 に示す．図 8 の上段の実験条件におけるストリーミング課題の回答結果を示している．横軸が時間であり，10 秒後に頭部運動を行うようにした．縦軸が 2 ストリームと知覚する確率である．はじめは 1 ストリームであるが，数秒後には 2 ストリームと聞こえる確率が上昇し，10 秒目に発生する頭部運動イベントによって，2 ストリームと聞こえる確率が少し低下し，運動後もう一度上昇していることが確認できる．この実験をテレヘッドを介さず，被験者が直接音を聴いた場合が重ねて描かれているが，両者にはほぼ差がないことが読み取れる．これは，遅延や騒音など，様々な影響はあるものの，複雑な知覚現象を扱う実験においても，テレヘッドが有効に機能していることを示唆する結果と言える．

6. おわりに

頭部形状と頭部運動を再現する音響テレプレゼンスロボット：テレヘッドを用いて，音像定位やストリーミング課題によって，ロボットを用いる実験の可能性について検討した．

・音像定位課題においては，頭部形状を作り込むことにより，定位精度は向上した．特に頭部運動を再現することで，定位精度は直接受聴とほぼ変わらない程度まで向上することが確認された．

・遅延弁別課題においては，遅延の追加に対する閾値が定位感のずれに依存していることを示唆する結果を得，80ms の遅延ではほとんど，定位感のずれに起因する遅延を知覚することは無いことを示唆する結果を得た．

・ストリーミング課題は，頭部形状，頭部運動，注意，空間知覚など，多彩な聴覚の知覚が関連する複雑な課題であるが，この様な課題においても，テレヘッドを用いた結果は，直接受聴と変わらなかった．

以上から，テレヘッドは音像定位の様な単純な聴覚実験や，ストリーミング課題の様な複雑なものまで，広範に活用可能であることを示唆する結果を得た．

ところで，既に本稿におけるストリーミング課題でも，被験者自らが頭部運動する意識的運動と，ロボットが勝手に動く意識的では無いが物理的には等しい効果をもたらす運動の比較などを行えることを確認している[10]．また，頭部運動を手元のコントローラで操作することにより，頭部運動の意識は保ったまま，頸部の体性感覚の効果のみを検証し，体性感覚が音像定位に寄与することを示唆する結果も報告されている[11]．

このように，人間の外界との関わり，特に環境理解について考えるとき，「身体性」の重要さは，その物理的存在に留まらないと筆者は考える．今回新たに検討した複雑な聴覚現象の一つであるストリーミング課題においても，意識・情動といった問題が大きく関わってくる．従来，ロボティクス，特に創発ロボティクスの分野では，Pfeifer らの議論に基づき，身体性について，外界を理解するために用いる物理的フィルタの様に捉えてきた．しかし，知覚現象が物理世界の脳内表現の解釈の一つであるならば，そのフィルタには，意識・情動といったものも含まれるのではないか．そうだとすれば，意識的に行った運動の効果をバーチャルリアリティシステムによって，精確に再現し，意識や体性感覚の影響をそれ

ぞれ個別に取り除く手法は，人間理解の発展に多大な寄与をもたらすのではないかと考えている．

参考文献

1) R. Pfeifer and C. Scheler, “知の創成”, 共立出版, 2001.
2) R. A. Brooks, “Elephants don’t play chess”, Robotics and Autonomous Systems, vol.6 pp3-15, 1990.
3) T. Deacon, “Symbolic species: The co-evolution of language and the brain”, W. W. Norton and Company Inc, 1997.
4) J. Blauert, “Spatial hearing”, MIT Press, 1982.
5) I. Toshima, H. Uematsu, and T. Hirahara, “A steerable dummy head that tracks three-dimensional head movement: TeleHead”, Acoustical Science and Technology, vol. 24, no. 5, pp. 327-329, 2003.
6) I. Toshima, S. Aoki, and T. Hirahara, “Sound localization using an auditory telepresence robot: TeleHead II”, Presence, MIT Press, vol. 17, no. 4, pp.392-404, 2008.
7) 戸嶋巌樹，青木茂明，“音響テレプレゼンスロボットの頭部運動再現における聴覚的時間的余裕の定量的評価”，第 24 回ロボット学会学術講演会，
8) I. Toshima, S. Aoki, “Possibility of head-shape simplification for an acoustical telepresence robot: TeleHead”, Journal of Robotics and Mechatoronics, Japan Society of Mechanical Engineers, vol. 21, no. 2, pp. 223-228, 2009.
9) 平原達也，森川大輔，岩谷幸雄，“インターネット接続したテレヘッドによる聴覚テレプレゼンス”，音響学会秋期研究発表会，pp. 651-652, 2011
10) 近藤洋史，Daniel Presnitzer，戸嶋巌樹，柏野牧夫，“音脈分凝のリセットに対する頭部運動の影響”，聴覚研究会資料，vol. 41, pp. 365-370, 2011.
11) 吉崎大輔，平原達也頭頸部の体性感覚情報が水平面音像定位に及ぼす影響日本音響学会 2011 年秋期研究発表会 講演論文集, pp.479-480

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-B102-3

UI-ALT: 音の選択聴取を可能とする実世界アバタのためのユーザインタフェース

UI-ALT: User Interface for Avatar-based Listenable Telepresence

○植田 俊輔，今井 倫太，中臺 一博，中村 圭佑

Shunsuke Ueda, Michita Imai, Kazuhiro Nakadai and Keisuke Nakamura

慶應義塾大学

Keio University

(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.

ueda@ayu.ics.keio.ac.jp

Abstract

In a telepresence situation, a remote user has difficulties in catching and joining conversations because the user has to listen to the mixture of sound sources via a user interface. To relax this problem, this paper proposes User Interface for Avatar-based Listenable Telepresence (UI-ALT). A remote user can see scenes and listen to conversations via a real world avatar like a telepresence robot having a camera and microphone array. The user selects a conversation by marking persons of interests as a circle or a line on a UI-ALT display. The user can listen only to the selected conversation even when several conversations occur simultaneously because sound source separation with the microphone array eliminates non-target sound sources. Through offline evaluation, we showed the effectiveness of UI-ALT in a telepresence situation.

1 はじめに

人間は雑音環境においても音を聴き分けることができる．例えば，パーティのような多くの雑音が存在する環境の中でも人間は自分が興味のある会話を選択的に聴き取ることが出来る．この現象は「カクテルパーティ効果[1]」という名称で知られている．しかし，テレプレゼンスロボットがこのような雑音環境に置かれた場合，遠隔ユーザは遠隔地でどのような会話が行われているのかを理解することは困難である．

近年，テレプレゼンスアバタとしてのロボットが様々な方法で研究されており[2][3][4]，Anybots 社の QB[4]のように実用化されている例もある．これらのロボットは遠隔地で存在感を示し，人間の代わりにタスクをこなすことが期待されている．しかし，これらのロボットは人間とのインタラクションに必要な音声情報をうまく処理することが出来ないため，高雑音環境での人間とのインタラクションが難しいと考えられる．日常環境の中には大抵音声を含む複数個の音源が存在しており，人間とインタラクションを行うにはこうした複雑な音環境の理解が必要となる．

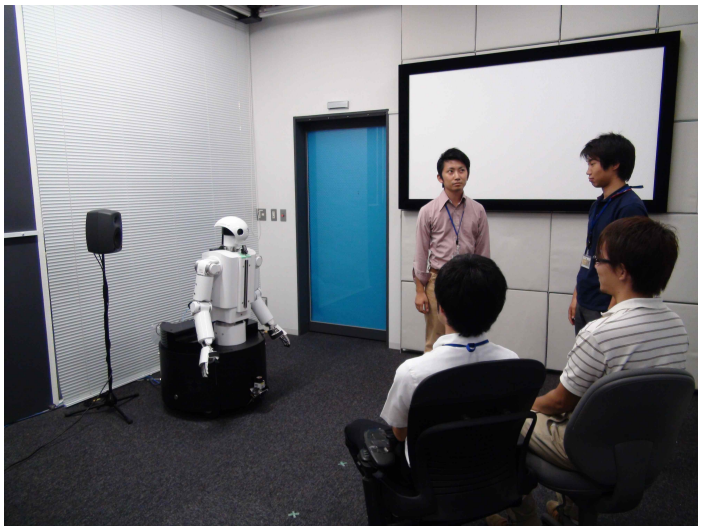


Figure 1: Avatar robot in a noisy room

本稿では，実世界アバタを対象として音の選択聴取を可能とするユーザインタフェース UI-ALT を提案する．UI-ALT はマイクロフォンアレイを搭載したアバタロボットを使用したインタフェースであり，マイクロフォンアレイ処理によって提供される音源定位および分離機能によりユーザは UI-ALT を通して望む方向の音を選択的に聴取することが出来る．つまり，UI-ALT を用いることで音の聴き分けを行うアバタロボットが実現可能である．

また，UI-ALT のユーザは UI 上で簡単なコマンドを入力することで音の選択聴取が出来る．水本らの研究[5]で

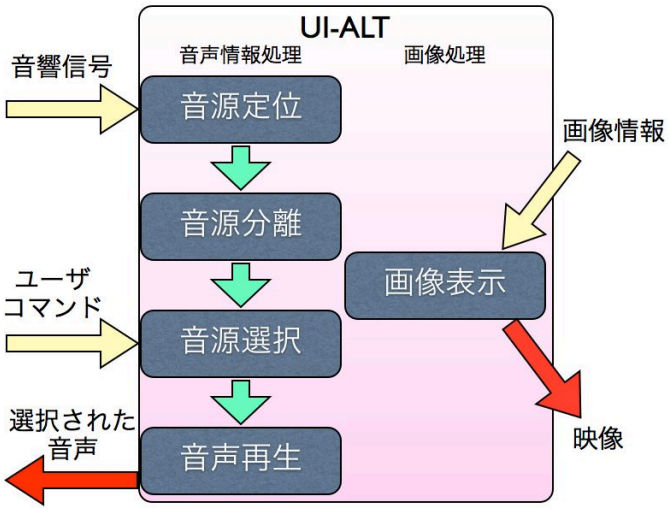


Figure 2: System architecture of UI-ALT

Figure 3: Snapshot of UI-ALT screen

は Willow Garage 社のテレプレゼンスロボット Texai に音の選択聴取が出来るユーザインタフェースを実装した．しかし水本らの研究では遠隔ユーザが分離された音を聴く際に，音の方向と幅の２つのパラメータを操作しなければならないため，実際にユーザが分離音を聴く際に煩雑な操作が要求される．このため，実際に遠隔ユーザはスムーズなインタラクションを行うことが出来ない．UI-ALT では，ユーザが UI の画面上で聴きたい方向を囲む，もしくは線を引くことで分離音が聴取出来るため，ユーザにとって簡単な操作で分離音を聴くことが出来る．

本稿は次の通りに展開する，第２節では UI-ALT のシステム構成について述べる．第３節では UI-ALT が応用可能なインタラクションの例について述べる．さらに第４節では，UI-ALT の有用性を示す為にオフラインで行ったディクテーション実験について述べ，最後に第５節でまとめと今後の課題を示す．

2　システムアーキテクチャ

UI-ALT のシステム構成図を図２に示す．

UI-ALT のユーザは，図３に示すように画面上で複数人が同時に喋っている中で会話を聴きたい方向の人に対してマウスを用いて線を引いたり丸で囲ったりすることでその方向の分離音を聴くことが出来る．この機能は，図２の中にあるオープンソースなロボット聴覚ソフトウェア HARK[6][7]を利用した音源定位・分離のモジュールによって実現される．

音源定位や分離，遠隔地のカメラ映像などはすべて ROS（Robot Operating System）[8][9]のメッセージで通信を行う．UI-ALT では音声データとカメラデータを同時に扱うため，処理が重くなってしまう可能性がある．そこで ROS が提供するメッセージを用いて通信を行うことにより，音声波形信号や音源 ID，カメラ画像情報など多種にわたるデータを小さい遅延で通信することが可能である．

以下の小節では UI-ALT の主要モジュールである音源定位，音源分離，音源選択の各モジュールについて詳しく述べる．

2.1　音源定位モジュール

入力である音響信号は最初に定位モジュールに送られる．定位モジュールではどの音がどの方向から来ているのかを推定することが出来る．音源の定位には HARK で提供されている雑音に頑健で，複数音源の定位が可能な MUSIC(MUltiple SIgnal Classification)[6]を用いる．MUSIC により，複数音源の水平方向の定位が可能となる．定位情報は入力音響信号とともに音源分離モジュールへ渡される．

2.2　音源分離モジュール

音源分離モジュールでは，選択的な会話の聴取を実現するために．定位情報と入力音響信号（混合音）から各音源信号を分離する．UI-ALT では HARK で提供されている GHDSS(Geometric-constrained Highorder Decorrelation-based Source Separation)[6]を用いて音源分離を行う．分離された音源情報は UI-ALT の音源選択モジュールへと送られる．

2.3　音源選択モジュール

音源選択モジュールはユーザのコマンドによって分離された音源を選択して音声再生モジュールに渡すモジュールである．ユーザがどの音源も選択していない場合は入力音響信号がそのまま再生モジュールに渡される．UI-ALT では図４に示すように選択したいグループを丸で囲う，選択したいグループの上に線を引くといった２種類の方法で音源選択をすることが出来る．

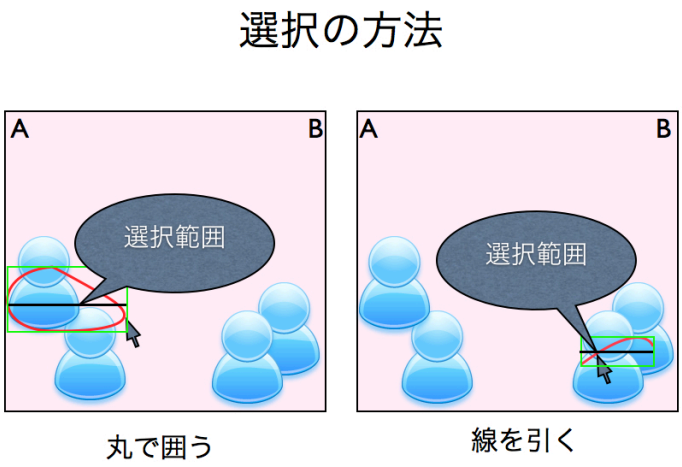


Figure 4: How to select sound source

ユーザがマウスモーションにより UI-ALT 上に円もしくは線を描き終わると，UI-ALT は以下の処理で音源選択を行う．

1. 描かれた円もしくは線の画像内における x, y 座標の最大値および最小値を取得する．
2. 選択範囲の x 座標の最大値および最小値をあらかじめ決められている USB カメラの画角から以下の式で角度に変換する．画像サイズは 640 × 480 であり，画像の中心が 0° である．

$$\theta = \pm \arctan\left(\frac{|x - 320| \times \tan\left(\frac{\text{カメラ画角}}{2}[deg]\right)}{320}\right) \quad (1)$$

3. 算出された角度範囲と音源の角度を比較して範囲に含まれていれば音源が選択されたと判断する．
4. 音声再生モジュールへ選択された分離音情報を送る．

UI-ALT では複数の音源を選択することも可能であり，複数選択された場合には選択された音源の数分の混合音が再生される．また音源選択を解除することも可能である．ユーザがマウスを右クリックすることで，音源選択状態をリセットして何も選択していない状態に戻すことが出来る．

3　応用可能なインタラクション場面

本節では，UI-ALT が実世界において応用可能であると考えられる場面について考察していく．具体例として以下に述べる 3 つの例を挙げる．

3.1　パーティ参加

ここでは，アバタロボットがパーティ会場にいて遠隔でユーザがパーティに参加する場面を考える．パーティ会場内では様々な場所で会話が行われていたり音楽が流れており，多様な音源が存在する．このため，遠隔ユーザはど

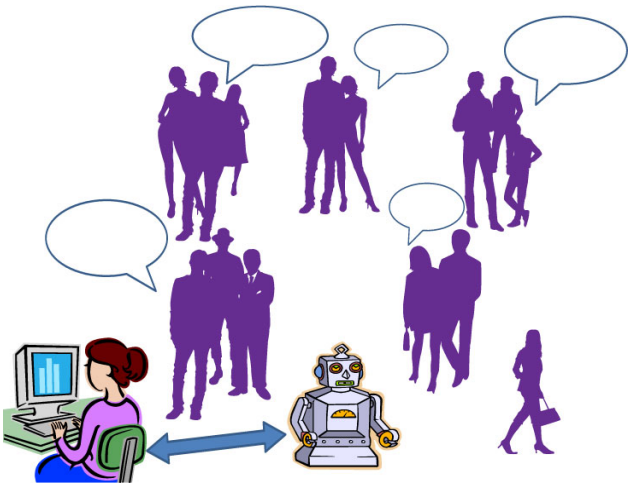


Figure 5: Avatar robot with UI-ALT at a party

のような会話が行われているのかを理解するのは難しい．仮に友人をパーティ会場で発見した場合でも遠隔ユーザは彼らが何を話しているのか理解することは難しい．そこでユーザは UI-ALT を用いて友人らを画面上で囲むことで友人らの会話の内容を聴くことができ，ユーザが実際にアバタロボットを操作して会話に参加することも可能になる．つまりユーザはまるで自分がそのパーティに参加しているかのような感覚を得ることができる．本例により，UI-ALT が可能とする音の選択聴取の有効性，またその結果として会話参加の容易性を表している．

3.2　レストランでの注文取り

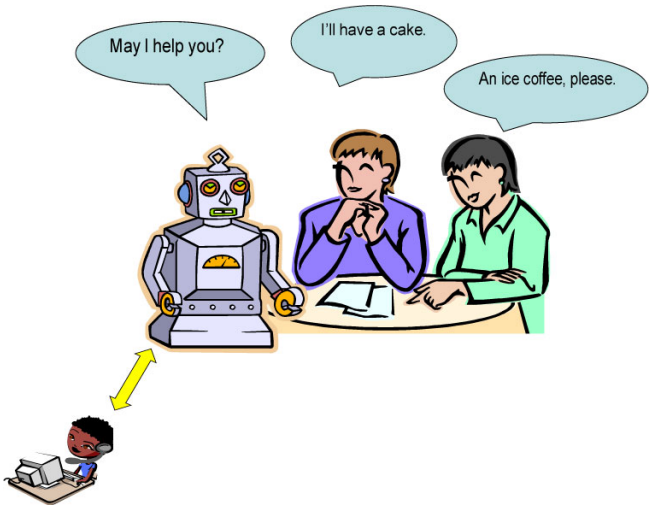


Figure 6: Avatar robot takes orders at a restaurant

ここでは，ファミリーレストランにおいてアバタロボットが従業員に変わって注文を取るという場面を考える．ファミリーレストランは家族連れなど様々な客層で賑わいをみせる場所であり，会話の音以外にも食事中に発生する音（フォークが皿に当たる音，グラスがぶつかる音など）が

ある高雑音環境である．遠隔ユーザはこのような雑音環境においても正しく注文を取るために，UI-ALT を用いて注文を取る人を画面上で選択することにより，遠隔ユーザは注文を正しく取るというタスクを遂行することができる．本例は，UI-ALT はファミリーレストランのような雑音環境において遠隔ユーザが対話タスクを遂行するために有用なシステムであることを表している．

3.3 会議

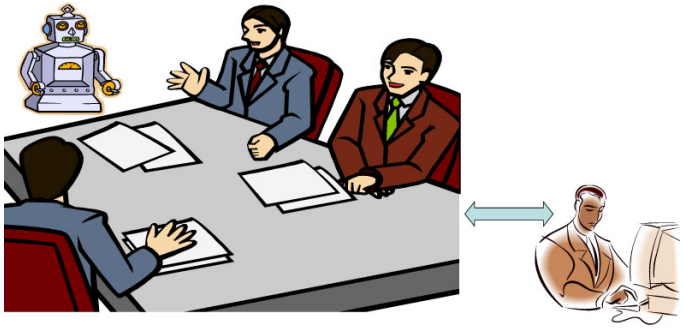


Figure 7: Avatar robot attends a meeting

ここでは，アバタロボットを通して遠隔でユーザが会議に参加する場面を考える．会議を行う際，時に活発な議論が行き過ぎて他の人の発言を聴かずに好き勝手に話し出してしまい，会議自体が収拾がつかないことがある．アバタロボットを通じて遠隔で会議の様子を見ているユーザにとっては会議室で発生しているすべての発言を聴き取ることは困難である．しかし，こうした発言の中に重要なキーワードが含まれている可能性もあるため，ユーザは出来るだけすべての発言を拾いたいと考える．UI-ALT を利用することで画面を見ながら気になる発言をしているユーザの発言を選択的に拾うことができる．UI-ALT は遠隔で会議のログを取る際にも有用であると考えられる．

以上で挙げた例から，日常環境におけるインタラクションにおいて，音声情報が必要不可欠であることがわかる．UI-ALT は雑音環境における人間とアバタロボットとのインタラクションに有用なユーザインタフェースとなりうる．

4 オフライン実験による評価

雑音環境における UI-ALT の有用性を示すために，本稿では UI-ALT を用いてユーザにディクテーションを行ってもらうオフライン実験を行った．本節では実験設定，実験結果，結果に対する考察を述べる．

4.1 実験設定

ユーザ実験を行う前に，別室でパーティ会場を想定した環境で音声と画像の録画を行った．実験室では雑音としてバックグラウンド音楽を流し，パーティに近い設定とした．4 人の大学生を実験室に集めて 2 人 1 組のグループを作ってもらい，図 8 のようにアバタロボットの正面から ±30° の方向に立ってもらった．音声の録音は頭部に 8 チャネルのマイクロフフォンアレイを搭載したアバタロボットを使用し，映像の録画には USB カメラを使用した．

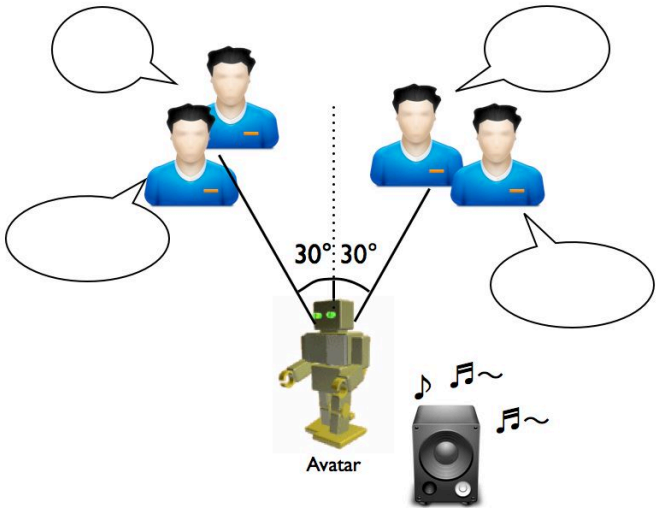


Figure 8: The location of the avatar robot and people during experiment. A loudspeaker plays background music.

ディクテーションのトピックとして両方のグループでお互いの自己紹介を行ってもらった．具体的な話題として，会話中にお互いの名前，出身，所属，趣味の 4 つの話題についてかならず触れてもらった．UI-ALT を使う場合と使わない場合を比較するために，同じようなシーンをグループ構成を変えて 2 種類録画を行った．各グループの発話の様子の例を図 9 に示す．

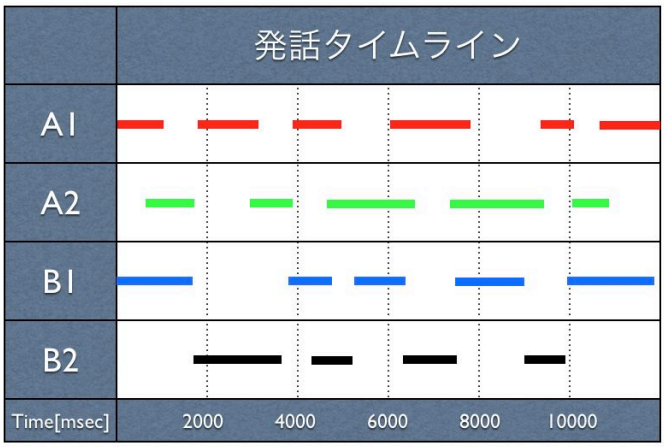


Figure 9: Timeline of each person's remark

本実験では UI-ALT のユーザとして 8 人の大学生に実

際に UI-ALT を用いてディクテーションタスクを行ってもらった．８人のうち４人は事前に UI-ALT の使い方を学ばずに使用してもらい，残りの４人は事前に１回だけ使い方を学んだ上で使用してもらった．実験では，各被験者は事前に撮影した２種類のビデオをランダムな順番で観てもらった．１度目は UI-ALT を使わずに映像と音声をそのまま流し，２度目は UI-ALT を用いて聴きたい会話を選択しながら実験を進めてもらった．被験者には映像内の２つのグループによる自己紹介の話題としてあげられていた内容を解答用紙に書き出してもらった．

本実験では我々は以下に挙げる項目について観察を行った．

- ディクテーションの正答率
- ユーザによる音源選択の仕方
- ユーザによる音源選択のスピード

4.2 結果

図 10 はディクテーションタスクにおける各被験者の正答率，事前練習を行わなかったグループの平均正答率，事前練習を行ったグループの平均正答率，および全体の平均正答率を UI-ALT を使った場合と使わなかった場合で比較した結果である．グラフの縦軸は正答率，横軸は各ユーザの ID を表す．UI-ALT を使った場合の全体の平均正答率は 76%であったのに対し，UI-ALT を使わなかった場合の平均正答率は 35%にとどまった．また，UI-ALT を事前に練習しなかったグループが UI-ALT を使った場合の正答率が 67%であったのに対し，UI-ALT を事前に練習したグループが UI-ALT を使った場合の正答率は 85%となった．平均正答率の結果を見ると，ユーザが UI-ALT を使った場合は使わなかった場合より２つのグループの会話の内容が理解出来ているということが言える．

ユーザによる音源選択の仕方については，一つのグループを長い時間選択しているユーザもいれば，頻繁に選択するグループを変えるユーザも見受けられた．選択のスピードについても，素早く選択しているユーザもいれば，ゆっくり選択しているユーザも見受けられた．

4.3 考察

実験結果より，UI-ALT を使った場合，ユーザのディクテーションの正答率にかなりの向上が見受けられる．このことから，UI-ALT は高雑音環境であっても会話内容の理解を支援するツールであると言える．

しかし，UI-ALT を事前に練習しなかったグループの中に，どちらの音声を選択してよいかわからずにビデオの再生が終わってしまい，ディクテーションタスクに回答出来なかったユーザも存在した．この現象の原因の一つとして考えられるのは UI-ALT の映像のフレームレートの低さである．今回の実験では遅延をなるべく小さく抑えるためにビデオのフレームレートを落として実験を行った．しかし，実験後に行ったアンケートからユーザは話者を選択する際に話者の口元や表情を見てある程度決めているという知見が得られた．ディクテーションタスクに回答出来なかったユーザはどちらのグループが何の話題について話していたのかが音声情報だけでは理解出来ず，映像のフレームレートも悪かったためにどちらのグループを選択してよいか混乱してしまったと考えられる．このことから，音源選択の際には視覚情報が聴覚情報と同じぐらい重要な役割を果たしているということが言える．

また，実験後のアンケート結果から，被験者のうちの半数が UI のマウス操作が複雑なため音源選択に苦労したという回答を得た．実験から，ユーザによって選択の仕方やスピードの違いが様々異なることが見受けられたが，ディクテーションタスクの正答率と比較してみると，素早く選択しているユーザほどより良い正答率を出しているという傾向が見られた．このことから，UI-ALT は音源選択の際に有効ではあるが，必ずしもすべてのユーザに対して直感的なインタフェースではないことがわかる．今後はユーザが望む音源を素早く選択出来るように最適な選択方法を調べていく必要がある．

5 まとめ

本稿では，実世界アバタを対象として，音の選択聴取機能を有するユーザインタフェース UI-ALT を提案した．UI-ALT は人間とアバタロボットとのインタラクションにおいて欠かすことの出来ない音声情報を扱えるインタフェースであるため，実世界の様々な環境に適用可能であると考えられる．本稿では実際に UI-ALT の応用が可能であると考えられる３つのインタラクションシナリオを示し，UI-ALT を用いることによって遠隔ユーザが雑音環境の音をアバタを通して聴く際に聴きやすくなったことをディクテーション実験により示した．

今後の課題として，まずインタフェースの改善が挙げられる．オフライン実験から，ユーザは話者を選択する際にある程度画面を見ながら選択しているという傾向が見られたので，UI の画像を見やすくする必要がある．また，選択の仕方も人それぞれであるということから，どのような選択の仕方が一番ユーザに取って使いやすいのかを調査する必要がある．

UI-ALT を用いたオンライン実験も計画している．今回のオフライン実験で得られた知見を基にアバタロボットを操作出来るようインタフェースを改良し，実際にパーティにアバタロボットを参加させて遠隔でユーザに参加してもらうといった実験を行っていく予定である．

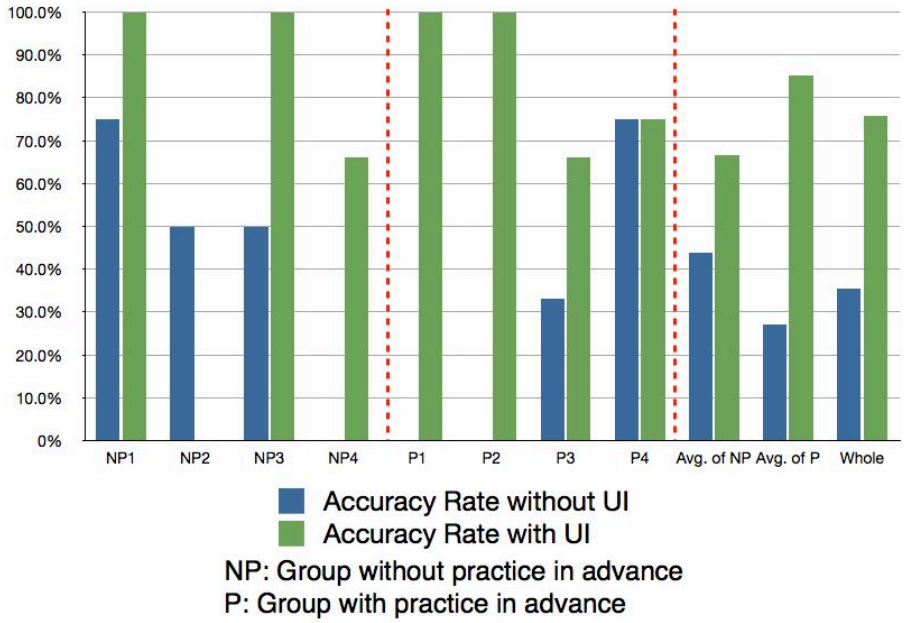


Figure 10: Result of accuracy rate in dictation task

参考文献

[1] Cherry E. Colin: Some Experiments on the Recognition of Speech, with One and with Two Ears, in *The Journal of the Acoustical Society of America*, *vol.25*, pp.975–979, 1953.

[2] Sigurdur Orn Adalegeirsson, Cynthia Brezeal: Mebot a robotic platform for socially embodied telepresence. in *Proc. of ACM/IEEE International Conference on Human-Robot Interaction(HRI)*, pp.15-22, 2010.

[3] Nishio, S, Ishiguro, H., Anderson, M., Hagita, N.: Repesentating personal presence with a teleoperated android: A case study with family. in *Proc. of AAAI 2008 Spring Symposium on Emotion, Personality, and Social Behavior*, pp.96-103, 2008.

[4] Anybots -Your Personal Avatar- : `http://www.anybots.com` .

[5] Takeshi Mizumoto, Takami Yoshida, Kazuhiro Nakadai, Ryu Takeda, Takuma Ohtsuka, Toru Takahashi, Hiroshi G. Okuno: Design and Implementation of Selectable Sound Separation on a Texai Telepresence System Using HARK in *Proc. of IEEE-RAS International Conference on Robotics and Automation(ICRA)*, pp.2130-2137, 2011.

[6] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, Huroshi Tsujino: Design and Implementation of Robot Audition System "HARK" in *Advanced Robotics*, *vol.24* pp.739-761, 2010.

[7] HARK Main Page: `http://winnie.kuis.kyoto-u.ac.jp/HARK/` .

[8] Morgan Quigley, Brian Gerkey, Ken Conley, Josh Faust, Tully Foote, Jeremy Leibs, Eric Berger, Rob Wheeler, Andrew Ng: ROS: an open-source Robot Operating System in *IEEE-RAS International Conference on Robotics and Automation (ICRA) Workshop on Open Source Software in Robotics*, 2009.

[9] ROS: `http://www.ros.org`

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-B102-4

SLAM に基づく非同期分散マイクロホンアレイのキャリブレーションの評価

Evaluation of a SLAM-based Calibration Method for Asynchronous Microphone Arrays

三浦弘樹 † **吉田尚水** † **中村佳佑** ‡ **中臺一博** †,‡

Hiroki MIURA, Takami YOSHIDA, Keisuke NAKAMURA, Kazuhiro NAKADAI

† 東京工業大学大学院　情報理工学研究科

‡(株) ホンダ・リサーチ・インスティチュート・ジャパン

Abstract

This paper evaluates an online calibration method for asynchronous microphone arrays. Conventional microphone array techniques require a lot of measurements of transfer functions to calibrate microphone locations, and a multi-channel A/D converter for inter-microphone synchronization. To solve these two problems, we proposed an online framework combining Simultaneous Localization and Mapping (SLAM) and beamforming and an implemented prototype system using an Extended Kalman Filter (EKF) showed the feasibility of the proposed framework in a simulated and a real environment. In this paper, we show the robustness of the proposed framework for different motion models, motion and observation errors to apply to real microphone array systems through numerical experiments.

1 はじめに

マイクロホンアレイ処理はロボット聴覚分野における音源定位や音源分離に有用であり，数多くの研究が報告されている [1, 2, 3, 4]．これらのマイクロホンアレイ処理には, 各マイクロホンの位置もしくは音源とマイクロホンアレイ間の伝達関数が既知であること，全チャネルを同期収録することが必要とされる．我々は，これらの問題をオンラインで解くため，拡張カルマンフィルタ (Extended Kalman Filter, EKF) に基づく Simultaneous Localization and Mapping (SLAM) と遅延和ビームフォーミングを組み合わせた手法を提案し, マイクロホンアレイの周りを人（音源）が十数回手を叩きながら歩くだけで非同期マイクロホンアレイのキャリブレーションが可能であることを実験により示した [5].

しかし，提案手法の評価は限定された環境で行われており，システムのパラメータがキャリブレーション性能にどのように影響するのかといった評価はされていなかった．

本稿では，提案手法の適用範囲を知り，実用化に向けた課題を明らかにするために，マイクロホンの初期配置，運動誤差，観測誤差に対するキャリブレーション性能の頑健性を評価する．状態遷移モデルには [5] で用いた長方形軌道に加え円運動を，各マイクロホンの初期配置には一様分布の場合と平均が真のマイク位置，標準偏差が (0.1, 0.5, 1.0) に従う乱数の場合を，運動誤差と観測誤差には実測値から求めた標準偏差，その 10 倍およびその 100 倍の場合をそれぞれ考慮し，その精度と収束速度を数値実験により評価した．

2 非同期分散マイクロホンアレイと問題の定式化

本稿では，非同期分散マイクロホンアレイを，各マイクロホンの位置が未知であり，各マイクロホンの時刻にずれがある (非同期) マイクロホンアレイとして定義する．非同期分散マイクロホンアレイを用いて，音源定位や音源分離といったマイクロホンアレイ処理が可能になれば，煩わしい伝達関数の計測作業や高価な多チャンネル同期 A/D デバイスが不要になり，より実用的な処理が実現できる．この非同期分散マイクロホンアレイを用いて，音源の位置，各マイクロホンの位置，同期時刻のずれを推定するキャリブレーション問題を Blind Alignment 問題と定義する．Blind Alignment 問題は，従来にも研究報告があり，例えば，Thrun らは，事前に各マイクロホンの位置が未知という条件の下, マイクロホン位置のオンラインキャリブレーションを実際にマイクロホンを用いて報告している [6]．しかし，彼らの手法では，音源位置は既知，マイクロホンは完全に同期されている必要があったまた，Ono

Table 1: Notation

N	マイクロホンの総数
K	発音の総数
c	音速
n	マイクロホンのインデックス
k	発音のインデックス
ω	周波数
l	EKF-SLAM における時間ステップ
$x_{s[k]}, y_{s[k]}, \tau_{s[k]}$	k 回目に発音した位置と時刻
$\boldsymbol{\xi}_{s[k]} = [x_{s[k]}, y_{s[k]}, \theta_{s[k]}]^T$	人の位置と向き
$\boldsymbol{\xi}_{mn} = [x_{mn}, y_{mn}, \tau_{mn}]^T$	マイクロホンの位置と同期時刻ずれ
$\boldsymbol{\xi}_m = [\boldsymbol{\xi}_{m1}, \ldots, \boldsymbol{\xi}_{mN}]^T$	マイクロホンの位置
$S_{[k]}(\omega)$	k 回目に発音した音
$X_{n[k]}(\omega)$	マイク n が観測した k 回目の音
$\boldsymbol{X}_{[k]}(\omega)$	$[X_{1[k]}(\omega), \ldots, X_{N[k]}(\omega)]^T$
$\boldsymbol{A}(\omega)$	音源とマイクロホンの間の伝達関数

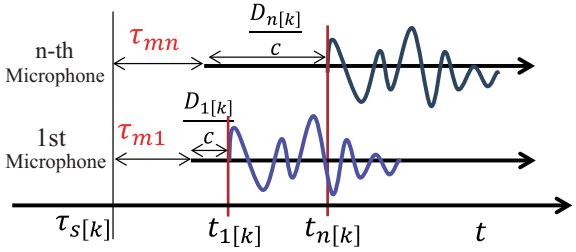


Figure 1: Observation model for each microphone

らは観測した音のみから音源位置，各マイクロホン位置，同期時刻ずれを推定する問題に取り組み，マイクロホンや発音回数など理論的に必要な条件を明らかにした [7]．しかし，彼らの手法はオフラインの手法である，計算量コスト大きい, 前もってキャリブレーションの推定回数を指定する必要があるといった問題があった．

これに対して，我々が研究を行っている手法 [5] は，オンラインで Blind Alignment 問題を解決することができる．つまり，位置が未知で，かつ，完全な同期収録が保証できない，非同期分散マイクロホンアレイのオンラインキャリブレーションが可能である．具体的には，SLAM を用いて，その地図推定を各マイクロホンの位置推定，自己位置推定を音源の位置推定に当てはめ，同期時刻のずれを含む推定誤差を最小になるように推定値を更新することによって，オンラインキャリブレーションを行う．

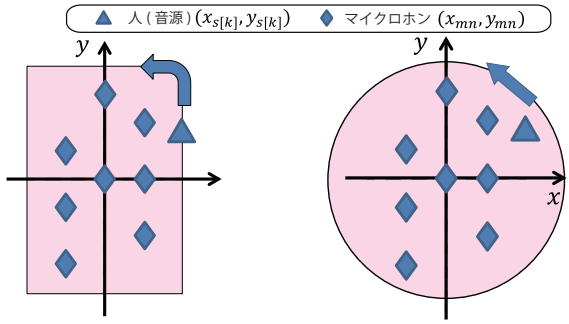
以降，手法の定式化を行う．まず，本稿では，音源はインパルスとして問題を定式化する．なお，本稿で用いる表記を Tab. 1 にまとめた．

2.1 伝達関数（ステアリングベクトル）モデル

マイクロホンで観測される信号は，マイクロホンと音源間の伝達関数を用いて，以下のように表すことができる．

$$\boldsymbol{X}_{[k]}(\omega) = \boldsymbol{A}(\omega) S_{[k]}(\omega) \tag{1}$$

この伝達関数 $\boldsymbol{A}(\omega)$ は，直接音のみを考慮することで，音源位置とマイク位置を用いて以下のように近似計算する



a) Rectangular motion　　b) Circular motion

Figure 2: Motion Models

ことができる．

$$\begin{aligned}\boldsymbol{A}(\omega) &\approx \boldsymbol{A}(\boldsymbol{\xi}_{s[k]}, \boldsymbol{\xi}_m, \omega) \\ &= [\exp(\ \ 2\pi j\omega t_{1[k]}), \ldots, \exp(\ \ 2\pi j\omega t_{N[k]})]^T. \end{aligned} \tag{2}$$

ここで，$t_{n[k]}$ は，マイクロホン n が k 回目に発せられた音を観測した時刻である．$t_{n[k]}$ は，Fig. 1 に示すように，音源が音を発した時刻 $\tau_{s[k]}$ を用いて，以下のように求めることができる ($D_{n[k]}$ はマイクロホン n と音源間の距離)．

$$t_{n[k]} = \tau_{s[k]} + \frac{D_{n[k]}}{c} + \tau_{mn}\ , \tag{3}$$

$$D_{n[k]} = \sqrt{(x_{s[k]} \quad x_{mn})^2 + (y_{s[k]} \quad y_{mn})^2}\ . \tag{4}$$

$\boldsymbol{A}(\boldsymbol{\xi}_{s[k]}, \boldsymbol{\xi}_m, \omega)$ は，音源定位で用いる際はステアリングベクトルとも呼ばれる．従来の音源定位手法 [8] ではこのステアリングベクトルを事前計測する必要があった．しかし，測定には設備が必要で，かつ時間がかかるため，簡単に計測することは難しい．提案手法では，$\boldsymbol{\xi}_{s[k]}$ と $\boldsymbol{\xi}_m$ が推定可能であり，$\boldsymbol{A}(\boldsymbol{\xi}_{s[k]}, \boldsymbol{\xi}_m, \omega)$ を事前計測なしに得ることができる．

2.2 状態遷移モデル

音源（人）の移動モデルは一般的には次の式で表される．

$$\boldsymbol{\xi}_{s[l+1]} = g(\boldsymbol{\xi}_{s[l]}, \eta_{[l]}) + \boldsymbol{w}_{s[l]} \tag{5}$$

ただし，$\eta_{[l]}$ は入力を表し，$\boldsymbol{w}_{s[l]}$ は平均 $\boldsymbol{0}$，分散 $[\sigma_x^2, \sigma_y^2, \sigma_\theta^2]$ の正規分布に従うモデル誤差を表す．ここで，$g(\boldsymbol{\xi}_{s[l]}, \eta_{[l]})$ は自由に設計できる．本稿では Fig. 2a)，b) に示すように長方形軌道を描く長方形運動モデルと円軌道を描く円運動モデルの 2 つを構築する．

なお，各マイクロホンの位置は動かないので，状態遷移モデルは音源（人）のみに対して構築する．

2.2.1 長方形運動モデル

長方形運動モデルは，以下の式で表される．

$$g(\boldsymbol{\xi}_{s[l]}, \eta_{[l]}) = \boldsymbol{\xi}_{s[l]} + \begin{bmatrix} \sin(\ \ _{s[l]}) & 0 \\ \cos(\ \ _{s[l]}) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_{s[l]} \\ u_{s[l]} \end{bmatrix}. \tag{6}$$

入力 $\eta_{[l]} = [v_{s[l]}, u_{s[l]}]^T$ は音源の移動速度と角速度を表し，角速度 $u_{s[l]}$ は長方形の四隅に音源が到達した時に 90 度回転させ，それ以外のときは直進 (0 度) である．

2.2.2 円運動モデル

円運動モデルは，半径一定の円を目標軌道としており，以下の式で表される．

$$g(\boldsymbol{\xi}_{s[l]}, \eta_{[l]}) = \begin{bmatrix} \cos(\Delta\ ) & \sin(\Delta\ ) & 0 \\ \sin(\Delta\ ) & \cos(\Delta\ ) & 0 \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{\xi}_{s[l]} + \begin{bmatrix} 0 \\ 0 \\ \Delta \end{bmatrix}. \tag{7}$$

入力 $\eta_{[l]} = \Delta$ は音源の 1 歩あたりの移動角度を表す．

2.3 観測モデル

観測は，k 番目のインパルスの到達時刻 $t_{n[k]}$ である．音を発した時刻 $\tau_{s[k]}$ は未知であるため，基準マイクロホン (マイクロホン 1) での観測時刻との差をとると，観測モデルは，以下のように相対時刻で表すことができる．

$$\boldsymbol{\zeta}_{[k]} = \begin{bmatrix} \frac{D_{2[k]} - D_{1[k]}}{c} + \tau_{m2} & \tau_{m1} \\ \vdots & \\ \frac{D_{N[k]} - D_{1[k]}}{c} + \tau_{mN} & \tau_{m1} \end{bmatrix} + \boldsymbol{\delta}_{[k]} \tag{8}$$

観測誤差 $\boldsymbol{\delta}_{[k]}$ は平均 0 分散 σ_r^2 の正規分布に従うものとする．

3 非同期分散マイクロホンアレイのキャリブレーション

提案法は，EKF-SLAM を用い，予測，観測，更新ステップを繰り返すことでキャリブレーションを行う．

予測ステップ 音源状態の平均 $\hat{\boldsymbol{\xi}}_{[l]}$ と分散 $\hat{\boldsymbol{P}}_{[l]}$ は以下のように計算される．

$$\hat{\boldsymbol{\xi}}_{s[l|l-1]} = g(\hat{\boldsymbol{\xi}}_{s[l-1]}, \eta_{[l-1]}) \tag{9}$$

$$\hat{\boldsymbol{P}}_{[l|l-1]} = \boldsymbol{G}_{[l]} \hat{\boldsymbol{P}}_{[l-1]} \boldsymbol{G}_{[l]}^T + \boldsymbol{F}^T \boldsymbol{R} \boldsymbol{F} \tag{10}$$

$$\boldsymbol{F} = [\boldsymbol{I}^{3\times 3}, \boldsymbol{O}^{3\times 3N}] \tag{11}$$

ここで $\boldsymbol{R}$ は $\boldsymbol{R} = \mathrm{diag}(\sigma_x^2, \sigma_y^2, \sigma_a^2)$ で表される共分散行列であり，$\boldsymbol{G}_{[l]}$ は次式で示される状態遷移モデルのヤコビアンである．

$$\boldsymbol{G}_{[l]} = \frac{\partial g(\boldsymbol{\xi_s}, \eta)}{\partial \boldsymbol{\xi_s}}|_{\boldsymbol{\xi_s} = \hat{\boldsymbol{\xi}}_{s[l-1]}} \tag{12}$$

$$= \begin{cases} \boldsymbol{I} + \boldsymbol{F}^T \begin{bmatrix} 0 & 0 & -v_{s[l]} \sin(\theta_{s[l]}) \\ 0 & 0 & v_{s[l]} \cos(\theta_{s[l]}) \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{F} & \text{(長方形運動モデル)}, \\ \boldsymbol{I} + \boldsymbol{F}^T \begin{bmatrix} \cos(\Delta\theta) & -\sin(\Delta\theta) & 0 \\ \sin(\Delta\theta) & \cos(\Delta\theta) & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{F} & \text{(円運動モデル)} \end{cases} \tag{13}$$

観測ステップ 各マイクロホンでの観測時刻から，式 (8) に従って，以下を得る．

$$\boldsymbol{h}(\hat{\boldsymbol{\xi}}_{[k|k-1]}) = \begin{bmatrix} \frac{\hat{D}_{2[k]} - \hat{D}_{1[k]}}{c} + \hat{\tau}_{m2} & \hat{\tau}_{m1} \\ \vdots & \\ \frac{\hat{D}_{N[k]} - \hat{D}_{1[k]}}{c} + \hat{\tau}_{mN} & \hat{\tau}_{m1} \end{bmatrix} \tag{14}$$

更新ステップ 予測ステップと観測ステップを元に，音源の位置・向きとマイクロホンの位置・同期時刻のずれの推定値を更新する．まず，$\boldsymbol{h}(\hat{\boldsymbol{\xi}}_{[k|k-1]})$ と $\boldsymbol{\zeta}_{[k]}$ の差を最小にするようにカルマンゲインを導出する．

$$\boldsymbol{K}_{[k]} = \boldsymbol{P}_{[k|k-1]} \boldsymbol{H}_{[k]}^T \left(\boldsymbol{H}_{[k]} \boldsymbol{P}_{[k|k-1]} \boldsymbol{H}_{[k]}^T + \boldsymbol{Q}_{[k]} \right)^{-1} \tag{15}$$

ここで，$\boldsymbol{H}_{[k]} = \frac{\partial \boldsymbol{h}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}|_{\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}_{[k|k-1]}}$ は観測モデルのヤコビアンであり，$\boldsymbol{Q}_{[k]}$ は $\boldsymbol{Q}_{[k]} = \mathrm{diag}(\sigma_r^2, \ldots, \sigma_r^2)$ で定義される共分散行列である．

求めたカルマンゲインを用いて，推定値を以下のように更新する．

$$\hat{\boldsymbol{\xi}}_{[k]} = \hat{\boldsymbol{\xi}}_{[k|k-1]} + \boldsymbol{K}_{[k]} \left(\boldsymbol{\zeta}_{[k]} \quad \boldsymbol{h}(\hat{\boldsymbol{\xi}}_{[k|k-1]}) \right), \tag{16}$$

$$\hat{\boldsymbol{P}}_{[k]} = (\boldsymbol{I} \quad \boldsymbol{K}_{[k]} \boldsymbol{H}_{[k]}) \hat{\boldsymbol{P}}_{[k|k-1]}. \tag{17}$$

4 キャリブレーション性能の評価

ここでは，以下の条件においてキャリブレーションの収束速度と収束後の音源位置・マイク位置・同期時刻のずれの推定精度を評価した．

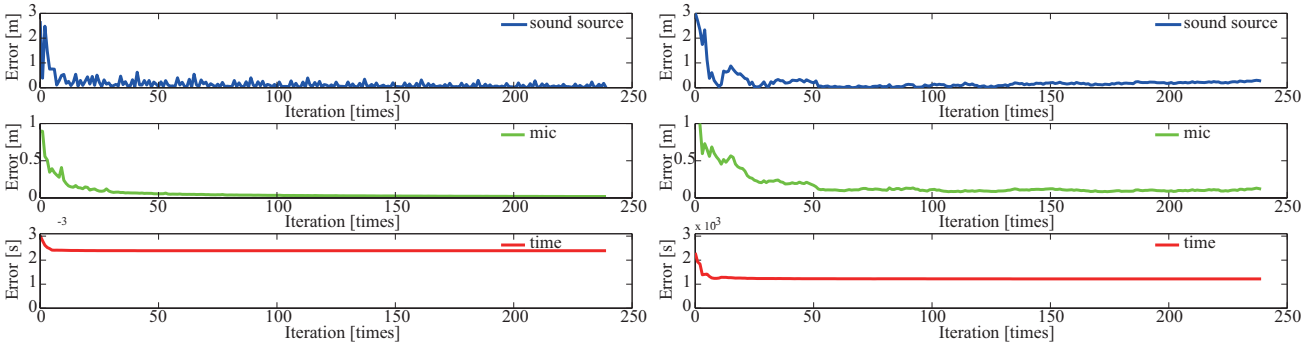
状態遷移モデル 長方形運動モデル・円運動モデル
各マイクの初期値 一様分布・真値を平均とした正規分布
運動誤差 実測値・その 10 倍・その 100 倍 (σ_x, σ_y)
実測値・その 25 倍・その 100 倍 (σ_θ)
観測誤差 実測値・その 10 倍・その 100 倍

なお，更新ステップによるマイクロホンの推定位置の変化量 $\hat{\xi}_m[k+1] \quad \hat{\xi}_m[k]$ が平均で 0.01 [m] 以下になったら収束したとみなし，それまでの発音回数を収束速度とした．

各パラメータは，77 回の拍手を収録した実測データ [5] から算出した値を基準にした．

実測データの収録条件 ハードウェアには，(株) システムインフロンティア社製の多チャンネル録音機器 RASP24 と MEMS マイクロホンを用い，8 ch，24 bits，16 kHz サンプリングで収録した．観測誤差には A/D コンバータの影響，配線長による影響，マイクと音源位置の計測誤差，そして観測された音の波形からの到達時刻の抽出精度が含まれる．しかし，A/D コンバータと配線長による影響は到達時刻の抽出による誤差に比べ小さいので無視するものとする．

実験時には 1.2 m × 2.4 m の机の上に 8 チャネルマイクロホンアレイを配置し，音源 (拍手) を一定間隔で動かすことによって 77 回録音した．得られた音の波形から到達時刻を各チャネルごとに抽出した．あらかじめ計測しておいた音源位置，各マイクロホン位置から到達時間差を計算し，マイクロホン 1 を基準とした観測モデルの観測誤差を計算した．Fig.4 はマイクロホンの観測誤差を分布を示すヒストグラムであり，平均 2.75×10^{-4} [s]，標準偏差



a) Rectangular motion　　b) Circular motion

Figure 3: Calibration results

2.1×10^{-3} [s] であった．サンプリング周波数が 16 kHz であるので，観測誤差の標準偏差は 16 サンプル程度である．

マイクロホンをランダムに部屋に配置し，長方形運動モデルと円運動モデルをそれぞれ用い，数値実験により評価する．さらに，実際の観測誤差を計測し，移動モデル誤差に対する頑健性を検証し，キャリブレーション性能を評価する．数値実験では，1.2 [m] × 2.4 [m] の部屋 (長方形運動モデル)，と半径 1.2 [m] の部屋 (円運動モデル) を想定して検証を行った．マイクロホン数は 8 であり，図 2 に従い配置した．基準となるマイクロホン 1 の位置を原点とし，回転方向の曖昧性を解消するため，マイクロホン 2 の位置は y 座標を 0，x 座標を正とした．音源はインパルスを想定し，部屋の隅に沿って反時計回りに移動する．初期位置は実際の初期位置である部屋の左下隅座標に対して，平均 0 [m]，標準偏差 0.5 [m] に従うガウス雑音を与えた．音源の移動は，1 歩あたり 0.3 [m] とし，5 歩進むごとに 1 回音を発するものした．状態遷移モデルの位置と角度の標準偏差はそれぞれ 0.1 [m]，1 [度]，観測誤差の標準偏差は 0.5×10^{-3} [s] (0.17 [m] に相当) とした．各マイクロホンの時刻のずれは固定であり，初期状態では，ずれは 0 [s]，標準偏差を 0.1 [s] とした．

4.1　状態遷移モデルのキャリブレーション評価

図 3a)，b) はそれぞれ長方形運動モデル，円運動モデルを用いたマイクロホンアレイのキャリブレーションの結果を示す．マイクロホン位置の平均誤差と同期時刻のずれの誤差には明確な差は無いが，長方形運動モデルを用いた場合の人位置の誤差は振動的になっている．これは人が部屋の壁にたどり着いたら直角に曲がるという非線形性の強いモデルを使用していることが原因と考えられる．一方円運動モデルでは長方形運動モデルほど振動的でないことがわかる．

4.2　分散パラメータに対するキャリブレーションの性能

実測した観測誤差の分散 $\sigma_r^2 = 4.41\times10^{-6}[s^2]$ を用いマイクロホンアレイのキャリブレーションの性能を評価する．状態遷移モデルは長方形運動モデルとし，誤差は期待値 0[m]，

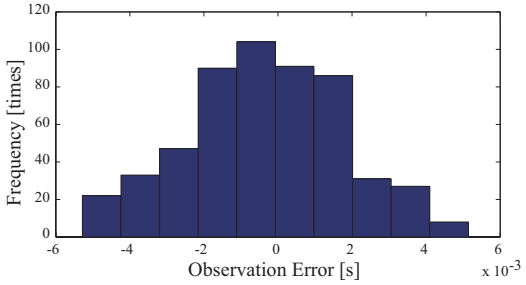


Figure 4: Histogram of Observation Errors

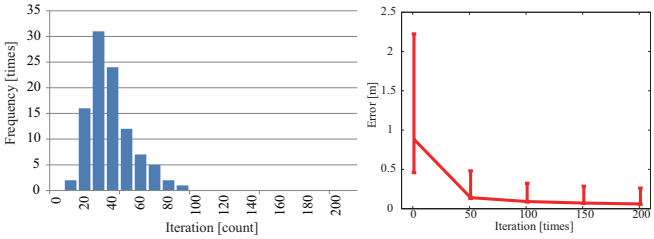
分散 $[\sigma_x^2, \sigma_y^2, \sigma_\theta^2] = [1.0\times10^{-2}[m^2], 1.0\times10^{-2}[m^2], 1.75\times10^{-2}[rad^2]]$ に従うガウス分布とする．マイクロホンの初期位置は一様分布に従い発生させ，数値シミュレーションにより性能検証した．シミュレーションは 100 回行い，その平均を求めた．

図 5–7 の左図はマイクロホン位置推定が収束したときの手を叩いた回数のヒストグラムを表す．横軸が収束するまでにかかったインパルス回数，縦軸が度数 (総計 100 回) である．ここで，収束とは変化率が 1.0×10^{-2} [m] を下回った時とする．

また，右図は，マイクロホンの位置推定誤差の平均を折れ線グラフで，最大値，最小値をエラーバーで示したものである．縦軸は各マイクロホン位置推定の平均誤差，横軸はインパルス回数であり 50 回ごとに平均計算を行った．

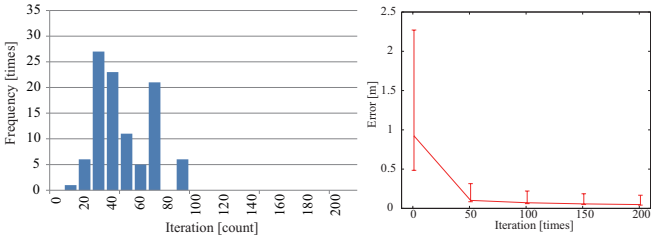
図 6 は，図 5 で用いた σ_r^2 を 10 倍，100 倍に変化させた場合の結果であり，図 7 は，図 5 の $[\sigma_x^2, \sigma_y^2, \sigma_\theta^2]$ を $[\times25, \times25, \times1]$ (a),b))，$[\times100, \times100, \times1]$ (c),d))，$[\times1, \times1, \times10]$ (e),f))，$[\times1, \times1, \times100]$ (g),h)) と変化させた場合の結果である．

図 6, 7 の左図から，収束までに必要なインパルス回数は，観測，状態遷移モデルに関わらず，実際の分散の 10 倍程度までなら，ほとんど変化がないことがわかる．100 倍程度になると，ヒストグラムの形が崩れ，収束までの時間が大きくなることがわかる．つまり，これらの分散は，実際の値の 10 倍程度までの値を設定する必要があると言える．一方，右図からは，インパルス回数に対するマイク

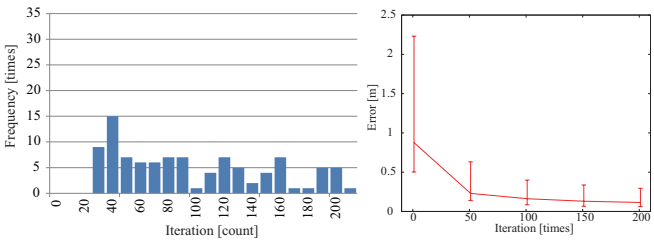


a) histogram　　b) average error

Figure 5: Reference Result



a) histogram $(\sigma_r^2 \times 10)$　　b) average error $(\sigma_r^2 \times 10)$



c) histogram $(\sigma_r^2 \times 100)$　　d) average error $(\sigma_r^2 \times 100)$

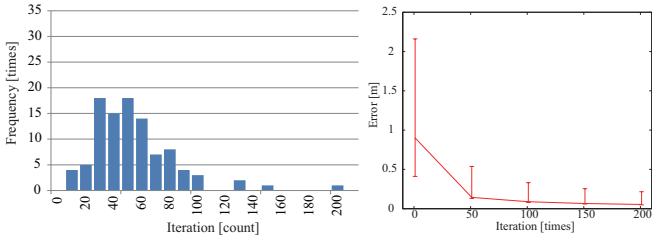
Figure 6: Changes of observation variences

ロホンの位置推定誤差の傾向は，どの場合も同様であるということが言える．50 回程度までは，推定性能が向上するものの，それを超えると推定性能の向上は緩やかになる．例えば，必要な精度を 0.2 [m] とすると，分散の値が 100 倍以上ずれていたとしても，50 回程度のインパルスで収束するということが言える．実際に，文献 [5] では，音源定位で用いるビームフォーミングの解像度が 0.2 [m] を用いていた．従って，これらの図から，観測，状態遷移モデルの誤差分散は，実際の値の 10 倍程度までなら，収束にほとんど影響しないこと，また，100 倍程度であっても，必要な解像度によっては，十分実用に耐えうることが分かった．

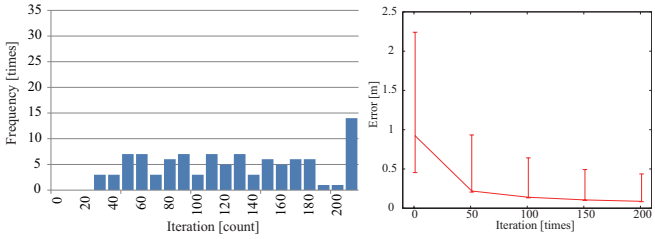
4.3　マイク初期位置に対するキャリブレーションの性能

図 8a)–f) はマイクロホンの初期配置を変化させた時のキャリブレーション性能を示している．左図，右図は，4.2 節と同様に，収束時のインパルス回数のヒストグラムとインパルス回数に対するマイクロホン位置推定の平均誤差を示している．

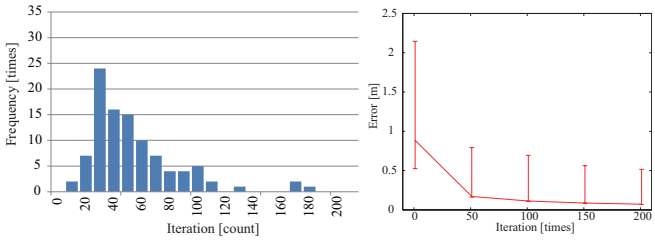
図 8a),b) は，実際のマイクロホンの位置に対する誤差



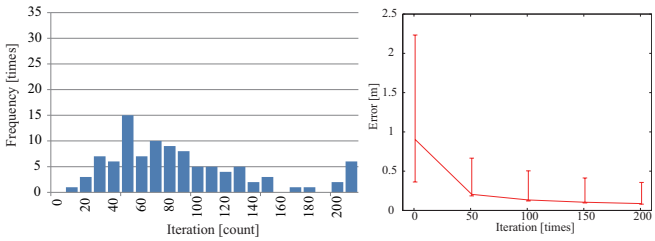
a) histogram $(\sigma_x^2 \times 25, \sigma_y^2 \times 25, \sigma_\theta^2 \times 1)$　　b) average error $(\sigma_x^2 \times 25, \sigma_y^2 \times 25, \sigma_\theta^2 \times 1)$



c) histogram $(\sigma_x^2 \times 100, \sigma_y^2 \times 100, \sigma_\theta^2 \times 1)$　　d) average error $(\sigma_x^2 \times 100, \sigma_y^2 \times 100, \sigma_\theta^2 \times 1)$



e) histogram $(\sigma_x^2 \times 1, \sigma_y^2 \times 1, \sigma_\theta^2 \times 10)$　　f) average error $(\sigma_x^2 \times 1, \sigma_y^2 \times 1, \sigma_\theta^2 \times 10)$

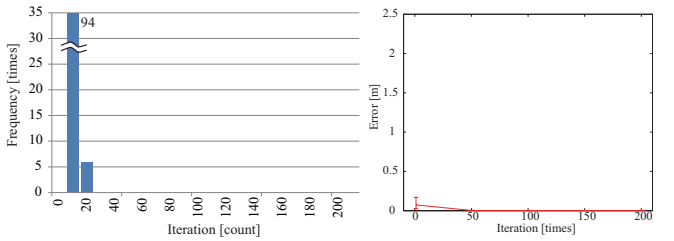


g) histogram $(\sigma_x^2 \times 1, \sigma_y^2 \times 1, \sigma_\theta^2 \times 100)$　　h) average error $(\sigma_x^2 \times 1, \sigma_y^2 \times 1, \sigma_\theta^2 \times 100)$

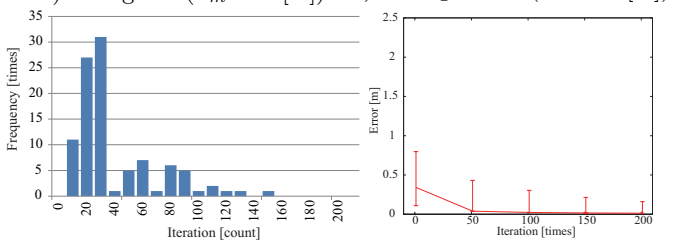
Figure 7: Changes of motion variences

の標準偏差 σ_m が 0.1 [m] になるようにマイクロホンの初期位置を設定した場合の結果であり，図 8c),d) は，σ_m =0.5 [m]，図 8e),f) は，σ_m =1.0 [m] の結果である．

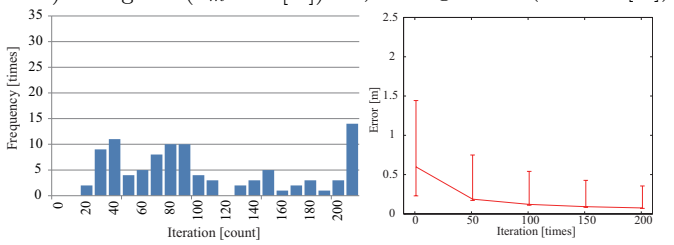
σ_m が 0.1 [m] と，初期位置が正解位置に比較的近い場合は，20 回以下のインパルスでほぼ収束することから，提案手法の正当性が示されている．σ_m が 0.5 [m] の場合であっても，収束までのインパルス回数は増加するものの 30 回程度で，大半が収束していることがわかる．一方，σ_m が 1.0 [m] と大きくなってしまうと，収束にかかるインパルス回数の度数分布はなだらかになり，一概に何回インパ



a) histogram (σ_m=0.1 [m]) b) average error (σ_m=0.1 [m])



c) histogram (σ_m=0.5 [m]) d) average error (σ_m=0.5 [m])



c) histogram (σ_m=1.0 [m]) d) average error (σ_m=1.0 [m])

Figure 8: Performance for initial microphone location errors

ルスを出力する必要があるか明確には言えない．

マイクロホンの位置推定誤差については，インパルス回数が増えるに応じて，収束していくこと，また 50 回程度で推定精度の向上が徐々に頭打ちになることがわかる．また，σ_m が 0.5 [m] 以下の場合は，マイクロホンの位置推定が 50 回程度で正確に行えるのに対して，σ_m が 1.0 [m] になると，なかなか初期位置の誤差が吸収しきれないことがわかる．

以上のことから，マイクロホンアレイの初期位置はキャリブレーションを行う際に重要なパラメータであり，本稿のケースでは，σ_m を 0.5 [m] 以下に設定することが望ましいと言える．

5 おわりに

本稿では，非同期分散マイクロホンアレイのオンラインキャリブレーション問題を解決するために提案している EKF-SLAM ベースの手法の評価を行った．状態遷移モデルに長方形運動モデルと円運動モデルを使用し，数値実験で両者を比較した．また，状態遷移モデルの誤差，観測誤差，そしてマイクロホンの初期位置を変化させ，提案手法のロバスト性，適用範囲の評価を行った．結果として，本手法を利用する際には，本稿のマイクロホンアレイ設定条件では，観測モデル，状態遷移モデルの誤差を実際の誤差の 10 倍程度に抑えるべきであるが，100 倍程度でも場合によっては十分実用に耐えうること，マイクロホンの初期位置は，実際の位置に対して標準偏差が 0.5 [m] 以下であれば，高精度なキャリブレーションが可能であることが示された．今後はカルマンフィルタの理論的解析をし，実環境で評価する必要がある．さらに，非線形な状態遷移モデルに対し，よりロバストな UKF やパーティクルフィルタの適用を試みる予定である．

謝辞

本研究の一部は科研費若手研究 (B)(22700165)，科研費 (S)(19100003)，新学術領域研究 (22118502)，特別研究員奨励費の補助を受けた．

参考文献

[1] J.-M. Valin, J. Rouat, and F. Michaud, “Enhanced robot audition based on microphone array source separation with post-filter,” in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*. IEEE, 2004, pp. 2123–2128.

[2] F. Asano, H. Asoh, and T. Matsui, “Sound source localization and signal separation for office robot “Jijo-2”,” in *Proc. of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI-99)*, 1999, pp. 243–248.

[3] S. Yamamoto, J.-M. Valin, K. Nakadai, T. Ogata, and H. G. Okuno, “Enhanced robot speech recognition based on microphone array source separation and missing feature theory,” in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*. IEEE, 2005, pp. 1489–1494.

[4] H. Saruwatari, Y. Mori, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, and T. Morita, “Two-stage blind source separation based on ICA and binary masking for real-time robot audition system,” in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*. IEEE, 2005, pp. 209–214.

[5] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, “SLAM-based online calibration of asynchronous microphone array for robot audition,” in *Proceedings of 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, IEEE, 2011, pp. 524–529.

[6] S. Thrun, “Affine structure from sound,” *Advances in Neural Information Processing Systems*, vol. 18, pp. 1353–1360, 2006.

[7] N. Ono, H. Kohno, N. Ito, S. Sagayama, “Blind alignment of asynchronously recorded signals for distributed microphone array,” in *2009 IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, IEEE, 2009, pp. 161–164.

[8] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, “Intelligent sound source localization for dynamic environments,” in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, IEEE, 2009, pp. 664–669.

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-B102-5

階層ベイズ推定を用いた有色雑音環境下での音源定位

Sound localization in colored noise using hierarchical Bayesian estimation

浅野太 [1,2]、**麻生英樹** [1]、**中臺一博** [2]
Futoshi Asano[1,2], Hideki Asoh[1] and Kazuhiro Nakadai[2]

産業技術総合研究所 [1], ホンダリサーチインスティチュートジャパン [2]
AIST[1], HRI[2]

Abstract

When sound localization is conducted in an enclosure such as a room, reverberation (sum of multiple reflections) behaves as an additive spatially colored noise, resulting in poor spatial resolution. The main reason for this is that the model used in the estimation is assumed to be spatial white. The authors reported that the joint estimation of the sound source parameters and the noise covariance using a Bayesian framework contributes to the improvement of the spatial resolution. In this report, a method of estimating the common factor of the noise covariance from a different observation data set using hierarchical modeling is discussed. This method is considered to be useful in the case such as a dynamic environment when a number of samples included in a single observation is limited.

1 はじめに

部屋などの閉空間内で音源定位を行う場合、部屋の残響 (壁などに多重反射した音源信号) が空間的に有色な雑音として加わるため、音源位置推定の分解能が低下する場合がある。これは、推定に用いているモデルが、雑音の空間的白色性を仮定しているためである。有色雑音に対しては、一般化固有値分解 (GEVD) を用いて雑音を白色化する方法が提案されているが [1]、これには有色雑音の共分散行列が既知であることが条件となる。残響の場合、残響だけを単独に観測することができないため、雑音の事前白色化は困難である。著者らは、ベイズ推定の枠組みを用いて、音源位置などのパラメタに加え、雑音の共分散行列を同時推定することにより、空間分解能が改善することを示した [2]。本報告では、共分散行列の推定を階層化 [3] することにより、動的環境のように一度に得られる観測サンプルが少ない場合でも、共分散行列を安定して推定する手法について考える。音源やセンサが閉空間内を移動したとしても、環境 (部屋) が同じであれば、部屋に特有の共振周波数は変化せず、このため、部屋内の 2 点間の伝達関数を共通の極零モデルを用いて表すことができることが報告されている [4]。本報告では、この知見に基づき、雑音の共分散行列を、音源やセンサ位置などが異なる複数のデータセットに共通する項と、個々のデータ (特定の音源-センサ配置) に特有の項に分解できるものと仮定する。このうち、データセットに共通する項を階層モデルを用いて推定する。

2 音源定位の問題

2.1 信号と雑音のモデル

観測値は、マイクロホン入力の短区間フーリエ変換 (STFT) により、$\boldsymbol{z}_{j,k} = [Z_1(\omega, j, k), \cdots, Z_M(\omega, j, k)]^T$ のように構成される。ここで、$Z_m(\omega, j, k)$ は、m 番目のマイクロホン入力の STFT である。j および k はブロックおよびフレームのインデックスを表す。フレームは、STFT を行う単位である。K 個の連続するフレームを $\boldsymbol{Z}_j = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_K]$ のようにまとめたものを、ここではブロックと呼ぶ。ブロック内では、音源位置 $\boldsymbol{\theta}_j = [\theta_{j,1}, \cdots, \theta_{j,N}]^T$ は定常と見なせるものと仮定する。観測値は、次式のようにモデル化されるものとする。

$$\boldsymbol{z}_{j,k} = \boldsymbol{A}_j(\boldsymbol{\theta}_j)\boldsymbol{s}_{j,k} + \boldsymbol{v}_{j,k} \tag{1}$$

ここで、$\boldsymbol{A}_j(\boldsymbol{\theta}_j) = [\boldsymbol{a}_j(\theta_1), \cdots, \boldsymbol{a}_j(\theta_N)]$ はアレイ・マニフォールド・ベクトル $\boldsymbol{a}_j(\theta_{j,n})$ を列ベクトルに持つマニフォールド行列、$\boldsymbol{s}_{j,k}$ は音源信号、$\boldsymbol{v}_{j,k}$ は雑音を表す。本報告では、$\boldsymbol{v}_{j,k}$ が部屋の残響である場合を考える。$\boldsymbol{v}_{j,k}$ は、以前のフレームにおける信号源 $\boldsymbol{s}_{j,l}(l < k)$ のレプリカであるが、遅延時間がある程度大きい場合は、$\boldsymbol{s}_{j,k}$ と $\boldsymbol{v}_{j,k}$ は無相関と考えてよい。この場合、共分散行列は次式のようにモデル化される。

$$\boldsymbol{R}_j = E[\boldsymbol{z}_{j,k}\boldsymbol{z}_{j,k}^H] = \boldsymbol{A}_j\boldsymbol{\Gamma}_j\boldsymbol{A}_j^H + \boldsymbol{K}_j \tag{2}$$

ここで、$\boldsymbol{\Gamma}_j = E[\boldsymbol{s}_{j,k}\boldsymbol{s}_{j,k}^H]$ および $\boldsymbol{K}_j = E[\boldsymbol{v}_{j,k}\boldsymbol{v}_{j,k}^H]$ は、音源および雑音の共分散行列である。雑音が白色の場合は、

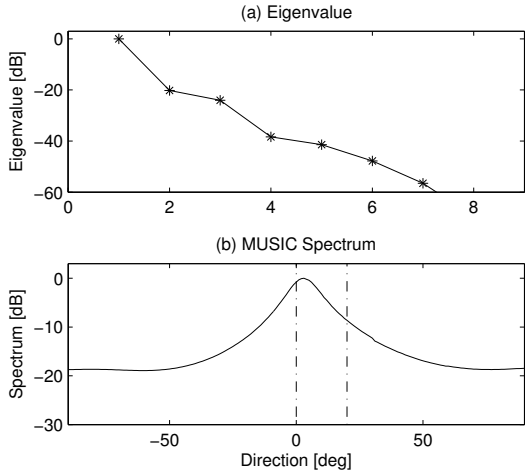


Figure 1: SEVD-MUSIC spatial spectrum. It is assumed that the noise $\boldsymbol{v}_k$ is spatially white with the covariance matrix $\boldsymbol{I}$.

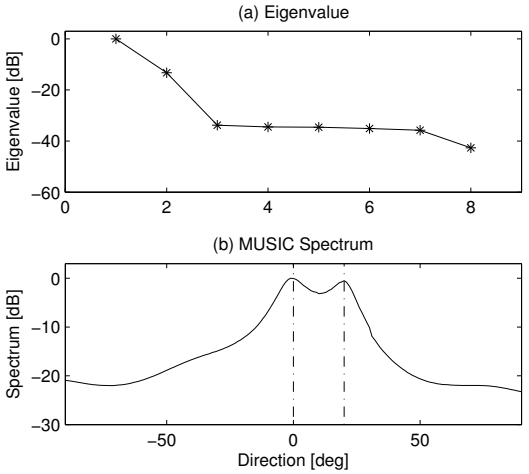


Figure 2: GEVD-MUSIC spatial spectrum. It is assumed that the noise $\boldsymbol{v}_k$ is spatially colored and its covariance matrix $\boldsymbol{K}$ is known.

雑音の共分散行列は、対角行列 $\boldsymbol{K}_j = {}_j^2\boldsymbol{I}$ となり、問題は大幅に簡略化される。本報告では、雑音の共分散行列が未知の非対角行列の場合を考える。

2.2 雑音の共分散行列推定の効果

ここでは、雑音の共分散に含まれる情報の効果を、空間スペクトル推定の一手法である MUSIC 法を用いて見ていく。図 1 は、残響時間 0.5 秒程度の会議室で測定したインパルス応答に、白色雑音を畳み込んで生成した観測信号に対して、共分散行列の標準固有値分解 (SEVD) を行い、MUSIC 空間スペクトルを算出したものである。音源方向は図 (b) 中の点線で示すように、$(0°, 20°)$ である。分析パラメタは、表 1 に示してある。インパルス応答測定に用いたマイクロホンアレイは、ロボット (HRP-2) に搭載された、8 素子のものである。同図 (a) から、音源数が $N = 2$ であるにも関わらず、音源数の指標となる大きな固有値は 1 つだけであり、対応した MUSIC 空間スペクトル (同図 (b)) にもピークは 1 つしか現れていない。この原因は、SEVD を用いた MUSIC 法では雑音の空間的白色性を仮定しており、実際の残響の空間的有色性との不整合によるものと考えられる。

Table 1: Parameters for analysis.

Parameter	Value
Sampling frequency	16 kHz
Number of microphones	8
Frame length(STFT length)	512 points
Frame shift	128 points
Block length（observation time)	32000 points
Frequency	1500 Hz

図 2 は、雑音 (残響) の共分散行列を既知として、共分散行列のペア $(\boldsymbol{R}, \boldsymbol{K})$ に対して、一般化固有値問題 (GEVD) を解き、これにより MUSIC スペクトルを求めたものである。GEVD に拡張することにより、雑音を空間的に白色化することができる。雑音が残響の場合、実際の応用では $\boldsymbol{K}$ は直接観測できないが、ここでは、ちょっとずるをして、インパルス応答から反射・残響の部分だけを切り出し、これにソース信号を畳み込んで観測値を生成して、$\boldsymbol{K}$ を求めた。この図から、固有値分布では、音源数 $N = 2$ に対応した大きな固有値が現れ、それ以外は、平坦な分布となっている。これは、GEVD による雑音の白色化の効果である。また、空間スペクトルも音源位置に対応した 2 つのピークが現れている。このことから、雑音の共分散行列 $\boldsymbol{K}$ の情報を利用することにより、有色雑音下での空間分解能の向上が期待される。

2.3 観測値の尤度と最尤法

ここでは、観測値に対する尤度を導入する。また、後述のベイズ推定の理解を助けるため、最尤法についても簡単に述べておく。

式 (1) において、$\boldsymbol{v}_{j,k}$ が多次元複素ガウス分布に従うものとし、また、観測値 $\boldsymbol{z}_{j,k}$ は互いに独立であると仮定すると、尤度関数は次式のようになる。

$$
\begin{aligned}
p(\boldsymbol{Z}_j|\boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j) &\propto |\boldsymbol{K}_j|^{-K} \\
&\exp\left(-\sum_{k=1}^{K} [\boldsymbol{z}_{j,k} - \boldsymbol{A}_j \boldsymbol{s}_{j,k}]^H \boldsymbol{K}_j^{-1} [\boldsymbol{z}_{j,k} - \boldsymbol{A}_j \boldsymbol{s}_{j,k}]\right) \\
&= |\boldsymbol{K}_j|^{-K} \exp\left\{-\mathrm{tr}\ \boldsymbol{C}_j \boldsymbol{K}_j^{-1}\right\} \qquad (3)
\end{aligned}
$$

ここで、

$$
\boldsymbol{C}_j = \sum_{k=1}^{K} [\boldsymbol{z}_{j,k} - \boldsymbol{A}_j \boldsymbol{s}_{j,k}] [\boldsymbol{z}_{j,k} - \boldsymbol{A}_j \boldsymbol{s}_{j,k}]^H \qquad (4)
$$

$\boldsymbol{S}_j = [\boldsymbol{s}_{j,1}, \cdots, \boldsymbol{s}_{j,K}]$ はブロック内の音源信号を表す。

本節では、簡単のため、$\boldsymbol{K}_j$ を既知の固定値と考え、式 (3) を $\boldsymbol{s}^*_{j,k}$ について偏微分して $\mathbf{0}$ とおくと、

$$\boldsymbol{A}_j^H \boldsymbol{K}_j{}^{-1} [\boldsymbol{z}_{j,k} - \boldsymbol{A}_j \boldsymbol{s}_{j,k}] = \mathbf{0} \tag{5}$$

これから、$\boldsymbol{s}_{j,k}$ の最尤推定値は次式で与えられる。

$$\hat{\boldsymbol{s}}_{j,k} = \left[\boldsymbol{A}_j^H \boldsymbol{K}_j{}^{-1} \boldsymbol{A}_j\right]^{-1} \boldsymbol{A}_j^H \boldsymbol{K}_j{}^{-1} \boldsymbol{z}_{j,k} \tag{6}$$

式 (6) を式 (3) に代入して対数をとり、$\boldsymbol{\theta}$ に無関係な項を取り除いて $\boldsymbol{\theta}$ についての対数尤度を求めると、次式のようになる。

$$\begin{aligned} LL(\boldsymbol{\theta}) &\propto -\mathrm{tr}\ \hat{\boldsymbol{C}}_j \boldsymbol{K}_j{}^{-1} \\ &= -\mathrm{tr}\ \boldsymbol{G}(\boldsymbol{\theta}) \boldsymbol{C}_{z,j} \boldsymbol{G}^H(\boldsymbol{\theta}) \boldsymbol{K}_j{}^{-1} \end{aligned} \tag{7}$$

ここで、

$$\hat{\boldsymbol{C}}_j = \sum_{k=1}^{K} [\boldsymbol{z}_{j,k} - \boldsymbol{A}_j(\boldsymbol{\theta}) \hat{\boldsymbol{s}}_{j,k}] [\boldsymbol{z}_{j,k} - \boldsymbol{A}_j(\boldsymbol{\theta}) \hat{\boldsymbol{s}}_{j,k}]^H \tag{8}$$

また、$\boldsymbol{G}(\boldsymbol{\theta})$ は、次式で定義されるように、音源方向を $\boldsymbol{\theta}$ と仮定したときに、残差 $\boldsymbol{z}_{j,k} - \boldsymbol{A}_j(\boldsymbol{\theta}) \boldsymbol{s}_{j,k}$ を与えるフィルタである。

$$\boldsymbol{G}(\boldsymbol{\theta}) = \boldsymbol{I} - \boldsymbol{A}_j(\boldsymbol{\theta}) \left[\boldsymbol{A}_j^H(\boldsymbol{\theta}) \boldsymbol{K}_j{}^{-1} \boldsymbol{A}_j(\boldsymbol{\theta})\right]^{-1} \boldsymbol{A}_j^H(\boldsymbol{\theta}) \boldsymbol{K}_j{}^{-1} \tag{9}$$

$\boldsymbol{C}_{z,j}$ は、次式で定義される観測値 $\boldsymbol{z}_{j,k}$ のサンプル共分散行列である。

$$\boldsymbol{C}_{z,j} = \sum_{k=1}^{K} \boldsymbol{z}_{j,k} \boldsymbol{z}_{j,k}^H \tag{10}$$

$\boldsymbol{\theta}_j$ の最尤推定値は、次式のようになる。

$$\hat{\boldsymbol{\theta}}_j = \arg\max_{\boldsymbol{\theta}_j} LL(\boldsymbol{\theta}_j) \tag{11}$$

3　パラメタの同時推定

複数のパラメタを同時推定する手法としては、変分ベイズ法 [5] やギブスサンプリング [3] などが考えられる。本報告では、観測値が音源方向 $\boldsymbol{\theta}$ の非線形関数となっていることから、ギブスサンプリングとメトロポリス・アルゴリズムを組み合わせて用いる [3, 6, 2]。3.1 節〜3.3 節では、サンプルを得るための条件付き分布について述べる。3.4 節では、これらの条件付き分布を用いて、パラメタのサンプルを反復して求める手続きについて述べる。

3.1　$\boldsymbol{s}_{j,k}$ の条件付き分布

ベイズの定理により次式が成り立つ。

$$p(\boldsymbol{s}_{j,k} | \boldsymbol{Z}_j, \boldsymbol{\theta}_j, \tilde{\boldsymbol{S}}_j, \boldsymbol{K}_j) \propto p(\boldsymbol{s}_{j,k}) p(\boldsymbol{Z}_j | \boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j) \tag{12}$$

ここで、

$$\tilde{\boldsymbol{S}}_j = [\boldsymbol{s}_{j,1}, \cdots, \boldsymbol{s}_{j,k\ \ 1}, \boldsymbol{s}_{j,k+1}, \cdots, \boldsymbol{s}_{j,K}] \tag{13}$$

$p(\boldsymbol{s}_{j,k})$ がガウス分布 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}_\mathbf{0})$ であると仮定すると、$\boldsymbol{s}_{j,k}$ の条件付き分布は次式のようになる。

$$p(\boldsymbol{s}_k | \boldsymbol{Z}, \boldsymbol{\theta}, \tilde{\boldsymbol{S}}, \boldsymbol{K}) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Phi}) \tag{14}$$

$$\boldsymbol{\Phi}_j = \boldsymbol{A}_j^H \boldsymbol{K}_j{}^{-1} \boldsymbol{A}_j + \boldsymbol{\Phi}_0{}^{-1}{}^{-1} \tag{15}$$

$$\boldsymbol{\mu}_{j,k} = \boldsymbol{\Phi}_j \boldsymbol{A}_j^H \boldsymbol{K}_j{}^{-1} \boldsymbol{z}_{j,k} \tag{16}$$

式 (6) と式 (14) を比較すると、条件付き平均値 $\boldsymbol{\mu}_k$ は、最尤推定値 $\hat{\boldsymbol{s}}_k$ に事前分布に関する項 $\boldsymbol{\Phi}_0{}^{-1}$ を加えたものとなっていることがわかる。

3.2　$\boldsymbol{K}_j$ の条件付き分布

共分散行列 $\boldsymbol{K}_j$ は、次式の inverse-Wishart 分布に従うものと仮定する。

$$\begin{aligned} p(\boldsymbol{K}_j) &= \text{inv-Wishart}\ \ \boldsymbol{K}_j; \nu_0, (\nu_0 \boldsymbol{K}_0)^{-1}{}^{-1} \\ &\propto |\boldsymbol{K}_j|^{-(\nu_0+M)} \exp\left\{-\mathrm{tr}(\nu_0 \boldsymbol{K}_0 \boldsymbol{K}_j{}^{-1})\right\} \end{aligned} \tag{17}$$

これから、$\boldsymbol{K}_j$ の条件付き分布も、次式に示すような inverse-Wishart 分布となる。

$$\begin{aligned} & p(\boldsymbol{K}_j | \boldsymbol{Z}_j, \boldsymbol{S}_j, \boldsymbol{\theta}_j) \\ &\propto p(\boldsymbol{K}_j) p(\boldsymbol{Z}_j | \boldsymbol{\theta}_j, \boldsymbol{S}_j, \boldsymbol{K}_j) \\ &\propto |\boldsymbol{K}_j|^{-(\nu_0+M)} \exp\left\{-\mathrm{tr}(\nu_0 \boldsymbol{K}_0 \boldsymbol{K}_j{}^{-1})\right\} \\ &\quad\ |\boldsymbol{K}_j|^{-K} \exp\left\{-\mathrm{tr}(\boldsymbol{C}_j \boldsymbol{K}_j{}^{-1})\right\} \\ &= |\boldsymbol{K}_j|^{-(\nu_0+M+K)} \exp\left\{-\mathrm{tr}([\nu_0 \boldsymbol{K}_0 + \boldsymbol{C}_j] \boldsymbol{K}_j{}^{-1})\right\} \\ &\propto \text{inv-Wishart}\ \ \nu_0 + K, [\nu_0 \boldsymbol{K}_0 + \boldsymbol{C}_j]^{-1} \end{aligned} \tag{18}$$

3.3　$\boldsymbol{\theta}_j$ の条件付き分布

$\boldsymbol{A}_j(\boldsymbol{\theta}_j)$ が $\boldsymbol{\theta}_j$ の非線形関数であることから、$\boldsymbol{\theta}_j$ のサンプルをその条件付き分布 $p(\boldsymbol{\theta}_j | \boldsymbol{Z}_j, \boldsymbol{S}_j, \boldsymbol{K}_j)$ から直接得るのは一般に困難である。このような場合、メトロポリス・アルゴリズム [3, 6] により、$\boldsymbol{\theta}_j$ のサンプルを得る手法が一般的である。メトロポリス・アルゴリズムでは、前回 (p) の反復によりサンプル $\boldsymbol{\theta}_j^{(p)}$ が得られているものとし、提案分布 $J(\boldsymbol{\theta}_j^* | \boldsymbol{\theta}_j^{(p)})$ から新たなサンプル $\boldsymbol{\theta}_j^*$ を得る。本報告では、次式の一様分布を提案分布として用いる。

$$J(\boldsymbol{\theta}_j^* | \boldsymbol{\theta}_j^{(p)}) = \mathcal{U}(\boldsymbol{\theta}_j^{(p)} - \ \ , \boldsymbol{\theta}_j^{(p)} + \ \ ) \tag{19}$$

ここで、 は、適当な定数ベクトルである。新たなサンプルは、次式の採択率が適当な閾値 r_{thr} を超えた場合に採用する。

$$r = \frac{p(\boldsymbol{Z}_j | \boldsymbol{\theta}_j^*, \boldsymbol{S}_j^{(p+1)}, \boldsymbol{K}_j^{(p+1)})}{p(\boldsymbol{Z}_j | \boldsymbol{\theta}_j^{(p)}, \boldsymbol{S}_j^{(p+1)}, \boldsymbol{K}_j^{(p+1)})} \frac{p(\boldsymbol{\theta}_j^*)}{p(\boldsymbol{\theta}_j^{(p)})} \tag{20}$$

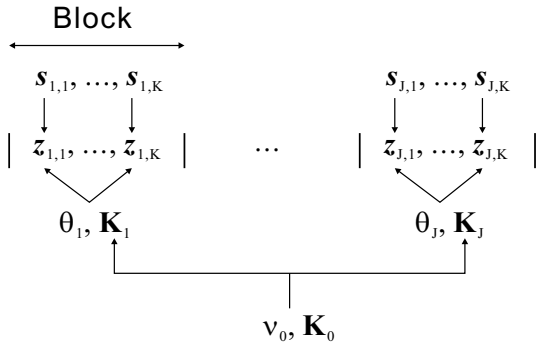


Figure 3: Hierarchical model of the covariance matrix.

3.4 ギブスサンプリング

ここでは、前節までで述べた各パラメタの条件付き分布から反復的にサンプルを得るギブスサンプリングについて述べる。ギブスサンプリングの手続きを以下にまとめる。

1. 初期値 $\boldsymbol{K}_j^{(1)}$ および $\boldsymbol{\theta}_j^{(1)}$ を設定する。
2. $\boldsymbol{s}_{j,k}$ のサンプルを得る。

$$\boldsymbol{s}_{j,k}^{(p+1)} \sim p(\boldsymbol{s}_{j,k}|\boldsymbol{Z}_j, \boldsymbol{\theta}_j^{(p)}, \tilde{\boldsymbol{S}}_{jk}^{(p)}, \boldsymbol{K}_j^{(p)}) \quad \forall k$$

3. $\boldsymbol{K}_j$ のサンプルを得る。

$$\boldsymbol{K}_j^{(p+1)} \sim p(\boldsymbol{K}_j|\boldsymbol{Z}_j, \boldsymbol{S}_j^{(p+1)}, \boldsymbol{\theta}_j^{(p)})$$

4. $\boldsymbol{\theta}_j$ のサンプルを得る。

$$\begin{aligned} \boldsymbol{\theta}_j^* &\sim J(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_j^{(p)}) \\ \boldsymbol{\theta}_j^{(p+1)} &= \begin{cases} \boldsymbol{\theta}_j^* & r > r_{thr} \\ \boldsymbol{\theta}_j^{(p)} & \text{otherwise} \end{cases} \end{aligned}$$

5. $p \leftarrow p+1$ として、ステップ 2 に戻る。

4 階層ベイス推定

4.1 共分散行列の階層モデル

冒頭で述べたように、共分散行列は、J 個の観測値 $\{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_J\}$ に共通する項と個々の観測値に特有の項に分解されるものと仮定する。このうち、共通する項を階層モデルを用いて推定する。共分散行列は、次式のサンプリングモデルに従うと仮定する。

$$\boldsymbol{K}_1, \cdots, \boldsymbol{K}_J \sim \text{ i.i.d. inv-Wishart}(\nu_0, (\nu_0 \boldsymbol{K}_0)^{-1}) \quad (21)$$

図 3 は、このモデルを図示したものである。式 (21) において、$\boldsymbol{K}_0$ が観測値に共通する項に相当する。ν_0 は、$\boldsymbol{K}_0$ に対する仮想的なサンプル数であり、式 (18) からわかるように、各データに特有の項に対する共通項の重みの役割を果たす。

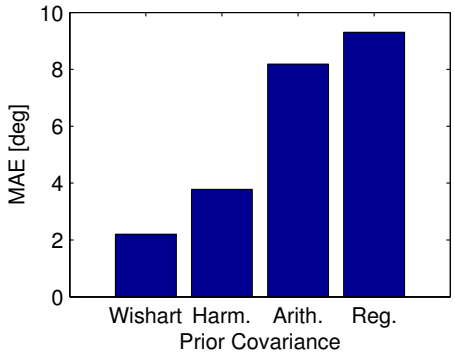


Figure 4: MAE for different estimation method for $\boldsymbol{K}_0$.

4.2 $\boldsymbol{K}_0$ の条件付き分布

$p(\boldsymbol{K}_0) = \text{Wishart}(\ , \boldsymbol{\Psi})$ であると仮定すると (inverse-Wishart ではないことに注意)、$\boldsymbol{K}_0$ の条件付き分布は次式のようになる。

$$\begin{aligned} & p(\boldsymbol{K}_0|\boldsymbol{K}_1, \cdots, \boldsymbol{K}_J, \nu_0) \\ & \propto p(\boldsymbol{K}_1, \cdots, \boldsymbol{K}_J|\boldsymbol{K}_0, \nu_0) p(\boldsymbol{K}_0, \nu_0) \\ & = p(\boldsymbol{K}_0) \prod_{i=1}^{J} p(\boldsymbol{K}_j|\boldsymbol{K}_0, \nu_0) \\ & \propto |\boldsymbol{\Psi}|^{\ \eta} |\boldsymbol{K}_0|^{\eta\ M} \exp\left\{-\text{tr}(\boldsymbol{K}_0 \boldsymbol{\Psi}^{\ -1})\right\} \\ & \quad \prod_{j=1}^{J} |\boldsymbol{K}_0^{\ -1}|^{\ \nu_0} |\boldsymbol{K}_j|^{\ (\nu_0+M)} \exp\left\{-\text{tr}(\nu_0 \boldsymbol{K}_0 \boldsymbol{K}_j^{\ -1})\right\} \\ & \propto |\boldsymbol{K}_0|^{\eta+J\nu_0\ \ M} \exp\left\{-\text{tr}\ \ \boldsymbol{K}_0 \boldsymbol{\Lambda}^{\ -1}\ \right\} \\ & = \text{Wishart}\,(\boldsymbol{K}_0;\ \ + J\nu_0, \boldsymbol{\Lambda}) \end{aligned} \quad (22)$$

ここで、

$$\boldsymbol{\Lambda} := \left(\boldsymbol{\Psi}^{\ -1} + \nu_0 \sum_{j=1}^{J} \boldsymbol{K}_j^{\ -1} \right)^{\ -1} \quad (23)$$

$\boldsymbol{\Psi}$ の影響が小さい場合、式 (23) は、$\{\boldsymbol{K}_j\}$ の調和平均を求めるような操作となり、この値（の定数倍）が式 (22) に示す Wishart 分布から得られるサンプルの平均値となる。

4.3 ν_0 の条件付き分布

ν_0 の事前分布に次式を仮定すると、

$$p(\nu_0) \propto \exp(-\alpha\nu_0) \quad (24)$$

条件付き分布は次式のようになる。

$$\begin{aligned} & p(\nu_0|\boldsymbol{K}_0, \boldsymbol{K}_1, \cdots, \boldsymbol{K}_J) \\ & \propto p(\nu_0) p(\boldsymbol{K}_1, \cdots, \boldsymbol{K}_J|\nu_0, \boldsymbol{K}_0) \\ & \propto \exp(-\alpha\nu_0) \prod_{j=1}^{J} \frac{|\nu_0 \boldsymbol{K}_0|^{\nu_0}}{\Gamma_M(\nu_0)} |\boldsymbol{K}_j|^{\ (\nu_0+M)} \\ & \quad \exp\left\{-\text{tr}(\nu_0 \boldsymbol{K}_0 \boldsymbol{K}_j^{\ -1})\right\} \end{aligned} \quad (25)$$

ここで、$\Gamma_M(\nu_0) = \ ^{M(M\ \ 1)/2} \prod_{m=1}^{M} \Gamma(\nu_0 - m + 1)$ である。$\Gamma(\cdot)$ はガンマ関数を表す。

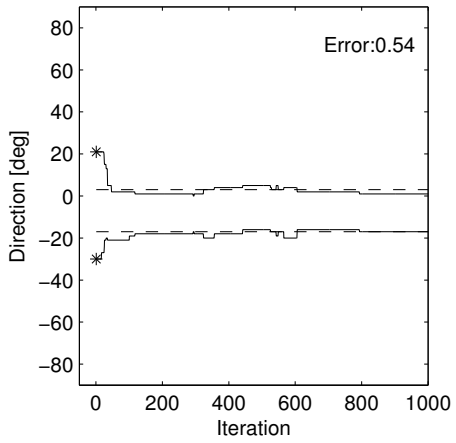


Figure 5: Variation of sample $\boldsymbol{\theta}^{(p)}$ in the Gibbs sampling.

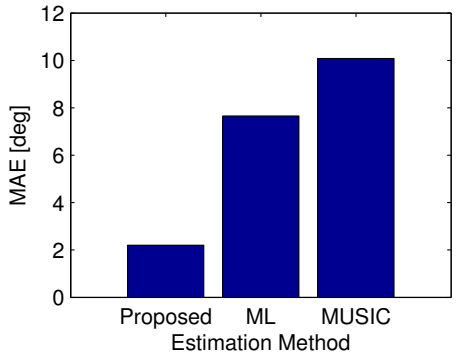


Figure 6: MAE for different parameter estimation method. H-Bayes:hierarchical Bayesian estimator、ML:maximum likelihood estimator、MUSIC:MUSIC estimator。

4.4 反復アルゴリズム

ここでは、$\boldsymbol{K}_0$ および ν_0 のサンプルを得る手続きをまとめる。

1. 初期値 $\boldsymbol{K}_0^{(1)}$ および $\nu_0^{(1)}$ を設定する。
2. 3.4 節で述べたアルゴリズムにより、$\{\boldsymbol{K}_1^{(p+1)}, \cdots, \boldsymbol{K}_J^{(p+1)}\}$ を得る。
3. $\boldsymbol{K}_0$ のサンプルを得る。
$$\boldsymbol{K}_0^{(p+1)} \sim p(\boldsymbol{K}_0|\boldsymbol{K}_1^{(p+1)}, \cdots, \boldsymbol{K}_J^{(p+1)}, \nu_0^{(p)})$$
4. ν_0 のサンプルを得る。
$$\nu_0^{(p+1)} \sim p(\nu_0|\boldsymbol{K}_0^{(p+1)}, \boldsymbol{K}_1^{(p+1)}, \cdots, \boldsymbol{K}_J^{(p+1)})$$
5. $p \leftarrow p+1$ として、ステップ 2 に戻る。

5 評価実験

5.1 実験 I - 静止音源

実験 1 では、静止音源の環境のシミュレーションとして、20 通りの音源配置 ($J = 20$) に対するインパルス応答をガウス雑音に畳み込み、観測値を生成した。インパルス応答は、2.2 節で用いたもので同様である。音源間隔を 20° とした音源のペアを用い、音源ペアの方向を乱数を用いて 20 通りに変化させた。分析パラメタは、ブロック長を 3200 点 (0.2 秒) とした以外は、表 1 と同じである。反復アルゴリズムの反復回数は 1000 回とした。パラメタ $\boldsymbol{\theta}$ の初期値としては、最尤法に白色雑音を加えたものを用いた。

図 5 は、反復におけるパラメタ θ の変化の例を表している。この図から、この例では、比較的少ない反復で、推定値が真値に収束しているのがわかる。

図 6 は、階層ベイズ推定を用いた音源定位を、最尤法、MUSIC 法と比較したものである。最尤法および MUSIC 法では、雑音を空間的に白色と仮定してる。図の縦軸は平均絶対誤差 MAE $= 1/N_{avg} \sum |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|$ であり、20 種の配置 ($J = 20$) と 30 回のトライアル ($N_{trial} = 30$) について平均した ($N_{avg} = J \quad N_{trial}$)。この図から、階層ベイズ推定を用いて雑音の共分散行列を推定した場合の方が、他の 2 つの推定法よりも誤差が小さいのがわかる。

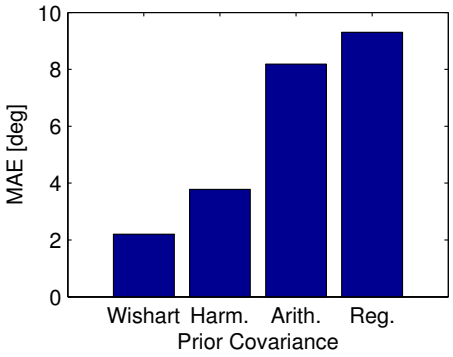


Figure 7: MAE for different estimation method for $\boldsymbol{K}_0$.

図 7 は、$\boldsymbol{K}_0$ のサンプルを得る方法を変えた場合の影響である。同図の“Wishart”は、4.2 節で述べた Wishart 分布からサンプルを得た場合、“Harm.”は式 (23) における $\boldsymbol{\Lambda}$ を用いた場合、“Arith.”は算術平均 $\sum_j \boldsymbol{K}_j$ を用いた場合、“Reg.”は $\boldsymbol{I}$ を用いた場合である。“Reg.”の場合は、共分散行列の階層推定は行われず、$\boldsymbol{K}_0 = \boldsymbol{I}$ の効果は、$\boldsymbol{K}_j$ の正則化となる。この図から、Wishart 分布からサンプルを得た場合、およびその平均値である $\boldsymbol{\Lambda}$ を用いた場合は、MAE が小さいことがわかる。このことから、$\boldsymbol{K}_0$ を $\{\boldsymbol{K}_j\}$ から階層推定することの本質は、式 (23) における調和平均のような操作であることがわかる。

図 8 は ν_0 を単一の値に固定して、MAE を算出したものである。この図から、おおむね $\nu_0 = 10^2$ 付近で MAE は最小値をとる。

図 9(b) は、階層ベイズ推定により求めた $\boldsymbol{K}_0$ の推定精度を評価するため、推定された $\boldsymbol{K}_0$ を用いて一般化固有値問題 $(\boldsymbol{C}_{z,j}, \boldsymbol{K}_0)$ を解き、MUSIC スペクトルを求めたものである。比較のために示した、標準固有値分解（雑音の白色性を仮定）の場合 (同図 (a)) と比較すると、空間分解能が向上しているのがわかる。

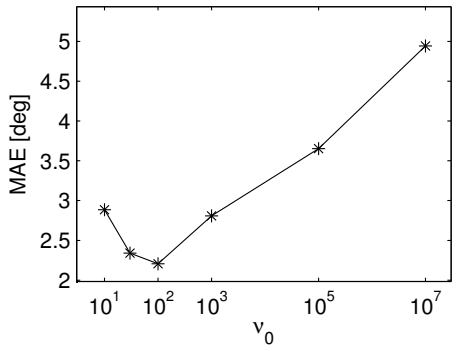


Figure 8: MAE for different ν_0 values.

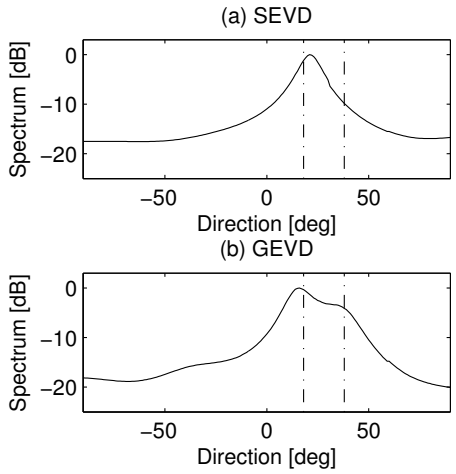


Figure 9: Evaluation of the estimate of $\boldsymbol{K}_0$ using MUSIC spatial spectral estimator.

5.2 実験 II - 移動音源

実験 II では、静止したロボットの周りを 2 名の話者が歩きながら発話した音声を収録し、観測値として用いた。用いたマイクロホンアレイは、ロボット (Honda Hearbo) に搭載した 8 素子のものである。2 名の話者は、約 30° の間隔を保ちながら、ロボットの周囲の半径 1.5m の円周上を等速運動した。分析条件は実験 I と同様である。観測データセットは、連続した 20 ブロックである。ギブスサンプリングの初期値には MUSIC 法の推定値を用いた。実験 I では、周波数を 1500Hz の単一周波数としたが、実験 II では、信号源が音声信号であり、周波数領域でのスパース性のため、周波数により結果が異なる。そこで、800Hz から 3000Hz までの 71 離散周波数について推定を行った。したがって、音源方向の推定値は、20 ブロック　71 周波数 =1420 の観測データごとに評価を行った。

図 10 は推定誤差をまとめたものである。上述の音声のスパース性のため、1420 の観測データセットについて MAE を計算すると、実効音源数が 0 や 1 の周波数とブロックの組み合わせの場合に、大きな誤差となり、手法の比較が難しい。そこで、ここでは、誤差が ±8° 以内に入るデータの全データに対する割合を評価値として用いた。実験 II では、利用できるアレイマニフォールドベクトルが 5° ごと

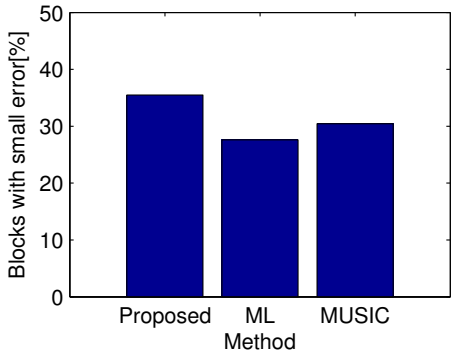


Figure 10: Percentage of blocks with small error .

である。このため、±8° の誤差は、マニフォールドベクトルに換算して概ね 0 ±1 ユニットの誤差に相当する。同図から、階層ベイス推定を行った場合は、ML 法、MUSIC 法に比べ、数%程度推定精度が改善しているのがわかる。

6 結論

本報告では、雑音の共分散行列を、小数のデータからでも安定的に推定する手法として、階層ベイズ推定を用いた手法を検討した。この結果、推定した雑音の共分散行列を用いて音源定位を行うことにより、空間分解能が向上することが示された。今後の課題としては、音源数 N も含めた同時推定の手法の開発が望まれる。

References

[1] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.

[2] F. Asano and H.Asoh, "Joint estimation of sound source location and noise covariance in spatially colored noise," in *Proc. Eusipco 2011*, 2011.

[3] P. D. Hoff, *A first course in Bayesian statistical methods*, Springer, 2009.

[4] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and zero modeling of head-related transfer function," *IEEE Trans. on Speech and Audio Processing*, vol. 7, no. 2, pp. 188–196, 1999.

[5] C. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[6] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisysinusoids via reversible jump mcmc," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2667–2676, 1999.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B102-6

音源定位手法 MUSIC のベイズ拡張

Bayesian Extension of MUSIC for Sound Source Localization

大塚 琢馬 †, 中臺 一博 ‡, 尾形 哲也 †, 奥乃 博 †

Takuma Otsuka†, Kazuhiro Nakadai‡, Tetsuya Ogata†, Hiroshi G. Okuno†

† 京都大学大学院情報学研究科, ‡(株) ホンダ・リサーチ・インスティチュート・ジャパン

†**Graduate school of Informatics, Kyoto University, ‡HONDA Research Institute Japan, Co., Ltd.**

†{ohtsuka, ogata, okuno}@kyoto-u.ac.jp, ‡nakadai@jp.honda-ri.com

Abstract

This paper presents a Bayesian extension of MUSIC-based sound source localization (SSL) method. SSL is important for the separation of simultaneous speech signals as well as for auditory scene analysis by mobile robots. One of the drawbacks of existing SSL methods is the necessity of careful parameter tunings, e.g., the sound source detection threshold depending on the reverberation time and the number of sources. Our contribution consists of (1) automatic parameter estimation in the variational Bayesian framework and (2) tracking of sound sources with reliability. Experimental results demonstrate our method robustly tracks multiple sound sources in a reverberant environment with RT20 = 840 (ms).

1 はじめに

音響情報は人間の知覚の重要な位置を占める. 例えば, 人は足音を聞くことで目に頼ることなく誰かが近づいてきている, あるいは遠ざかっているといった状況を理解することができる. ロボットや計算機による周囲の音響情報の理解, つまり, 「音環境理解」の実現は, 聴覚障碍者の補助や, 人間の音に対する気づきを向上することができると期待される [Kubota et al., 2008].

音源定位はマイクロフォンアレイを用いた同時発話混合音声の分離 [Nakadai et al., 2010], 遠隔ロボットのオペレータへの音源方向提示 [Mizumoto et al., 2011], 移動ロボットによる音源検出と位置推定 [Sasaki et al., 2010]など, 音環境理解にとって重要な要素技術である. 図 1 に示すような, 複数音源, ロボットの移動, 音源移動など, 動的に音環境が変化する状況においても, 手間のかかるパラメータ設定をしなくてもロボットが頑健に各音源を定位, 追跡することが望まれる.

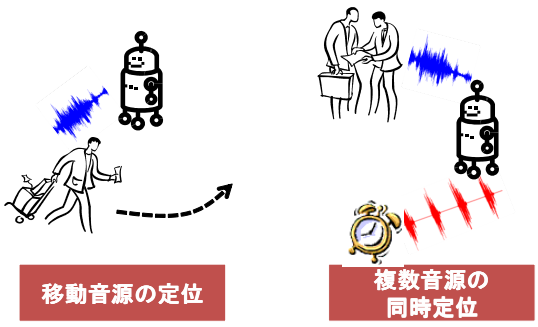


Figure 1: 動的環境下での音源定位

マイクロフォンアレイを用いた音源定位法はビームフォーミングに基づく手法 [Doclo et al., 2001] と, MUltiple SIgnal Classification (MUSIC) に基づく手法 [Schmidt, 1986; Asano et al., 2001; Danès et al., 2010] がロボットによく応用される. 我々は次の理由より, MUSIC 法を利用する. (1) MUSIC の方が雑音に頑健である, (2) 音源数がマイクロフォン数未満という条件下では, 比較的安定して複数音源の定位が可能である.

通常の MUSIC 法では, 音源が到来しているかどうかを MUSIC スペクトルと呼ばれる音源到来評価関数に対して閾値を設定して判定する. 多くの場合, 適切な閾値は環境中の音源数や残響時間などに依存するため, 状況に応じて最適な閾値の設定が重要である. MUSIC 法を用いた場合の環境中の音源数推定問題は, 赤池情報量規準の利用 [Danès et al., 2010] や, サポートベクターマシンの適用 [Yamamoto et al., 2006] によってこれまで取り組まれてきた. しかし, これらの手法で音源数が推定できたとしても, 適切な音源検出閾値の設定問題は依然として残っている. この問題に対する典型的な対策としては, マイクロフォンアレイを設

置した環境で録音した音響信号から計算した MUSIC スペクトルを見ながら手動で閾値を設定するという方法であった.

本稿では, MUSIC 法による音源定位のベイズ拡張を行い, 従来法で必要とされた閾値に相当する情報を自動的に学習することを試みる. これにより, 閾値設定の手間を省くと共に, 試行錯誤により設定された閾値の精度と同等以上の定位精度を実現する. 本手法は次の 2 つのステップから成る. (1) マイクロフォンアレイが置かれた環境で録音した数十秒程度の音響信号から, 音源存在閾値に相当するパラメータを学習する. 学習には変分ベイズ隠れマルコフモデル (VB-HMM) [Beal, 2003] に基づくパラメータ推定アルゴリズムを用いる. (2) VB-HMM により学習したパラメータを用いた複数音源の逐次的定位を行う. 逐次定位では, 観測モデルが VB-HMM より複雑になるため, パーティクルフィルタ [Arulampalam et al., 2002] を用いる.

2 MUSIC 法を用いる音源定位

まず本稿が扱う問題を述べ, MUSIC スペクトルの算出法を説明する. 本稿での水平面上の音源到来方向推定問題を, 図 2 に示した. 今回用いたマイクロフォンアレイは, マイクロフォンがロボットに円状に 8 本配置されており, 水平面上に 5° 刻みの解像度での定位を行う. 以下に本稿で扱う問題設定を示す.

入力 M チャネルの音響信号と, 各周波数ビンごとに D 方向からの伝達関数,
出力 N 個の音源到来方向,
仮定 同時に検出可能な最大音源数 N_{max} はマイクロフォンの数未満 ($N \leq N_{max} < M$).

水平面一周を 5° 刻みに定位するので, $D = 72$ である.

次に, MUSIC スペクトルの算出法について簡単に述べる. より詳細は文献 [Schmidt, 1986; Danès et al., 2010] などに記述されている. MUSIC 法は時間周波数領域[1] において適用される.

$\mathbf{x}_{\tau,\omega} \in \mathbb{C}^M$ を M チャネル音響信号の時間フレーム τ, 周波数ビン ω における複素振幅ベクトルとする. 各周波数ビン ω, ΔT (sec) 間隔の時刻 t に対して, (1) 入力信号の自己相関行列 $\mathbf{R}_{t,\omega}$ の計算, (2) $\mathbf{R}_{t,\omega}$ の固有値分解, (3) 固有ベクトルと伝達関数から MUSIC スペクトルの計算を行う.

(1) 入力信号の自己相関行列は時間 ΔT で観測したサンプル値の相関として計算する.

$$\mathbf{R}_{t,\omega} = \frac{1}{\tau(t) - \tau(t-\Delta T)} \sum_{\tau=\tau(t-\Delta T)}^{\tau(t)} \mathbf{x}_{\tau,\omega}\mathbf{x}_{\tau,\omega}^H, \tag{1}$$

ただし, $(\cdot)^H$ はエルミート転置, $\tau(t)$ は時刻 t に対応する時間フレームを表す. 入力ベクトル $\mathbf{x}_{\tau,\omega}$ の M 個の要素は

[1] 我々の実装では, サンプリング周波数 16000 (Hz) で, 窓長 512 (pt), シフト幅 160 (pt) の短時間フーリエ変換を行っている.

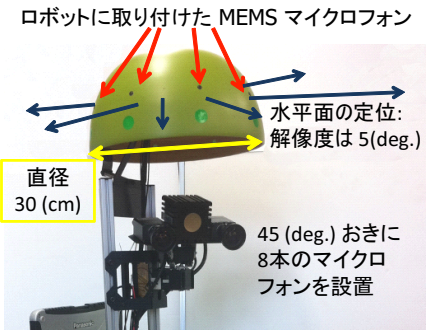


Figure 2: 使用したロボット Kappa. 本稿では水平面上の定位を扱う (青矢印). 8 本のマイクロフォンがロボットのボウルに沿って付けられている (赤矢印).

各チャネルに対応する.

(2) $\mathbf{R}_{t,\omega}$ を次のように固有値分解する.

$$\mathbf{R}_{t,\omega} = \mathbf{E}_{t,\omega}^H \mathbf{Q}_{t,\omega} \mathbf{E}_{t,\omega}, \tag{2}$$

ここで, $\mathbf{E}_{t,\omega}$ は固有ベクトル, $\mathbf{Q}_{t,\omega}$ は固有値から成る対角行列である. $\mathbf{E}_{t,\omega} = [\mathbf{e}_{t,\omega}^1 \ldots \mathbf{e}_{t,\omega}^M]$ と, $\mathbf{R}_{t,\omega}$ の M 個の固有ベクトルで表せ, $\mathbf{Q}_{t,\omega} = \mathrm{diag}(q_{t,\omega}^1 \ldots q_{t,\omega}^M)$ となる. ただし, 固有値 $q_{t,\omega}^m$ は降順に並べられているものとする.

入力信号に N 個の音源が含まれる場合, 固有値 $q_{t,\omega}^1$ から $q_{t,\omega}^N$ までが, 音源のエネルギーに対応する大きな値を持つ. それに対し, 残りの固有値 $q_{t,\omega}^{N+1}$ から $q_{t,\omega}^M$ まではマイクロフォンに伴う観測ノイズなどによる小さな値を取る. ここで重要なポイントは, $\mathbf{e}_{t,\omega}^{N+1}$ から $\mathbf{e}_{t,\omega}^M$ のノイズに対応する固有ベクトルは, 音源到来方向に対応する伝達関数ベクトルと直交するという点である [Schmidt, 1986].

(3) MUSIC スペクトルは以下のように計算する.

$$P_{t,d,\omega} = \frac{\|\mathbf{a}_{d,\omega}^H \mathbf{a}_{d,\omega}\|}{\sum_{m=N_{max}+1}^{M} \|\mathbf{a}_{d,\omega}^H \mathbf{e}_{t,\omega}^m\|}, \tag{3}$$

ただし, $\mathbf{a}_{d,\omega}$ は方向 d, 周波数ビン ω に対応する M 次元の伝達関数ベクトルである. これらの伝達関数はマイクロフォンアレイを用いて事前に測定したものである. 今, 観測されいる最大の音源数は N_{max} 個と仮定している. そのため, $\mathbf{e}_{t,\omega}^{N_{max}+1}$ から $\mathbf{e}_{t,\omega}^M$ までの固有ベクトルは, 常に音源到来方向 d に対応する伝達関数 $\mathbf{a}_{d,\omega}$ と直交する. 従って, 式 (3) の分母は音源到来方向の d に対しては 0 となる. つまり, MUSIC スペクトル $P_{t,d,\omega}$ は ∞ に発散する. ただし, 実際には, 壁からの反射音などの影響で MUSIC スペクトルは発散せず鋭いピークとして観測されることが多い.

周波数ビンごとの MUSIC スペクトルを合算する.

$$P'_{t,d} = \sum_{\omega=\omega_{min}}^{\omega_{max}} \sqrt{q_{t,\omega}^1} P_{t,d,\omega}, \tag{4}$$

ここで, $q_{t,\omega}^1$ は周波数ビン ω における最大固有値である. 我々の実装では, 音声信号を対象とするため, $\omega_{min} = 500$ (Hz), $\omega_{max} = 2800$ (Hz) とした.

従来法では各方向 d に対して, $P'_{t,d} > P_{thres}$ のように閾値

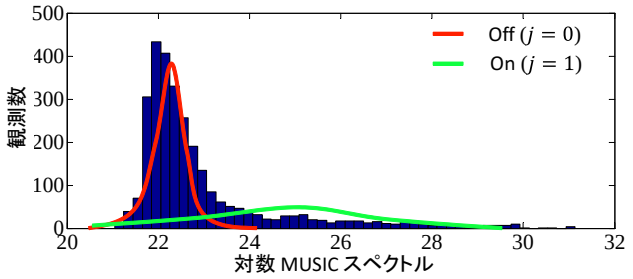


Figure 3: 対数 MUSIC スペクトルの分布. 青: 対数 MUSIC スペクトルのヒストグラム, 赤線: 音源がない場合のガウス分布; 緑線: 音源が存在する場合のガウス分布.

P_{thres} を用いて音源存在判定を行う. しかし, 適切な P_{thres} は残響時間や, 最大音源数 N_{max} に依存するため, 実験的に設定されることが多かった.

3 音源定位のベイズ拡張

MUSIC 法による音源定位のベイズ拡張アルゴリズムは次の 2 ステップから成る. (1) パラメータ学習: VB-HMM を用いて, 対象環境で録音した音響信号からパラメータの事後分布を計算する. (2) オンライン音源定位: パーティクルフィルタを用いて, 学習したパラメータの事後分布を元に複数音源の存在事後確率計算を行う. HMM では状態ベクトルとして D 次元の 2 値ベクトルを用い, 各次元の値が, その方向の音源が存在するか否かを示す. 音源存在閾値 P_{thres} に相当する情報が VB-HMM のパラメータの事後分布として自動的に学習される.

観測モデルはガウス混合モデル (GMM) を用いる. MUSIC スペクトルをガウス分布に従う観測値とみなし, 音源の有無に対応するガウス分布を利用する. ガウス分布を用いる理由は, 複数の周波数ビンの値を加算して対数をとった MUSIC スペクトルが近似的にガウス分布とみなせるためと, ガウス分布を用いることで計算が容易となるためである. 図 3 は対数スケールの MUSIC スペクトルである. 音源が存在しない (Off) のときのガウス分布は狭い MUSIC スペクトルの領域に形成され, 音源が存在する (On) ときの分布は値の広い領域を覆っている. VB-HMM の学習を通じて, 図 3 に示すようなガウス分布のパラメータである平均, 精度 (分散の逆数) の事後分布が計算される.

逐次的な音源定位には以下の 2 要件を満たす観測モデルを利用するためパーティクルフィルタを用いる. (1) 各時刻で同時に存在する音源数は高々 N_{max} 個. (2) $P'_{t,d}$ の極大点にしか音源は存在しない. 詳しい説明は 3.2 節に記す.

3.1 VB-HMM を用いたパラメータ学習

本手法は次の対数 MUSIC スペクトルを観測とする.

$$x_{t,d} = 10\log_{10} P'_{t,d}. \quad (5)$$

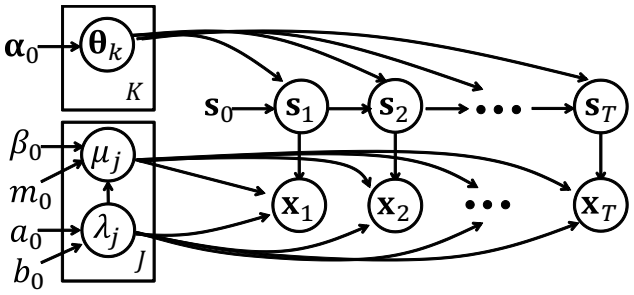


Figure 4: VB-HMM のグラフィカルモデル

$s_{t,d}$ を音源存在を表す 2 値変数し, $s_{t,d} = 1$ のときは時刻 t, 方向 d に音源が存在するものとする.

図 4 に VB-HMM の確率変数間の条件付き独立性を示すグラフィカルモデルを示す. 通常の HMM と VB-HMM との違いは, 状態遷移確率のパラメータ θ_k や, 観測確率のパラメータ μ, λ が固定値ではなく, 確率変数として扱われる点である. これらのパラメータの確率分布を学習し, オンライン音源定位時にはパラメータを積分消去することで, 最尤推定に基づく通常の HMM よりも学習初期値などに頑健な結果を得る.

3.1.1 観測モデル

VB-HMM で用いる観測モデルを以下に示す.

$$p(\mathbf{x}_t|\mathbf{s}_t,\mu,\lambda) = \prod_{d=1}^{D}\prod_{j=0}^{1}\mathcal{N}(x_{t,d}|\mu_j,\lambda_j^{-1})^{\delta_j(s_{t,d})}, \quad (6)$$

ただし, $\delta_y(x)$ は $x = y$ のとき $\delta_y(x) = 1$, さもなくば $\delta_y(x) = 0$ を表す. また, $\mathcal{N}(\cdot|\mu,\lambda^{-1})$ は, 平均 μ, 精度 λ の正規分布の確率密度関数を表す. パラメータ μ と λ には共役事前分布として, 正規-ガンマ分布を用いる.

$$p(\mu,\lambda|\beta_0,m_0,a_0,b_0) = \prod_{j=0}^{1}\mathcal{N}(\mu_j|m_0,(\beta_0\lambda_j)^{-1})\mathcal{G}(\lambda_j|a_0,b_0), \quad (7)$$

ただし, $\mathcal{G}(\cdot|a,b)$ は形状 a, 尺度 b のガンマ分布である.

3.1.2 状態遷移モデル

状態遷移モデルは基本的に, 各方向ビン d について, 前状態で音源がない場合 $s_{t,d} = 0$ と音源がある場合 $s_{t,d} = 1$ から, 次状態で音源が出現する, 継続する, 消滅するといった遷移を考える. 本稿ではさらに, 移動する音源についても考慮するために, 表 1 のように前状態の組み合わせから成る 4 つの場合を考える. すなわち, 前時刻の同方向ビン $s_{t-1,d}$ に音源が存在するかどうかと, 前時刻の隣接方向ビン $s_{t-1,d\pm1}$ のいずれかに音源が存在するかによって分類する. 例えば, θ_1 は前時刻に当該方向 d 及び隣接ビン $d\pm1$ に音源が存在しない状態から音源が出現する確率, θ_2 は, 前時刻に方向 d に音源が存在しないが, 隣接ビン $d\pm1$ には音源が存在したため, その音源が方向 d に移動してきて $s_{t,d} = 1$ となる確率を表す. 状態遷移確率は以下の通り.

Table 1: 隣接状態を考慮した状態遷移の場合分け

前状態 $s_{t-1,d}$	隣接前状態 $1-s_{t-1,d-1}s_{t-1,d+1}$	音源存在確率 $p(s_{t,d}=1\|s_{t-1,d-1:d+1})$
0 (off)	0	θ_1
0 (off)	1	θ_2
1 (on)	0	θ_3
1 (on)	1	θ_4

$$p(\mathbf{s}_t|\mathbf{s}_{t-1},\theta) \quad = \quad \prod_{d=1}^{D}\prod_{k=1}^{4}\prod_{j=0}^{1}\left(\theta_k^{s_{t,d}}(1-\theta_k)^{1-s_{t,d}}\right)^{f_k(\mathbf{s}_{t-1},d)} \tag{8}$$

ここで, $f_k(\mathbf{s}_{t-1},d)$ は表 1 に従って, 方向ビン d の周りの前状態の値 $s_{t-1,d-1}, s_{t-1,d}, s_{t-1,d+1}$ によって条件 k に合致するときに $f_k(\cdot,d)=1$ その他の場合は 0 を返す条件識別関数である. 初期状態としては, 音源は存在しない, すなわちすべての d に対して $s_{0,d}=0$ とする.

状態遷移パラメータである $\theta=[\theta_1,...,\theta_4]$ には, 式 (8) の共役事前分布としてベータ分布を用いる.

$$p(\theta|\alpha_0) \quad = \quad \prod_{k=1}^{4}\mathcal{B}(\theta_k|\alpha_{0,1},\alpha_{0,0}), \tag{9}$$

ただし, $\mathcal{B}(\cdot|c,d)$ はパラメータ c,d を持つベータ分布の確率密度関数である.

3.1.3 事後分布の推定

VB-HMM の学習は, 事後分布 $p(\mathbf{s}_{1:T},\theta,\mu,\lambda|\mathbf{x}_{1:T})$ を以下のように因数分解可能な分布に近似して推定する.

$$\begin{aligned} p(\mathbf{s}_{1:T},\theta,\mu,\lambda|\mathbf{x}_{1:T}) \quad &\approx \quad q(\mathbf{s}_{1:T},\theta,\mu,\lambda), \\ &= \quad q(\mathbf{s}_{1:T})q(\theta)q(\mu,\lambda), \end{aligned} \tag{10}$$

$(\cdot)_{1:T}$ は, 時刻 1 から T までの確率変数の集合を表す. 式 (10) で近似される分布は, 下記の観測変数 $\mathbf{x}_{1:T}$ の対数エビデンスの下限 $\mathcal{L}(q)$ を最大化するよう更新する [Beal, 2003; Bishop, 2006].

$$\log p(\mathbf{x}_{1:T}) = \log\int p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)d\mathbf{s}_{1:T}d\theta d\mu d\lambda \geq \mathcal{L}(q),$$

$$\begin{aligned} \mathcal{L}(q) =& \mathbb{E}_{\mathbf{s}_{1:T},\theta,\mu,\lambda}\left[\log p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)\right] \\ &-\mathbb{E}_{\mathbf{s}_{1:T},\theta,\mu,\lambda}\left[\log q(\mathbf{s}_{1:T})q(\theta)q(\mu,\lambda)\right]. \end{aligned} \tag{11}$$

ただし, $\mathbb{E}_{\mathbf{s}_{1:T},\theta,\mu,\lambda}[\cdot]$ は分布 $q(\mathbf{s}_{1:T})q(\theta)q(\mu,\lambda)$ に関する期待値である. 式 (11) が極大値に収束するまで, 各分布は以下のように交互に更新される.

$$\begin{aligned} \log q(\mathbf{s}_{1:T}) =& \mathbb{E}_{\theta,\mu,\lambda}\left[\log p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)\right] \\ \log q(\theta) =& \mathbb{E}_{\mathbf{s}_{1:T},\mu,\lambda}\left[\log p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)\right] \\ \log q(\mu,\lambda) =& \mathbb{E}_{\mathbf{s}_{1:T},\theta}\left[\log p(\mathbf{x}_{1:T},\mathbf{s}_{1:T},\theta,\mu,\lambda)\right] \end{aligned}$$

事前分布の共役性により, 事後分布は結局以下のように分布のパラメータを更新することと等価である. $q(\theta)=\prod_k q(\theta_k)$ はそれぞれの k に対し, 式 (12) に示すパラメータ $\hat{\alpha}_{k,1}, \hat{\alpha}_{k,0}$ を持つベータ分布となり, $q(\mu,\lambda)=\prod_j q(\mu_j,\lambda_j)$ は, 式 (13), (14) のように, パラメータ $\hat{\beta}_j,\hat{m}_j,\hat{a}_j,\hat{b}_j$ を持つ正規ガンマ分布となる.

$$\hat{\alpha}_{k,j} = \alpha_{0,j} + \sum_{t,d}\langle s_{t,d,j}f_k(\mathbf{s}_{t-1},d)\rangle, \tag{12}$$

$$\hat{\beta}_j = \beta_0 + w_j, \hat{m}_j = (\beta_0 m_0 + w_j\bar{x}_j)/(\beta_0+w_j), \tag{13}$$

$$\hat{a}_j = a_0 + \frac{w_j}{2}, \hat{b}_j = b_0 + \frac{w_j S_j^2}{2} + \frac{\beta_0 w_j(\bar{x}_j-m_0)^2}{2(\beta_0+w_j)}, \tag{14}$$

ただし, 変数 $s_{t,d,j}$ は, $s_{t,d}=0$ のとき, $s_{t,d,0}=1$, また, $s_{t,d}=1$ のとき, $s_{t,d,1}=1$ となる変数である. 式 (13), (14) に用いられる正規分布の十分統計量は

$w_j=\sum_{t,d}\langle s_{t,d,j}\rangle$, $\bar{x}_j=\frac{\sum_{t,d}\langle s_{t,d,j}\rangle x_{t,d}}{w_j}$, $S_j^2=\frac{\sum_{t,d}\langle s_{t,d,j}\rangle(x_{t,d}-\bar{x}_j)^2}{w_j}$.

と定義する. また, $\langle\cdot\rangle$ は式 (10) の分布による期待値演算子である. $q(\mathbf{s}_{1:T})$ に対応する, 各時刻の状態変数と状態遷移の期待値 $\langle s_{t,d,j}\rangle$, $\langle s_{t,d,j}f_k(\mathbf{s}_{t-1},d)\rangle$ は次のように計算する.

$$\langle s_{t,d,j}\rangle \propto \alpha(s_{t,d,j})\beta(s_{t,d,j}), \tag{15}$$

$$\langle s_{t,d,j}f_k(\mathbf{s}_{t-1},d)\rangle \propto \tilde{\alpha}(s_{t-1,d,k})\tilde{p}(s_{t,d}|\mathbf{s}_{t-1})\tilde{p}(x_{t,d}|s_{t,d})\beta(s_{t,d,j}), \tag{16}$$

ただし, $\alpha(s_{t,d,j})$ と $\beta(s_{t,d,j})$ はそれぞれ前向き・後ろ向き再帰式により計算される.

$$\alpha(s_{t,d,j}) \propto \sum_{k=1}^{4}\tilde{\alpha}(s_{t-1,d,k})\tilde{p}(s_{t,d}|\mathbf{s}_{t-1})\tilde{p}(x_{t,d}|s_{t,d}), \tag{17}$$

$$\beta(s_{t,d,j}) = \sum_{j'=0}^{1}\beta(s_{t+1,d,j'})\tilde{p}(s_{t+1,d,j'}|s_{t,d,j})\tilde{p}(x_{t,d}|s_{t,d}). \tag{18}$$

式 (16) 遷移, 観測確率の幾何平均は次の通り.

$$\tilde{p}(s_{t,d}=j|\mathbf{s}_{t-1}) \propto \quad \prod_{k=1}^{4}\exp\left\{\psi(\hat{\alpha}_{k,j})-\psi(\hat{\alpha}_{k,0}+\hat{\alpha}_{k,1})\right\}^{f_k(\mathbf{s}_{t-1},d)}, \tag{19}$$

[2]

$$\tilde{p}(x_{t,d}|s_{t,d}) \propto \prod_j \exp\left\{\frac{\psi(\hat{a}_j)-\log\hat{b}_j-1/\hat{\beta}_j}{2} - \frac{a_j(x_{t,d}-\hat{m}_j)^2}{2\hat{b}_j}\right\}^{s_{t,d,j}} \tag{20}$$

式 (15), (16) はともに, 添字 j,k を動かしたとき総和が 1 になるように正規化されている. $\tilde{\alpha}(s_{t-1,d,k})$ は, 状態遷移の条件 k に関する前向き確率である. 本節で示されたパラメータ更新式 (12)–(16) が収束するまで計算される. 初期値としては, $\langle s_{t,d,j}\rangle$ と $\langle s_{t,d,j}f_k(\mathbf{s}_{t-1},d)\rangle$ の値を, 観測変数 $x_{t,d}$ の値を m_0 の値を閾値として処理することで, 0 ないし 1 を与えることで行う.

3.2 パーティクルフィルタによるオンライン音源定位

本節ではパーティクルフィルタ [Arulampalam et al., 2002] を用いた, オンライン音源定位手法を述べる. オンライン推定では, 式 (12)–(14) で求めたパラメータの事後分布を利用する. パーティクルフィルタの推定対象は, MUSIC スペクトルの時系列データが与えられたときの, 各方向ビンにおける音源存在事後確率である. この分布を P 個のパーティクルを用いて以下のように近似計算する.

$$p(\mathbf{s}_t|\mathbf{x}_{1:t}) \quad \approx \quad w_p\mathbf{s}_t^p, \tag{21}$$

[2] $\psi(\cdot)$ はディガンマ関数.

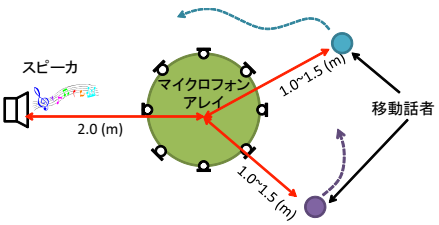


Figure 5: 実験条件: マイクロフォンアレイの周囲を動く移動話者と固定音源

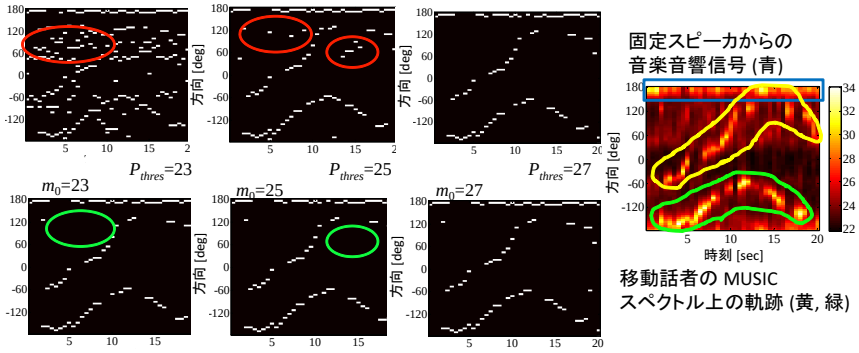


Figure 6: 音源定位結果: 白が音源が存在する方向, 時間ビン. 上図: 固定閾値 P_{thres} による定位結果. 下図: 初期値 m_0 を変えた場合の本手法による定位. 右図: 観測された対数 MUSIC スペクトル. 音楽音響信号が 180 [deg] 付近に存在し, 2 人の話者が移動している.

ただし, w_p はパーティクル p の重み, $\mathbf{s}_t^p$ は状態ベクトルの値である. これらの w_p と $\mathbf{s}_t^p$ は次のように得る.

(1) 提案分布から $\mathbf{s}_t^p$ をサンプルする.

$$\mathbf{s}_t^p \sim q(\mathbf{s}_t|\mathbf{x}_t, m, a, b), \tag{22}$$

$$q(\mathbf{s}_t^p|\mathbf{x}_t, \hat{m}, \hat{a}, \hat{b}) \propto \prod_d \prod_{j=0}^{1} C(x_{t,d})^{s^p_{t,d,1}} \exp(-\Delta^2_{d,j}/2)^{s^p_{t,d,j}}, \tag{23}$$

ただし, $x_{t,d}$ が極大値を取る d のとき, $C(x_{t,d}) = 1$ でその他の場合は $C(x_{t,d}) = 0$ となる. この項は, 時間 t の中で, $x_{t,d}$ の極大方向 d にのみ音源が存在する, つまり $s_{t,d} = 1$ となるよう導入されている. 提案分布の重みにはマハラノビス距離 $\Delta^2_{d,j} = (x_{t,d} - \hat{m}_j)^2 \hat{a}_j / \hat{b}_j$ を用いる.

(2) 各パーティクル p について, 重み w_p を算出.

$$w_p \propto \frac{\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p)\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p)}{q(\mathbf{s}_t^p|\mathbf{x}_t, \hat{m}, \hat{a}, \hat{b})}, \tag{24}$$

$$\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p) = \prod_d C(x_{t,d})^{s^p_{t,d,1}} \int p(\mathbf{x}_t|\mathbf{s}_t^p, \mu, \lambda) q(\mu, \lambda) d\mu d\lambda, \tag{25}$$

$$\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p) = \int p(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p, \theta) q(\theta) d\theta. \tag{26}$$

式 (25),(26) にある状態遷移, 観測確率は, VB-HMM で計算された式 (6),(8) の事後分布で積分消去することで計算できる. これにより, VB-HMM で学習されたパラメータの曖昧性を考慮したオンライン定位を行う. なお, 式 (25) の $C(x_{t,d})^{s^p_{t,d,1}}$ の項は, 式 (23) と同様に, $x_{t,d}$ の極大方向 d にのみ音源の存在を許す項である. 分布の共役性を用いると, この積分計算は次のように解析的に求まる.

$$\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p) = \prod_d C(x_{t,d})^{s^p_{t,d,1}} \mathcal{S}t(x_{t,d}|\hat{m}_j, \frac{\hat{\beta}_j \hat{a}_j}{(1+\hat{\beta}_j)\hat{b}_j}, 2\hat{a}_j)^{s^p_{t,d,j}}, \tag{27}$$

$$\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p) = \prod_d \prod_k \left(\hat{\alpha}_{k,s_{t,d}} / (\hat{\alpha}_{k,0} + \hat{\alpha}_{k,1})\right)^{f_k(\mathbf{s}^p_{t-1}, d)}, \tag{28}$$

ただし, $\mathcal{S}t(\cdot|m, \lambda, \nu)$ は平均 m, 精度 λ, 自由度 ν の Student t-分布である. さらに, 最大の音源数を N_{max} に抑えるため, 状態ベクトル $\mathbf{s}_t^p$ に存在する音源数が N_{max} を超える場合には観測確率は 0 とする.

全パーティクルの重み計算後, 各パーティクルの重み w_p は $\sum_{p=1}^P w_p = 1$ となるよう正規化する. この手順に従い, 式 (21) の音源存在の事後分布を計算する. 我々の実装手法では, 各ステップごとにパーティクルが持つ重みに比例してリサンプリング処理が行われる.

4 評価実験

評価実験では, VB-HMM によるパラメータ分布推定とパーティクルフィルタを用いたオンライン音源定位から成る本手法と, 従来の固定閾値を用いて音源定位する手法を比較する. オフラインでの VB-HMM での学習は, 1 人の話者がマイクロフォンの周囲を発話しながら動く音響信号で行った. オンラインの音源定位実験に使用した音源の配置を図 5 に示す. マイクロフォンアレイの周囲を移動する 2 話者と, 固定されたスピーカから音楽が再生されている. オフライン, オンラインで用いられた信号の長さはともに 20 (sec) である. パラメータの設定は次の通り. 観測信号の自己相関行列を計算する窓幅 $\Delta T = 500$ (msec), $N_{max} = 3$, $\alpha_0 = [1, 1]$, $\beta_0 = 1$, $a_0 = 1$, $b_0 = 500$. パーティクル数は $P = 500$ とした. 実験で使用した室内の残響時間は $RT_{20} = 840$ (msec) であった.

図 6 にオンライン音源定位の結果を示す. 従来法の閾値は $P_{thres} = 23, 25, 27$ に設定されており, 本手法の初期値は $m_0 = 23, 25, 27$ に設定されている. パーティクルフィルタの定位結果の図では, 事後分布の音源存在確率が 0.95 以上のビンを音源が存在するとして白く表示している. 従来法においては, 閾値を低く設定した場合は図 6 の赤枠で示すように音源の誤検出が頻発する. 対して, 本手法では緑枠で示すように, 学習の初期値に対して頑健に妥当な音源定位結果を示している. また, 本手法において音源存在確率の閾値を 0.95–1.00 まで動かして結果を検証したが, こ

れらの値を閾値に対しても頑健に同様の結果を示すことを確認した. この結果から, 本手法におけるオフライン学習, オンライン定位の枠組みが, 自動的に音源定位に適したパラメータに収束することが確認できる. さらに, 今回の実験条件から, 本手法は学習時に 1 音源しか用いなくても, 複数音源に対して安定したオンライン定位が可能であることが確認された.

4.1 議論と今後の課題

実験を通じて, 本手法は学習初期値や, 学習時とオンライン推定時の音源数ミスマッチに頑健であることを示した. しかし, 本手法には次の制約が存在する. (1) 音源ごとの軌跡は直接は推定されない. (2) 音声のポーズ等に応じて定位が途切れる. 混合音に含まれる各音源の定位結果を元に音源分離を行うシステムでは (例 [Nakadai et al., 2010]), 安定した音源分離のために音源ごとの軌跡, ポーズ等を接続した定位が重要である.

(1) 本稿で示した状態空間モデルでは, 各時間フレームで音源が存在する方向ビンを推定する. 音源ごとのトラッキング結果が必要な場合, 連続する時間フレームで近い定位結果をスムージングするといった後処理や, あるいは, 複数音源の移動を状態遷移モデルに組み込む必要がある.

(2) 音声では文の終わり等にポーズがしばしば入り, 対応する時間フレームの MUSIC スペクトルの値は減少する. 図 6 で示された本手法による定位結果でも, 0°, 6 (sec) 付近の話者の定位が途切れている様子が示されている. この問題も, 後処理で途切れた軌跡をつなげるといった方法や, 音のポーズを明示的に状態モデルに取り込むといった手法の改良による対処が考えられる.

5 まとめ

本稿では MUSIC 法に基づく音源定位法のベイズ拡張を述べた. 本手法は, (1) VB-HMM によるパラメータの自動学習, (2) パーティクルフィルタを用いたオンライン音源定位から成る. 評価実験では, $RT_{20} = 840$ (msec) の残響環境下で, 1 音源の音響信号の学習に対し, 3 音源同時音源定位を実現した. 今後の展開としては, 実際に移動ロボットに本手法を適用して, ロボット位置と環境中に存在する音源位置の推定を通した音環境理解システムの構築などが挙げられる.

謝辞: 本研究の一部は科研費特別研究員奨励金/基盤 (S), JST-ANR BINAAHR, GCOE の支援を受けた.

参考文献

[Arulampalam et al., 2002] M. Arulampalam et al.: A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking, *IEEE Trans. on Signal Proc.*, Vol. 50, No. 2, pp. 174–189, 2002.

[Asano et al., 2001] F. Asano et al.: Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition, *Proc. of Eurospeech*, pp. 1013–1016, 2001.

[Beal, 2003] M. J. Beal: Variational Algorithms for Approximate Bayesian Inference, *Ph.D. thesis*, Gatsby Computational Neuroscience U., Univ. Colledge London, 2003.

[Bishop, 2006] C. M. Bishop: Chapter 10, Approximate Inference, *Pattern Recognition and Machine Learning*, Springer, 2006.

[Danès et al., 2010] P. Danès and J. Bonnal: Information-Theoretic Detection of Broadband Sources in a Coherent Beamspace MUSIC Scheme, *Proc. of IROS*, pp. 1976–1981, 2010.

[Doclo et al., 2001] S. Doclo and M. Moonen: GSVD-based optimal filtering for multi-microphone speech enhancement, *Microphone arrays*, pp. 111–132, Springer, 2001.

[Kubota et al., 2008] Y. Kubota et al.: Design and Implementation of 3D Auditory Scene Visualizer towards Auditory Awareness with Face Tracking, *Proc. of IEEE Int'l Symboposium on Multimedia (ISM-2008)*, pp. 468–476, 2008.

[Mizumoto et al., 2011] T. Mizumoto et al.: Design and Implementation of Selectable Sound Separation on a Texai Telepresence System using HARK, *Proc. of ICRA*, pp. 2130–2137, 2011.

[Nakadai et al., 2010] K. Nakadai et al.: Design and Implementation of Robot Audition System "HARK", *Advanced Robotics*, Vol. 24, No. 5–6, pp. 739–761, 2010.

[Sasaki et al., 2010] Y. Sasaki et al.: Map-Generation and Identification of Multiple Sound Sources from Robot in Motion, *Proc. of IROS*, pp. 437–443, 2010.

[Schmidt, 1986] R. O. Schmidt: Multiple Emitter Location and Signal Parameter Estimation, *IEEE Trans. on Antennas and Propagation*, Vol. 34, No. 3, pp. 276–280, 1986.

[Yamamoto et al., 2006] K. Yamamoto et al.: Detection of Overlapping Speech in Meeting using Support Vector Machines and Support Vector Regression, *IEICE Trans. Fundamentals*, Vol. E89-A, No. 8, pp. 2158–2165, 2006.

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-B102-7

AUDIO TRACKING FOR SMALL MEETINGS USING LASER RANGE FINDERS AND LOCAL AUDIO SCANS

Jani Even, Panikos Heracleous, Carlos Ishi, Takahiro Miyashita and Norihiro Nogita

ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan
even@atr.jp

ABSTRACT

This papers presents a system designed for separating and tracking the voices of a few persons talking around a table. During the meeting, the locations of the participants are monitored by a human tracker system based on laser range finders (LRFs). Then using a uniform circular array (UCA) of microphones, audio localization is performed to estimate the most powerful sound source, usually the mouth, in the neighborhood of each of the detected participants. Finally, beamforming is applied to obtain an audio stream for each of the detected participants. The use of LRF based human tracker enables the system to assign a continuous audio track to each of the participants. Experimental results using real meeting data show the efficiency of the proposed approach.

1. INTRODUCTION

An important task in meeting transcription is speaker diarization (i.e. to find "Who talked when") [1]. It is quite common to use a microphone array or distributed microphones to obtain one stream for each active participant (for example with audio beamforming in [2] or using the directions of arrival [3]). Then in order to create the diary of the meeting, speaker identification and activity detection is performed using these streams.

In this paper, we propose an extension of the multi-modal approach to this problem we presented in [4]. Contrary to other approaches, the speaker localization is not performed using only the audio signals. In addition to the audio signals recorded by a uniform circular microphone array (UCA), laser range finders (LRFs) are used to obtain distances. First the locations of the participants are estimated by a human tracker system based on laser range finders (LRF) [5] then these positions are refined using the audio signals to scan with a beamformer the neighborhood of each of the positions given by the human tracker (this was not done in [4]). This scan, referred to as local scan, is based on the broadband MUSIC algorithm (see [6] for details on the different broadband MUSIC approaches). In particular, the power of the MUSIC pseudo spectrum is used to determine the activity of the participants.

THIS WORK WAS SUPPORTED BY KAKENHI (21118003).

After localization and activity detection, the audio data are processed in order to obtain an enhanced audio stream for each of the active participants at all time. A specificity of the proposed method is that silent participants are also assigned an audio stream because they are detected by the LRF based human tracker. Experiments were conducted in a realistic meeting situation to demonstrate the efficiency of the proposed method. In order to underline the gain of using the human tracker, a conventional broadband MUSIC algorithm was also used. For this algorithm, the tracking is performed by using spatial information but also Gaussian mixture models (GMMs) [7] that were trained before hand.

2. METHOD

2.1. Localization

The motion of the participants in the meeting area is monitored using 4 LRFs mounted on poles around the meeting area's perimeter. To reduce the errors due to noise and occlusion, each person is tracked with a particle filter using a linear motion model with random perturbations (see [5]). The human tracker gives the position $\{x, y\}$ of the torso of each of the participants in the room. However, the positions that matter are not the positions of the participants but the positions of their mouths. Consequently, the positions given by the human tracker have to be refined. In particular, the z coordinates have to be estimated.

For this purpose, a local audio scan is applied around each of the positions given by the human tracker to estimate the position of the mouth (see Fig. 1). This local audio scan is based on the MUSIC algorithm.

The raw audio signals, referred to as the *observed signals* in the remainder, are acquired by a uniform circular array (UCA) of $m = 16$ microphones positioned on a table in the middle of the meeting area. The position of the microphone array is assumed to be known.

For the localization purpose, the frequency domain observation is obtained by using a short time Fourier transform with a hanning window of 51 points, a shift of 25 points and an fft size of 64 points. The localization is performed every 200 ms corresponding to 128 frequency frames. The vector

of observed signals in the fth frequency bin is

$$\mathbf{X}_L(f,k) = \left[\; X_1(f,k), \quad \cdots \quad , X_m(f,k) \; \right]^T$$

where k denotes the frame index.

For a selected number of frequency bins, the narrow band MUSIC pseudo power spectrum $\mathbf{P}_{nb}(f,x,y,z)$ is obtained by

- performing a singular value decomposition of the observation covariance $\mathbf{\Gamma}(f) =< \mathbf{X}_L(f,k)\mathbf{X}_L^H(f,k) >_k$,
- creating the projector $\mathbf{P}_K(f)$ on the space spanned by the K least powerful singular values,
- scanning the space around the LRF position by using a beamformer $\mathbf{W}(f,x,y,z)$
- estimating the pseudo power by

$$\mathbf{P}_{nb}(f,x,y,z) = \frac{1}{\mathbf{W}(f,x,y,z)\mathbf{P}_K(f)\mathbf{W}^H(f,x,y,z)}.$$

Then the broadband MUSIC pseudo spectrum is obtained by averaging the narrow band pseudo spectra

$$\mathbf{P}_{bb}(x,y,z) =< \mathbf{P}_{bb}(f,x,y,z) >_f \; .$$

For each of the positions $\{x_0, y_0\}$ given by the human tracker, the updated position $\{x,y,z\}$ gives the maximum of the broadband MUSIC pseudo spectrum estimated in the space around $\{x_0, y_0\}$. The participant is considered active if that maximum pseudo spectrum is above a threshold ϵ_p.

For each of the 200 ms block, the proposed method detect if a participant is active and at the same time give a refined estimate of the mouth position of this active participant. Note that the activity of the participants along the 200 ms blocks is tracked by the human tracker even if the participants are silents.

2.2. Audio stream

At any time, an audio stream is assigned to each of the Q active participants. The desired streams are obtained by processing the observed signals in the frequency domain. For the beamforming purpose, the frequency domain observation is obtained by using a short time Fourier transform with a hanning window of 401 points, a shift of 200 points and an fft size of 512 points. The beamforming is performed every 200 ms corresponding to 16 frequency frames. The vector of observed signals in the fth frequency bin is

$$\mathbf{X}(f,k) = \left[\; X_1(f,k), \quad \cdots \quad , X_m(f,k) \; \right]^T$$

where k denotes the frame index.

First, the refined positions (in the microphone array referential) are used to estimate a set of delay and sum (DS) beamformers. Only considering the delays for a direct path propagation we can write the set of DS beamformers as

$$\mathbf{Y}_{\text{DS}}(f,k) = \begin{bmatrix} \mathbf{w}_1(f,k) \\ \vdots \\ \mathbf{w}_Q(f,k) \end{bmatrix} \mathbf{X}(f,k)$$

where $\mathbf{Y}_{\text{DS}}(f,k)$ are the beamformed audio streams and the $Q \times m$ matrix has general term

$$w_{ij}(f,k) = e^{-j2\pi f \frac{r_{ij}(k) - r_{i1}(k)}{c}}$$

with c the celerity of the sound and $r_{ij}(k)$ the distance between the mouth of the ith participant and the jth microphone (the first microphone is used as reference).

Then an audio stream for each of the participants is obtained by applying a linearly constrained minimum variance (LCMV) beamformer.

The LCMV beamformer weights for the ith participant are given by

$$\mathbf{w}_{\text{LCMV},i}(f,k) = \frac{\mathbf{w}_i(f,k)\mathbf{K}^-1(f)}{\mathbf{w}_i(f,k)\mathbf{K}^-1(f)\mathbf{w}_i^H(f,k)}$$

where $\mathbf{K}(f)$ is the estimate of the noise and interference covariance and $\mathbf{w}_i(f,k)$ is the steering vector pointing to the ith participant.

The estimate of the noise and interference covariance is composed of two parts

$$\mathbf{K}(f) = \mathbf{\Gamma}(f) + \sum_{j=1, j\neq i}^{Q} \mathbf{w}_j^H(f,k)\mathbf{w}_j(f,k)\sigma_j^2(f).$$

The first term $\mathbf{\Gamma}(f)$ is the estimate of the noise covariance obtained when only the noise is present. The second term represents the contribution of the other participants (the interferences). It is a sum of the contributions made by each interfering participants. The interfering participants are represented by point sources located at the positions given by the human tracker. For each of these point sources, the DS beamformer is used to obtain the power which is estimated by

$$\sigma_j(f) \quad = \quad \text{var}\left\{ \mathbf{Y}_{\text{DS}}^{(j)}(f,k) \right\}.$$

Note that for a silent participant, this power is likely to be small.

Finally, the audio stream of the ith participant is

$$Y_{\text{LCMV},i}(f,k) = \mathbf{w}_{\text{LCMV},i}(f,k)\mathbf{X}(f,k)$$

The LCMV beamformers provide an audio stream for each of the detected participants that contains less interference from the other participants and fewer environmental noise than the DS beamformer streams. In the remainder, we refer to $Y_{\text{LCMV},i}(f,k)$ by $Y_i(f,k)$ for convenience.

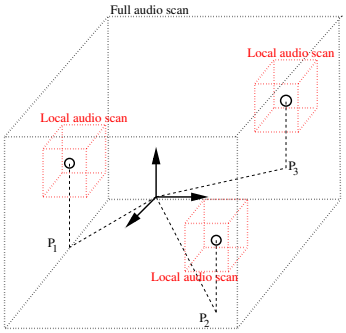


Fig. 1. Full and local audio scans for three positions P_1, P_2 and P_3

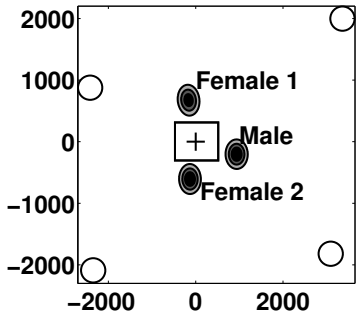


Fig. 3. Microphone array (cross), pole mounted LRFs (circles), table (rectangle) and probability densities of the three speakers position (distances are in mm)

3. EXPERIMENTS

Two different cases were compared where the audio stream of each participant is obtained by: using localization based only on audio signals (MUSIC; the full scan in Fig. 1) and based on human tracker and audio signals (LRF + MUSIC; the local scans in Fig. 1).

3.1. Experimental setup

The experiment setup is described in Fig. 3. The four circles in the corners represent the pole mounted LRFs used by the human tracker, the cross gives the position of the microphone array and the probability densities of the positions of the three speakers during the experiment also appear (note that the densities are sharp even if the speakers were not told to limit their movements). The experiment setup consists of four pole mounted LRFs (Fig.2 right) in the corner of the monitored area and of a table top UCA (Fig.2 left).

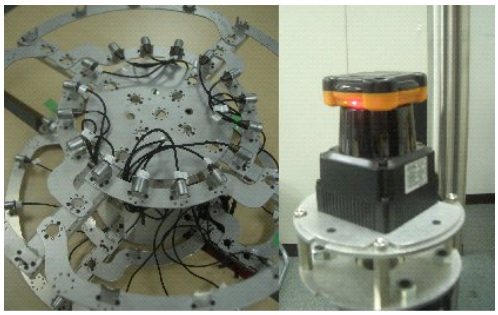


Fig. 2. Table top microphone array (left) and pole mounted LRF (right).

3.2. Data set

In this experiment, three participants were considered (2 females and 1 males). In the remainder of the paper, the speakers are designated by the letters $\{a, b, c\}$. Two test sets were recorded in a room while monitoring the speaker movement with the LRF based human tracker system. The three participants were sitting around a table (the participants were not given any instruction concerning their movements). A first test set, referred to as *reading set* is obtained by letting the participants read some sentences from the JNAS database. First b and c are reading at the same time then after a short pause a and b are reading at the same time. The second test set, referred to as *conversation test*, is extracted form a real conversation between the three participants and includes speech and interjections. The activity of the participants was hand labeled for both of the test sets. The observed signal from microphone 1 is given for each test set in Fig.4.

3.3. Conventional broadband MUSIC

To show the advantage of using human tracker system for the diarization, a conventional broadband MUSIC approach was also used.

For the conventional broadband MUSIC algorithm, only one broadband pseudo spectrum is obtained by scanning the whole space then in the selected frequency bins. Then the number of audio sources is determined by finding the local maxima of the pseudo power spectrum that are above the threshold ϵ_p. The localization is also performed every 200 ms using 128 frames. Then audio streams are obtained for each of the detected audio sources using the same beamforming technique as for the LRF + MUSIC case.

However, a big difference is that the detected audio sources from each of the 200 ms blocks have to be combined together to create the audio tracks. For audio sources active in consecutive blocks, the distance between the sources is used to combine them: sources that did not move much are considered the same. For combining sources that are inactive for several blocks, it was necessary to use speaker identification based on GMMs [7]. The features extracted from the audio streams are the MFCCs (12 MFCCs and the

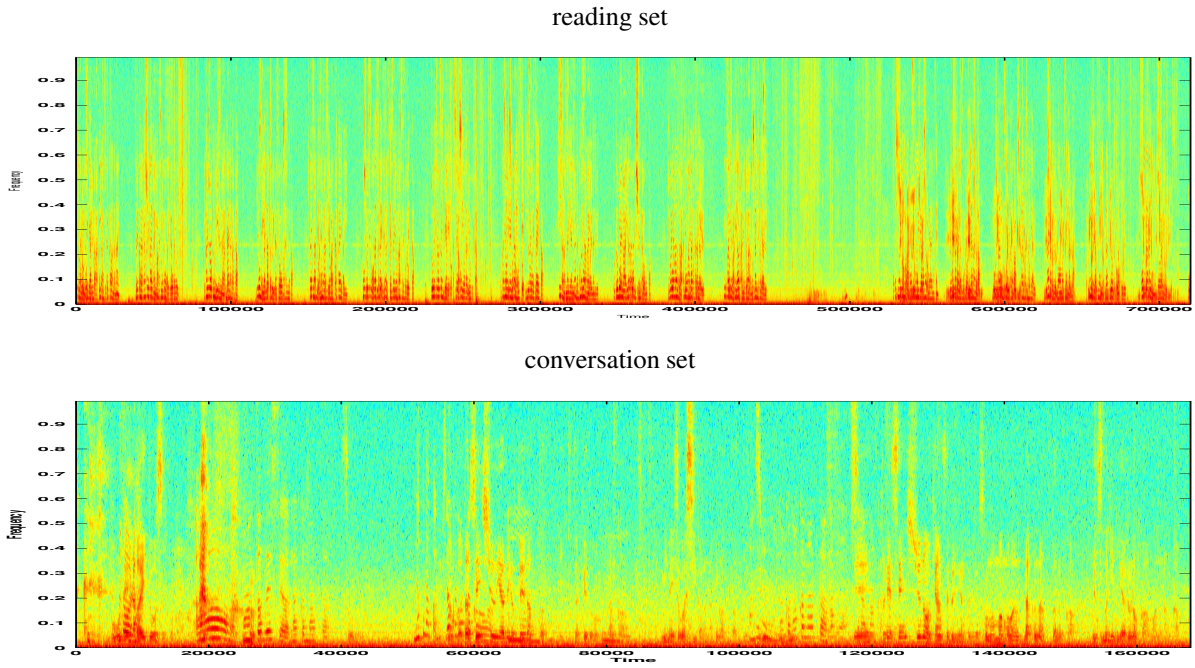


Fig. 4. Observations for reading set (top) and conversation set (bottom).

log spectral energy, their derivatives and their accelerations). For each speaker a common training set of 100 Japanese sentences from the JNAS database [8] was recorded using a close talking microphone while sitting at the table in the experiment room. A set of GMMs was trained for each of the speakers using these 100 utterances and a general GMM was also trained using the 300 utterances (referred to as GGMM). The GMMs for all the speakers are designated by $\{\lambda_a, \lambda_b, \lambda_c\}$ and the GGMM by $lambda_G$. The number of mixtures was set to 512 after testing several values. Training and testing were performed with HTK 3.41 using the whole utterances.

The GMMs are used to determine for each block which of the participants is active. In this paper, for a given block the likelihoods are normalized using the following likelihood ratio (For decision based on likelihood, it is usually necessary to apply a normalization [9, 10])

$$\mathcal{L}(Y_q|\lambda_i) = \log p(Y_q|\lambda_i) - \log p(Y_q|\lambda_G).$$

where λ_G is the general GMMs estimated on all training utterances.

The decision rule is to select for each of the block the speaker whose model has the largest likelihood

$$\mathcal{L}(Y_q|\lambda_j) \quad = \quad \max_i \mathcal{L}(Y_q|\lambda_i)$$

as the active speaker.

Table 1. Deletion, insertion and correct percentages for the MUSIC method.

	reading set			conversation set		
	del	ins	cor	del	ins	cor
a	0.0	6.5	93.5	0.5	23.7	75.8
b	1.1	23.7	75.2	10.2	14.3	75.5
c	0.0	18.0	81.9	10.8	8.2	81.1
avg.	0.4	16.1	83.5	7.2	15.4	77.4

Table 2. Deletion, insertion and correct percentages for the LRF + MUSIC method.

	reading set			conversation set		
	del	ins	cor	del	ins	cor
a	0.2	9.2	90.6	1.4	18.3	80.3
b	0.6	29.7	69.7	2.5	17.3	80.2
c	0.4	12.2	87.4	1.8	13.7	84.5
avg.	0.4	17.0	82.6	1.9	16.4	81.7

3.4. Diarization

The results of the meeting diarization are given in terms of deletion, insertion errors in Table 1 and 2:

- An insertion error occurs when a speaker is detected for the audio stream of a silent participant.
- A deletion error occurs when an active participant is not detected.

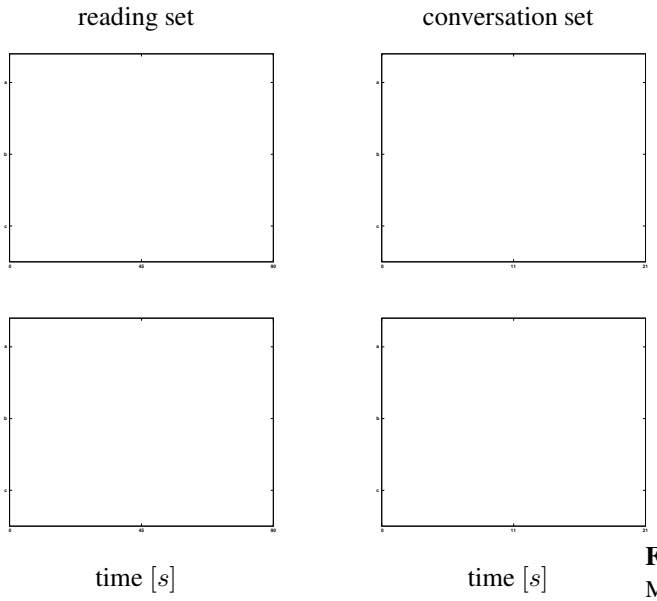


Fig. 5. Result of the diarization for the two test sets with MUSIC (top) and LRF + MUSIC (bottom).

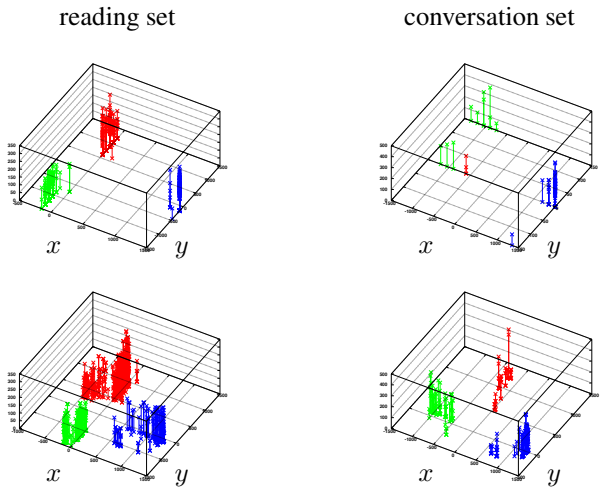


Fig. 6. MUSIC Pseudo power in space for the two test sets with MUSIC (top) and LRF + MUSIC (bottom).

These percentages are computed by comparing the hand labeled activity with the activity given by both of the methods. Figure 5 gives a graphical representation of the diarization results. For each of the sub-figure, one row correspond to one of the three participants. The color code shows the deleted samples (red), the inserted samples (blue) and the correctly detected samples (green).

We can see that using both the LRFs and the audio data for the localization gives the best performance for the conversation set but for the reading set there is not much difference (it is also faster than the full audio scan).

3.5. Localization

Figure 6 shows the repartition of the detected block power in the space for the three participants (a in blue, b in red and c in green) in the two data sets. We can especially see that for the conversation set, the MUSIC method has a bad estimate for the speakers b and c that are the two female speakers as the GMMs trained on reading conditions are not good for the interjections present in the conversation set.

4. CONCLUSION

This paper presents a multi-modal approach to the diarization problem that combines LRF base human tracker with microphone array. In particular using LRF is an efficient way to perform the tracking the participants and merge the detected audio blocks together.

5. REFERENCES

[1] J.G. Fiscus, J. Ajot, and J.S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," *Lecture note in computer science*, vol. 4625, pp. 373–389, 2008.

[2] F. Asano et al., "Detection and separation of speech events in meeting recordings using a microphone array," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. ID 27616, 2007.

[3] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A doa based speaker diarization system for real meeting," *HSCMA 2008, Trento, Italy*, pp. 29–32, 2008.

[4] J. Even, P. Heracleous, C. Ishi, and N. Hagita, "Multimodal front-end for speaker activity detection in small meetings," *Proceedings of 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 536–541, 2011.

[5] D.F. Glas et al., "Laser tracking of human body motion using adaptive shape modeling," *IROS 2007, San Diego, USA*, pp. 602–608, 2007.

[6] S. Argentieri and P. Danès, "Broadband variations of the music high-resolution method for sound source localization in robotics," *IROS-2007, San Diego, USA*, pp. 2009–2014, 2007.

[7] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transaction on speech and audio processing*, vol. 3, no. 1, pp. 72–82, 1995.

[8] K. Ito et al., "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoust. Soc. of Japan*, vol. 20, pp. 196–206, 1999.

[9] A. Rosenberg, J. DeLong, C. Lee, B.H. Juang, and F. Soong, "The use of cohort normalized scores for speaker verification," *Proc. ICSLP*, pp. 599–602, 1992.

[10] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech communication*, vol. 17, no. 1-2, pp. 109–116, 1995.

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-B102-8

Kinectにおけるリアルタイム・ブラインド空間サブトラクションアレーの実装と評価

Implementation and Evaluation of Real-Time Blind Spatial Subtraction Array on Kinect

大沼 侑司[1*]　鎌土 記良[1]　宮崎 亮一[1]
猿渡 洋[1]　鹿野 清宏[1]
Yuji Onuma[1]　Noriyoshi Kamado[1]　Ryoichi Miyazaki[1]
Hiroshi Saruwatari[1]　Kiyohiro Shikano[1]

[1] 奈良先端科学技術大学院大学
[1] Nara Institute of Science and Technology

Abstract: In this paper, we propose a new noise-robust hands-free speech recognition system with a 'kinect' for the robot audition based on the real-time blind spatial subtraction array (BSSA). Kinect is a multi-modal interface which consists of sensor devices such as the motion detector, the colored image sensor and the microphone array. In our previous study, we have developed the hands-free speech recognition system with a linear microphone array based on BSSA. The proposed system in this paper is improved to obtain not only acoustical information but also visual information such as an accurate direction of the speakers by using kinect for improving the recognition rate of the first utterance. In this paper, as the first step, we implemented BSSA on kinect and we assessed the performance of the noise reduction via a speech recognition tests under an actual environment to verify the feasibility of the microphone array in kinect. The results of the experiments clarify that the proposed system markedly improves the speech recognition performance in typical noisy environments.

1　はじめに

人と音声コミュニケーションを行うロボット対話システムでは，ユーザからはなれた位置にマイクロホンを設置して音声認識を行うハンズフリー音声認識が必要不可欠である．しかし，実環境下においては，周囲に存在する環境雑音や残響，さらにはファンノイズやロボット自体が発する音声などによって，音声認識の性能が低下する問題がある．従って，ロボットが高精度に音声認識を行うためには，雑音環境下においても目的音声を高精度に抽出可能なシステムの実現が必要不可欠であると言える．しかし，システムの設置される環境によっては，周囲の環境雑音は非定常なものであり，単純な Wiener Filter (WF) を用いた雑音抑圧では十分な雑音抑圧性能を得られないことも考えられる．また，本稿で用いるマルチモーダル・インターフェースである Kinect [7] のマイクロホンアレーを含め，特に安価に提供可能なマイクロホンアレーにおいては，たとえ同時期に製造された同型のマイクロホン素子であっても素子誤差が存在する．雑音抑圧に用いる手法によっては，素子誤差は雑音抑圧の性能に悪影響をもたらしたり，システムの運用前にマイクロホン素子のキャリブレーションを必要とさせる可能性がある．また，Kinect に搭載されているようなロボット聴覚として実用的な小型のマイクロホンアレーでは，遅延和アレー (Delay and Sum : DS) [1, 2, 3] などの大規模アレーを必要とする手法などは実用的では無い．

我々は，小型のマイクロホンアレーでも実環境下において現実的な計算コストで効果的に雑音抑圧を行うことのできる手法として，ブラインド空間的サブトラクションアレー (Blind Spatial Subtraction Array : BSSA) [4] を提案している．これは，マイクロホン素子誤差や残響の影響による雑音推定精度の劣化を抑制可能な独立成分分析 (Independent Component Analysis : ICA) に基づいた手法である．ICA は，拡散性雑音の多い環境下では，音声信号の推定よりも拡散性雑音の推定精度が高いということが知られている．BSSA は，ICA

*連絡先：奈良先端科学技術大学院大学 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
E-mail: yuji-o@is.naist.jp

のこの特徴を利用した手法であり，DS により目的信号を強調した音声から，ICA により推定した雑音をスペクトル減算 (Spectral Subtraction：SS) することで雑音を抑圧する．これにより，ICA のみを用いた場合よりも高精度の雑音抑圧が可能である．

一方，近年では様々なセンサー情報を用いたマルチモーダル・インターフェースのロボット知覚センサへの応用が盛んに行われている．我々は過去に，ロボット視覚より得られた話者方位を ICA の初期値推定に用いることで，従来は不可能であった対話ロボットの初期応答時の雑音抑圧精度の低下を防ぐ手法の提案を行い，その有効性を示している [5]．このように，音声のみならず，ロボット周囲の様々な周辺情報を活用することで，従来より高精度な雑音抑圧手法の実現が期待できる．Kinect への BSSA の実装を行うことで，同一インターフェース内で取得可能な人体の動きなど，マイクロホンアレー以外の情報を多分に活用した，より高精度な雑音抑圧システムの実現が期待できる．また，デバイス自体が小型であるため，ロボット聴覚への応用も期待できる．そこで，本稿では，ロボット視覚情報を応用したロボット聴覚インターフェースを構築することを目的とし，まず，その第一段階として Kinect のマイクロホンアレーへのリアルタイム BSSA [6] の実装を行う．

また，実装したシステムの有効性を確認するため，雑音環境下における実環境音声認識実験を行い，実験結果についての考察を通してその有効性について検討する．

2 ブラインド空間的サブトラクションアレー [4]

2.1 概要

ICA は拡散性雑音が存在する環境下において，点音源で近似される目的の音声信号を推定するよりも，拡散性の雑音信号を推定する方が優れた推定精度を示すことが知られている．そこで，高精度に目的音声を抽出する手法として BSSA が提案されている．BSSA における処理の流れを図 1 に示す，BSSA ではマイクロホンアレーに入力された信号は以下のように処理される．

- DS により目的音声スペクトル $y_{\mathrm{DS}}(f,\ )$ を強調する (主パス).
- ICA により雑音信号スペクトル $z(f,\ )$ を推定する (参照パス).
- 主パスの出力から参照パスの出力を SS で減算し，目的音を強調する．

詳細な信号処理については以下で説明する．

2.2 主パスでの目的音強調

本研究での受音系は，直線上に配置されたマイクロホンアレーである．マイクロホンアレーで観測されるマルチチャネル信号 $\boldsymbol{x}(t) = [x_1(t), \ldots, x_J(t)]^T$ に対して短時間離散フーリエ変換を行うと，以下のような時間周波数領域信号 $\boldsymbol{x}(f,\ )$ が得られる．

$$\boldsymbol{x}(f,\ ) = \boldsymbol{h}(f)s(f,\ ) + n(f,\ ) \tag{1}$$

ここで，f は周波数ビンを，　は時間フレームインデックスを表す．

主パスにおける目的音響長は DS に基づいて行われる．DS により目的音を強調した主パス出力 $y_{\mathrm{DS}}(f,\ )$ は以下のように表される．

$$y_{\mathrm{DS}}(f,\ ) = \boldsymbol{g}_{\mathrm{DS}}(f)^T \boldsymbol{x}(f,\ ) \tag{2}$$

$$\boldsymbol{g}_{\mathrm{DS}}(f) = \left[g_1^{(\mathrm{DS})}(f), \ldots, g_J^{(\mathrm{DS})}(f) \right]^T \tag{3}$$

$$g_j^{(\mathrm{DS})}(f) = \frac{1}{J} \exp\left(-i2\ \ \frac{f}{M} f_s d_j \frac{\sin\theta_U}{c} \right) \tag{4}$$

ここで，$\boldsymbol{g}_{\mathrm{DS}}(f)$ は DS のフィルタ係数ベクトル，θ_U は目的音方位，f_s はサンプリング周波数，$d_j (j = 1, \ldots, J)$ はマイクロホン位置を示す．また，M は DFT 点数，c は音速である．

2.3 参照パスでの雑音推定

参照パスでは，ICA により雑音を推定する．ICA は目的音信号と推定雑音信号が互いに独立となるように，分離フィルタの最適化を行う．ICA による観測信号の分離処理は以下のように表現される．

$$\boldsymbol{o}(f,\ ) = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\ ) \tag{5}$$

ここで $\boldsymbol{o}(f,\ ) = [o_1(f,\ ), \ldots, o_K(f,\ )]^T$ は分離信号ベクトル，K は出力音源数，$\boldsymbol{W}_{\mathrm{ICA}}(f)$ は分離行列を表している．

また，ICA に基づく分離フィルタ $\boldsymbol{W}_{\mathrm{ICA}}(f)$ は以下の更新式に基づいて最適化される．

$$\begin{aligned}\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) =& \mu \left[\boldsymbol{I} - \left\langle \boldsymbol{\varphi}\left(\boldsymbol{o}(f,\ )\right) \boldsymbol{o}^H(f,\ ) \right\rangle_\tau \right] \cdot \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) \\ &+ \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f)\end{aligned} \tag{6}$$

ここで p は更新回数，μ は更新係数，$\boldsymbol{M}^H$ は行列 $\boldsymbol{M}$ の複素共役転置，$\boldsymbol{I}$ は単位行列，$\langle\rangle_\tau$ は時間平均，$\varphi(\cdot)$ は非線形関数ベクトルを表している．

参照パスでは雑音の推定を行うため，分離信号ベクトルから，目的音推定信号 $o_U(f,\ )$ を以下のように取り除いた信号ベクトル $\boldsymbol{q}(f,\ )$ を得る．

$$\begin{aligned}\boldsymbol{q}(f,\ ) = \big[& o_1(f,\ ), \ldots, o_{U\ \ 1}(f,\ ), 0, \\ & o_{U+1}(f,\ ), \ldots, o_K(f,\ ) \big]^T\end{aligned} \tag{7}$$

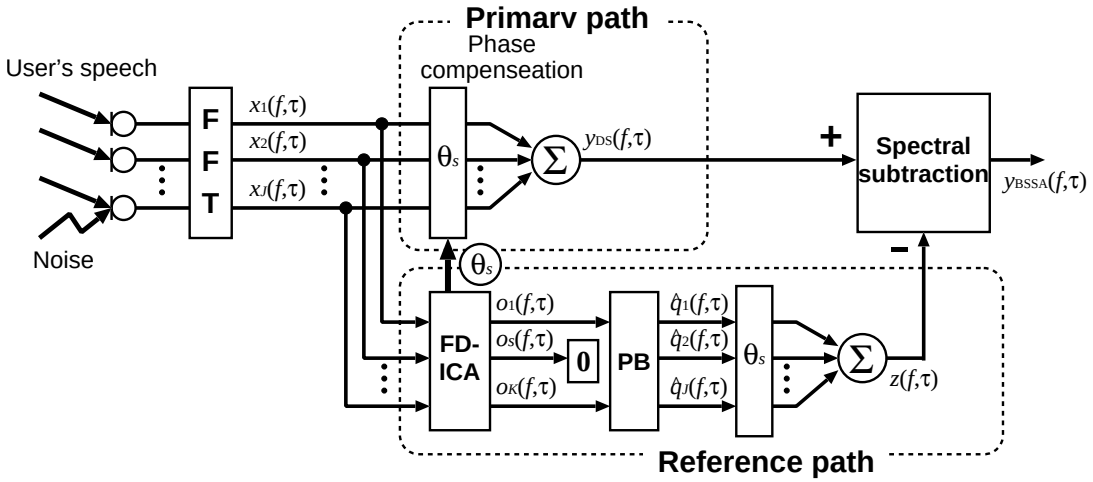


図 1: Block diagram of BSSA.

次に射影法 (Projection Back : PB) によって，利得の正規化を行う，この処理は以下の式によって与えられる．

$$\hat{\boldsymbol{q}}(f,\ ) = \boldsymbol{W}^{+}_{\mathrm{ICA}}(f)\boldsymbol{q}(f,\ ) \tag{8}$$

ここで，$\boldsymbol{M}^{+}$ は行列 $\boldsymbol{M}$ の Moore-Penrose 型一般逆行列を表す．最後に，下式のように，主パスと同様に DS を適用し，推定雑音 $z(f,\ )$ を得る．

$$z(f,\ ) = \boldsymbol{g}^{T}_{\mathrm{DS}}(f)\hat{\boldsymbol{q}}(f,\ ) \tag{9}$$

2.4 雑音抑圧処理部

最後に，雑音抑圧がスペクトル領域における減算によって行われ，出力 $y_{\mathrm{BSSA}}(f,\ )$ を得る．これは以下のように表現される．

$$y_{\mathrm{BSSA}}(f,\ ) = \begin{cases} \sqrt[n]{|y_{\mathrm{DS}}(f,\ )|^n - \beta\cdot|z_{\mathrm{ICA}}(f,\ )|^n} \\ \qquad \text{if} \left|y_{\mathrm{DS}}(f,\ )\right|^n - \beta\cdot\left|z_{\mathrm{ICA}}(f,\ )\right|^n \geq 0 \\ \gamma\cdot y_{\mathrm{DS}}(f,\ ) \qquad (\text{otherwise}) \end{cases} \tag{10}$$

ここで，SS の指数乗ドメインを表す．

この減算処理は，式 (10) 中の条件によって，二つの処理に分岐する．もしも，スペクトル上での減算結果が正の値を持つ場合は，$y_{\mathrm{BSSA}}(f,\ )$ はスペクトル減算係数 β の関数となる．ここで β は通常 1 より大きな値に設定され，推定雑音スペクトルを多めに減算 (オーバーサブトラクション) することにより，頑健な雑音抑圧処理を実現している．一方，スペクトル領域上での減算結果が負の値を持つ場合，小さな正の値を持つ γ によりフロアリングが行われる．一般に音声認識のデコーダは位相情報にそれほど敏感ではないため，スペクトル上で雑音抑圧処理を行う BSSA は音声認識に有効である．

2.5 リアルタイムアルゴリズム

BSSA において， DS や SS の処理はリアルタイムに動作させることが可能であるが， ICA によって雑音推定フィルタを最適化する部分については計算量が多いため，リアルタイムに動作させることが困難である．そこで， ICA 部分についてはリアルタイムに分離フィルタを更新するのではなく，過去のある時間区間のデータで学習した分離フィルタを，次の時間区間に適用させる．具体的な処理の流れを図 2 に示す．また，入力された信号は以下の手順で処理される．

[STEP 1] 入力信号をフレーム毎に高速フーリエ変換 (Fast Fourier Transform : FFT) を用いて時間周波数信号に変換する．

[STEP 2] ICA による分離フィルタの最適化部分は過去の 1.5 秒の入力信号データを用い，次の 1.5 秒の間には分離フィルタの更新を行う．この分離フィルタは，さらに次の 1.5 秒のための分離フィルタとして用いられる．これは， ICA の分離フィルタの学習には非常に多くの計算量が必要で，学習中のデータに最適化された雑音推定フィルタを，そのデータ自身に適用することが困難なためである．

[STEP 3] STEP 2 における ICA の学習と平行して，入力信号を BSSA の主パスと参照パスに分けて処理を行う．主パスでは DS を用いて目的音を強調する．参照パスでは，過去のデータから ICA により更新された雑音推定フィルタを基に雑音信号を推定する．

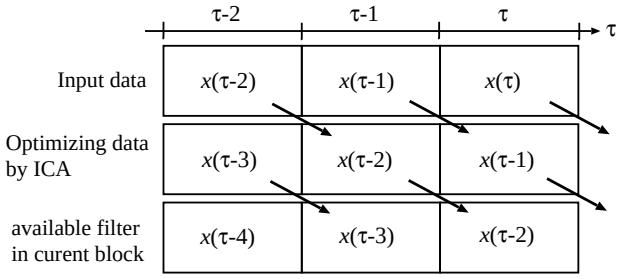


図 2: Configuration of updating separation filter in BSSA.

[STEP 4] STEP 3 より得られた，主パスの出力から，参照パスの出力 (推定雑音) をスペクトル減算することにより目的音を強調した信号を得る．

この処理では ICA により最適化された雑音推定フィルタの更新はリアルタイムではなく 1.5 秒毎に行われるが， DS や分離フィルタによるフィルタリング，スペクトル減算はリアルタイムで動作するため，システム全体ではリアルタイムで動作しているように見える．

3 Kinect を用いたリアルタイム雑音抑圧処理システムの提案

3.1 Kinect の概要

本稿では，ロボット聴覚情報を応用したロボット聴覚インターフェースを構築することを目的とし，まず，その第一段階として Kinect のマイクロホンアレーへのリアルタイム BSSA [6] の実装を行う．まず，Kinect の概要について述べる．Microsoft Kinect (Kinect) [7] は，モーションキャプチャや音声認識機能を，同社のコンシューマ向ゲーム機器である Xbox 360 に付加するために開発されたマルチモーダル · インターフェースであり，本稿で使用するマイクロホンアレーのほか，RGB カメラ，深度センサなどの Kinect 周囲における周辺情報を取得するためのセンサ群が搭載されている．

また，マイクロホンアレーに限れば，Kinect 内部にはこの出力信号を処理する機構は設けられて居らず，出力信号は Universal serial bus (USB) 経由で外部デバイスへと転送される仕組みとなっている．そのため，USB 経由で Kinect のマイクロホンアレー出力信号を PC 上で取得することができ，マイクロホンアレー出力信号を用いた自由なプログラミングが可能であることが特徴となっている．

現在，Microsoft は， Kinect を Windows 上で動作させるための統合開発環境である Kinect for Windows SDK [8] を一般公開しており，Windows 上で Kinect のセンサ情報を用いたアプリケーションの開発が可能となっている．本 SDK をインストールすることで，Kinect が接続された Windows PC 上では，Kinect のマイクロホンアレーが 4 チャネル入力の USB オーディオデバイスとして認識され，一般のオーディオ · アプリケーション · プログラム · インタフェース (API) でプログラムを記述することができるようになる．

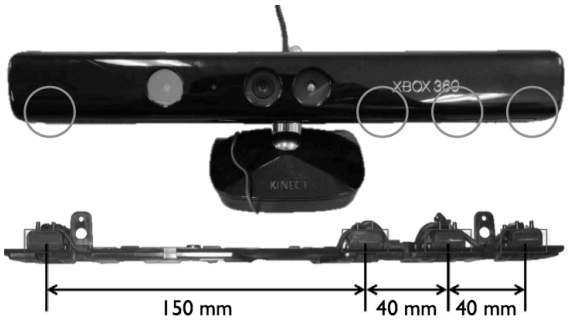


図 3: Microphone array of Kinect.

3.2 Kinect のマイクロホンアレーとその内部構造

図 3 に，Kinect に搭載されいてるマイクロホンアレーを示す．Kinect のマイクロホンアレーは， 4 つの単一指向性マイクロホン Ringford Products 製 CZ034GU により構成されており，各素子は一直線上にの不当間隔で並べられている．Kinect を正面から見たときに，右側に間隔が 40 mm で 3 つのマイクロホンが，左側に間隔が 150 mm で 1 つのマイクロホンが配置されている．

図 4 に，Kinect におけるマイクロホンアレー出力信号が USB に出力されるまでの内部構成のブロック図を示す．マイクロホンアレーからの出力信号は，アンバランス伝送で 2 つの 2 チャネル · プリアンプ内蔵型 A/D コンバータにそれぞれ入力される．その後，I2C 経由でオーディオストリームコントローラに信号が渡され，各 2 ch の出力信号が サンプリング周波数 16 kHz, 分解能 16bit の 4 ch オーディオデータにパッキングされ，USB 経由で出力される．

3.3 実装

Kinect でロボット聴覚に用いることのできる雑音抑圧処理を行うため，Kinect の後段に USB 経由で PC を接続し，Microsoft Visual C++ 2010 を用いて PC 上にリアルタイム BSSA による雑音抑圧処理システムの実装を行う．実装に用いた PC は Intel 製 Core i7 1.86 GHz の CPU と 8 GB のメモリを備え，OS は

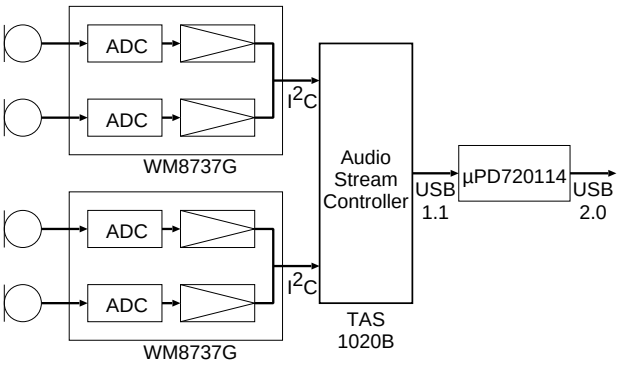


図 4: Block diagram of the microphone array input of Kinect.

Windows 7 Ultimate とする．提案システムは平均して約 40 MBytes のメモリを使用する．また，提案システムでは，Kinect より入力された信号は，本稿の第 2 章にて述べたリアルタイム BSSA の処理による雑音抑圧処理が行われたあと，リサンプラを介して任意のオーディオデバイスへ出力できるようにシステムの構成を行った．今回の実装では，Kinect の制約により入力信号のサンプリング周波数は 16 kHz，量子化ビット数は 16 bits となる．また，STFT のフレーム長は 512 点，シフトサイズは 128 点とする，ICA による分離フィルタ更新のための信号分析窓長は 256 点とする．

4 実環境における評価実験

4.1 実環境雑音の模擬

公共の場における実環境音声認識実験は困難であるため，実験室内に実環境を模擬した拡散性雑音環境を構築する．実際の駅で単一指向性マイクロホン 8 本で収録した雑音を，実験室に設置した 8 個のラウドスピーカーから再生し，駅の雑音環境を模擬する．収録された雑音には，駅の背景雑音や電車走行音をはじめ，発券機，自動改札機，車，人の足音，風などの駅における様々な雑音を含んでおり，非定常な雑音となっている．

4.2 実験条件

Kinect 上に構築したリアルタイム BSSA のロボット聴覚としての有効性について検討を行うため，実環境で音声認識実験を行った．図 5 に実験に使用した環境を示す．実験に用いた目的音は， 46 話者による 200 文を読み上げたもので， Kinect の正面 1.0 m の位置に設置したスピーカーから再生される．各スピーカーと Kinect は高さ 1.2 m の位置に設置する．床から天井までの高さは 2.7 m とする．雑音は， Kinect と目

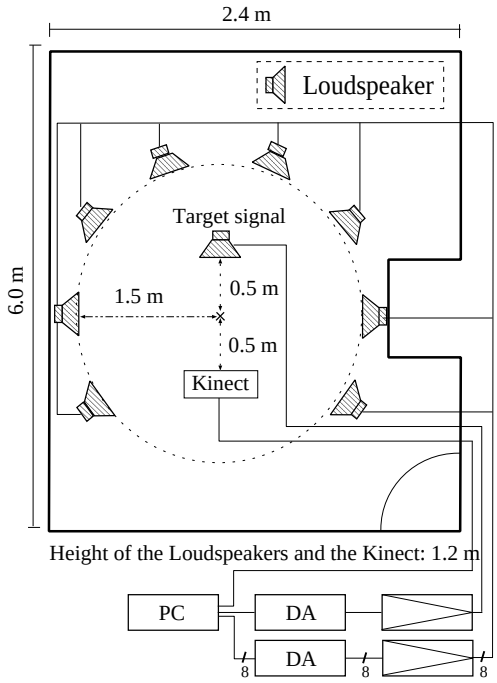


図 5: Acoustical environment used in real-world experiment.

的音を再生するスピーカの周囲を取り囲むよう半径 1.5 m の円上に設置した， 8 個のラウドスピーカーから実収録された駅雑音を再生する．

あらかじめ騒音計を用いて Kinect の設置位置にて，目的音の音圧は 65 dBA に，駅雑音の音圧は，目的音声の音圧との SN 比が平均 5 dB, 10 dB, 15 dB となるよう音量を調整してから実験を行う．

この環境において，提案システムによる雑音抑圧処理を行った場合と，雑音抑圧処理を行わなかった場合の音声を収録した．収録した音声を音声認識器にかけ，音声認識を行い，雑音抑圧処理前と処理後の収録音声で単語正解率と単語正解精度の比較を行った．音声認識実験の詳細条件を表 1 に示す．

BSSA の主パスである DS 部分では Kinect マイクロホンアレーの 4 チャネル分すべての信号を使用し，参照パスの ICA は中央 2 チャネルの出力を用いる．雑音抑圧処理部の SS では，指数乗のドメインは 2.0 乗，減算係数 β は 1.4 ，フロアリング係数 γ は 0.2 を用いて評価を行う．

4.3 実験結果

図 6 に音声認識実験の結果を示す．(a) に単語正解率，(b) に単語正解精度を示す．図 6 より，単語正解率，単語正解精度共に無処理の場合と比べて，本実験環境下では 10% 以上の精度の改善が見られることが

表 1: Experimental conditions for speech recognition.

テストデータ	JNAS [10] テストセット 男女 46 話者 200 文
音声認識タスク 音響モデル	新聞記事読み上げ 語彙数：20k
音響モデルの 学習データ	JNAS 260 話者 1 話者あたり 150 文
認識デコーダ	Julius ver. 4.2 [9]

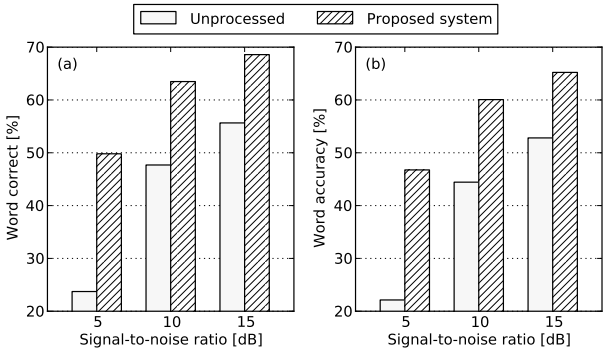


図 6: Result of speech recognition test in real-world experiment. (a) word correct, and (b) word accuracy.

わかる．先攻研究 [6] によると， 2 cm 間隔 4 チャネルのマイクロホンアレーを用いたシミュレーション実験と， 2.1 cm 間隔 8 チャネルの実環境音声認識実験の結果は，今回の結果とほぼ同等の結果を示しており，これらのシステムと比較し，遜色のない性能を示す提案システムは，実環境においても有効であるといえる．したがって，提案システムによる雑音抑圧処理は有効であるといえる．

5 おわりに

本稿では，ロボット視覚情報を応用したロボット聴覚インターフェースを構築することを目的とし，まず，その第一段階として，マイクロホンアレーやモーションセンサなどを内蔵したマルチモーダル・インターフェースである Kinect のマイクロホンアレーへのリアルタイム BSSA [6] の実装を行った．また，実装したシステムを用いて，実環境における音声認識実験によるシステムの評価を行った．実験結果より，提案システムを用いることで，雑音環境下において音声認識率が約 10% 以上改善される事を確認した．

今後は Kinect のマイクロホンアレーだけでなく，モーションセンサの情報を用いてロボットに話しかけてきた話者の動的位置検出を行い，ICA のフィルタ生成を補助するマルチモーダルなインタフェースとしての可能性を検討していく予定である．

謝辞

本研究の一部は，科学技術振興機構・戦略的創造研究推進事業 (CREST) の支援を受けた．

参考文献

[1] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” Journal of the Acoustical Society of America, vol.78, no.5, pp.1508-1518, 1985.

[2] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, “Microphone array based speech recognition with difference talker-array positions,” ICASSP ’97, pp.227-230, 1997.

[3] H. F. Silverman, and W. R. Pattterson, “Visualizing the performance of large-aperture microphone arrays,” ICASSP ’99, pp.962-972, 1999.

[4] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, K. Shikano, “Blind spatial subtraction array fort speech enhancement in noisy environment,” *IEEE Trans. Audio, Speech, and lang. Process.*, vol.17, no.4, pp.650-664, 2009.

[5] H. Saruwatari, N. Hirata, T. Hatta, R. Wakisaka, K. Shikano, T. Takatani, “Semi-Blind Speech Extraction for Robot Using Visual Information and Noise Statistics,” Proc. of the 11th IEEE ISSPIT2011, 2011.

[6] 高橋 祐, 猿渡 洋, 鹿野清宏, “独立成分分析を導入した空間的サブトラクションアレーによるハンズフリー音声認識システムの開発,” 電子情報通信学会論文誌. D, vol.93, no.3, pp.312-325, 2010.

[7] Microsoft, “Kinect - Xbox.com,” http://www.xbox.com/ja-JP/kinect

[8] Microsoft, “Microsoft Kinect SDK for Developers| Develop for the Kinect | Kinect for Windows,” http://kinectforwindows.org/

[9] Julius development team, “大語彙連続音声認識エンジン julius,” http://julius.sourceforge.jp/

[10] 音声資源コンソーシアム, http://research.nii.ac.jp/src/index.html

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-B102-9

ブラインド音源分離のための Infinite Sparse Factor Analysis の複素拡張

Complex Extension of Infinite Sparse Factor Analysis for Blind Source Separation of Speech Signals

柳楽浩平 **Kohei NAGIRA**　**高橋徹** **Toru TAKAHASHI**　**尾形哲也** **Tetsuya OGATA**　**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
{knagira, tall, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

We present a method of blind source separation (BSS) for speech signals using a complex extension of infinite sparse factor analysis (ISFA) in the frequency domain. In real environment, microphone array embedded in robot captures sound mixture contaminated by delayed signals (i.e. reflections, short-time reverberations, and time lags of signals arriving at microphones). Our method achieves robust separation of sound mixture that contains such delayed signals. Our method uses complex normal distributions to estimate source signals and mixing matrix. Experimental results indicate that our method outperforms the conventional ISFA and in the average signal-to-distortion ratio (SDR).

1 はじめに

音声信号のブラインド音源分離は遠距離音声認識[Wölfel and McDonough, 2009; Seltzer *et al.*, 2004]やロボット聴覚システム[Nakadai *et al.*, 2010; Valin *et al.*, 2004] などの様々な領域で応用されており，それゆえに活発な研究トピックの一つとなっている．実環境においては，マイクからのシステムへの入力信号は複数話者の混合音声となり，さらに反射音や残響なども同時に入力される．このような混合信号からそれぞれの話者の音声を認識するために，混合音を分離する必要がある．

音声信号の音源分離に対する主な要求条件は以下の通りである．

要件 1. 事前情報を用いない分離

要件 2. アクティビティの同時推定

要件 3. 時間遅れ信号に対する頑健性

音源位置やマイク配置などの事前情報を用いない音源分離はブラインド音源分離[Belouchrani *et al.*, 1997]と呼ばれる．独立成分分析 (Independent component analysis: ICA) [Hyvärinen *et al.*, 2001] はブラインド音源分離によく利用される手法である．実環境下でのブラインド音源分離を達成する手法としてよく用いられるものに周波数領域の ICA[Sawada *et al.*, 2002]があるが，各音源のアクティビティの推定は行わないため，要件 2 を満たさない．

Infinite sparse factor analysis (ISFA) [Knowles and Ghahramani, 2007] はノンパラメトリックベイズに基づいたブラインド音源分離手法である．ISFA は音源分離と音源のアクティビティの同時推定を行うため，要件 1，2 を満たす．しかしながら従来の ISFA では反射音や残響，各信号のマイクへの到来時間差などの時間遅れ信号を含んだ混合音声をモデル化しておらず分離が困難であるため，要件 3 を満たさない．

我々の研究の目的は以上 3 つの要求を満たすブラインド音源分離システムの開発である．本稿では ISFA の複素拡張を用いて，これらの要求条件を満たす BSS 手法を提案する．

2 ISFA によるブラインド音源分離

本章では本稿で取り上げる問題を明らかにし，従来の ISFA について説明したのち，解決すべき問題点について述べる．

2.1 ブラインド音源分離の問題設定

本稿で扱うブラインド音源分離問題を要約すると以下のようになる．

入力: D 本のマイクに入力される K 音源の混合信号
出力: 元の K 個の音源信号とそれらのアクティビティ
仮定: $K \leq D$
残響時間は短時間フーリエ変換 (Short time Fourier transform: STFT) の窓幅より短い

D 個のマイクを用いてシステムに K 個の音源からの混合信号を入力し，音源方向やインパルス応答などの事前情報を用いずに元の K 個の音源信号を分離して出力する．

2.2 音声信号のブラインド音源分離

ここで，音声信号がマイクに入力される際の音声の混合過程について述べる．音源とマイクの間には距離があるので，音源から生じた音はすぐにマイクに届くのではなく，距離の分だけ遅延して入力される．また，壁などで反射した後にマイクに届く音などの間接音も同時に入力される．つまり，複数音源からの音声が混合される際，各音源からの音声信号それぞれの直接音及び間接音が同時に入力されることになる．このような混合過程は次の式のような時間領域での畳み込み混合モデルとして定式化できる．

$$\overline{\mathbf{x}}(t) = \sum_{j=0}^{J} \overline{\mathbf{A}}(j)\overline{\mathbf{s}}(t-j) \tag{1}$$

ここで，t は時刻を表し，$\overline{\mathbf{x}}(t)$，$\overline{\mathbf{s}}(t)$，$\overline{\mathbf{A}}(j)$ はそれぞれ観測信号，音源信号，伝達関数を表す．J は残響時間を意味しており，本稿では J が STFT の窓幅よりも短いことを仮定している．無響室などでの混合音声の場合この仮定が満たされるが，一般的な部屋での混合信号は必ずしもこれを満たさない．

畳み込み混合信号のブラインド音源分離問題を解く際には STFT がよく利用される．STFT により，式 (1) は以下のような式に変換される．

$$\mathbf{x}(f,t) = \mathbf{A}(f,t)\mathbf{s}(f,t) \tag{2}$$

f は周波数帯域のインデックスである．つまり，時間領域での畳み込み混合が周波数領域の瞬時混合に変換できる．この変換で，変換前は実数信号であったのに対し、変換後では複素信号を扱う必要が生じる．STFT を施したのちに，各周波数ごとに独立に分離処理を行い，分離結果に対して逆 STFT を施し元の音声信号を復元する．

2.3 従来の ISFA

Infinite sparse factor analysis [Knowles and Ghahramani, 2007]はノンパラメトリックベイズに基づくブラインド音源分離手法である．ここでは ISFA のモデルについて述べる．

はじめに，ISFA の混合モデルについて説明する．K，D，N をそれぞれ音源数，マイクの数，音源信号の長さとする．瞬時混合モデルは以下のように表される．

$$\mathbf{X} = \mathbf{A}(\mathbf{Z} \odot \mathbf{S}) + \mathbf{E}, \tag{3}$$

ここで，$\mathbf{Z} = [\mathbf{z}_1, \cdots \mathbf{z}_N]$，$\mathbf{X} = [\mathbf{x}_1, \cdots \mathbf{x}_N]$，$\mathbf{S} = [\mathbf{s}_1, \cdots \mathbf{s}_N]$, $\mathbf{E} = [\varepsilon_1, \cdots \varepsilon_N]$, $\mathbf{x}_t = [x_{1t}, x_{2t}, \cdots, x_{Dt}]^{\mathrm{T}}$ は時刻 t での混合信号ベクトル，$\mathbf{s}_t = [s_{1t}, s_{2t}, \cdots, s_{Kt}]^{\mathrm{T}}$ は音源信号ベクトル，$\varepsilon_t = [\varepsilon_{1t}, \varepsilon_{2t}, \cdots, \varepsilon_{Dt}]^{\mathrm{T}}$ はガウス性雑音のベクトルとする．また，$\mathbf{A}$ は $D \times K$ の混合行列，$\mathbf{z}_t = [z_{1t}, z_{2t}, \cdots, z_{Kt}]^{\mathrm{T}}$ は時刻 t での各音源のアクティビティを表す．z_{kt} 二値の変数であり，音源 k が時刻 t で音が鳴っている場合は $z_{kt} = 1$ となり，そうでない場合は $z_{kt} = 0$ となる．演算子 $\odot$ は要素ごとの積を表している．ISFA は観測信号 $\mathbf{X}$ のみを用いて音源信号 $\mathbf{S}$ とそれらのアクティビティ$\mathbf{Z}$，混合行列 $\mathbf{A}$，その他のパラメータを同時に推定する．

2.4 従来法の問題点

従来の ISFA[Knowles and Ghahramani, 2007]では複素数を扱えないため，STFT によって得られる混合音声の複素スペクトルに対して従来の ISFA を適用できず，畳み込み混合信号の分離ができない．これは音声信号のブラインド音源分離を行うにあたって解決すべき主要な問題の一つである．なぜなら，上記の通り音声信号の混合過程は伝達関数の畳み込みを用いて表されるからである．

3 ISFA の複素拡張

時間遅れ信号を含んだ混合信号を分離するために，周波数領域で ISFA を用いることを考える．我々の従来手法[柳楽ら, 2011]では，入力信号の実部と虚部を別々に実数 ISFA に入力していたが，実部と虚部の統合の際に別音源の実部と虚部が統合される可能性があり，推定精度が低下するという問題点があった．本稿では ISFA 自身を複素信号を扱えるように拡張することで周波数領域での ISFA を実現する．

Table 1 は本手法の推論アルゴリズムである．本手法は Metropolis-Hastings アルゴリズムと Gibbs サンプリングに基づいている．ベイズの定理から，潜在変数の事後分布は事前分布と尤度関数の積から得られる．以下では，各パラメータの事前分布とこのモデルの尤度関数を示し，それぞれの事後分布について述べる．

3.1 事前分布

各変数の事前分布は以下の通りである．

$$\varepsilon_t \sim \mathcal{N}_C(0, \sigma_\varepsilon^2 \mathbf{I}) \qquad \sigma_\varepsilon^2 \sim \mathcal{IG}(p_1, p_2), \tag{4}$$

$$s_{kt} \sim \mathcal{N}_C(0, 1), \tag{5}$$

$$\mathbf{a}_k \sim \mathcal{N}_C(0, \sigma_{\mathbf{A}}^2 \mathbf{I}) \qquad \sigma_{\mathbf{A}}^2 \sim \mathcal{IG}(p_3, p_4), \tag{6}$$

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha) \qquad \alpha \sim \mathcal{G}(p_5, p_6). \tag{7}$$

ここで，$\mathbf{a}_k$ は $\mathbf{A}$ の k 番目の列，p_1，p_2，p_3，p_4，p_5，p_6 はハイパーパラメータである．$\mathcal{N}_C(\mu, \sigma^2)$ は平均 μ，分散 σ^2 の一変量複素正規分布を表す．$\mathcal{G}(b, \theta)$ と $\mathcal{IG}(b, \theta)$ は形状母数 b，尺度母数 θ のガンマ分布と逆ガンマ分布を表す．それぞれの分布の確率密度関数は以下のようになって

Table 1: Algorithm for estimating model parameters of complex ISFA

1. 混合行列 $\mathbf{A}$，音源のアクティビティ $\mathbf{Z}$，音源信号 $\mathbf{S}$ を事前分布を元に初期化
2. 各時刻 t について以下を実行
   - 2-1 各音源 k ごとに式 (17) をもとに z_{kt} をサンプル
   - 2-2 $z_{kt}=1$ なら式 (13) から s_{kt} をサンプルそうでない場合は $s_{kt}=0$
   - 2-3 この時刻で初めて active になる音源の数 κ_t を決め，初期化
3. 各音源 k ごとに混合行列 $\mathbf{a}_k$ を式 (21) からサンプル
4. 全時刻通して inactive になっている音源があれば除去
5. σ_ε^2，$\sigma_{\mathbf{A}}^2$，α を式 (22)，(23)，(24) をもとに更新
6. 2 へ戻る

いる．

$$\mathcal{N}_C(x;\mu,\sigma^2) = \frac{1}{\pi\sigma^2}\exp\left(-\frac{|x-\mu|^2}{\sigma^2}\right), \quad (8)$$

$$\mathcal{G}(x;b,\theta) = \frac{x^{b-1}}{\Gamma(b)\,\theta^b}\exp\left(-\frac{x}{\theta}\right), \quad (9)$$

$$\mathcal{IG}(x;b,\theta) = \frac{x^{-(b-1)}}{\Gamma(b)\,\theta^b}\exp\left(-\frac{1}{\theta x}\right). \quad (10)$$

IBP(α) はパラメータ α の Indian buffet process (IBP) [Griffiths and Ghahramani, 2006] を表す．IBP は潜在的に無限個の音源を扱うことができる確率過程である．IBP の概要は以下のように表される．

1. 時刻 $t=1$ において
   初めから鳴っている音源の数を Poisson(α) からサンプリングする．
2. 時刻 $t=i$ において
   - 音源 k は確率 $\frac{m_k}{i}$ で active になる．ここで m_k は時刻 $t=1$ から $i-1$ までで音源 k が active になった時間の数を表す．
   - 既存の音源が active かどうかを決定した後，時刻 i で始めて active になる音源の数を Poisson($\frac{\alpha}{i}$) からサンプリングする．

α は母数 α のポアソン分布を表す．IBP にはサンプル順序の交換可能性があり，注目している時刻 t が最後にサンプルされると考えてよい．つまり，時刻 t 以外のアクティビティが与えられた状態で時刻 t のアクティビティを推定できる．これより，IBP に基づくアクティビティの事前分布は以下のようになる．

$$P(z_{kt}|\mathbf{z_{-kt}}) = \frac{m_{k,-t}}{N} \quad (11)$$

ただし，$m_{k,-t}=\sum_{s\neq t} z_{ks}$ を表し，$\mathbf{z}_{-kt}$ は $\mathbf{z}_t$ の要素のうち z_{kt} を取り除いたものを表す．

3.2 尤度関数

複素 ISFA の尤度関数は以下のように表される．

$$\begin{aligned} P(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{Z}) &= \prod_{t=1}^{N} P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_t,\mathbf{z}_t) \\ &= \prod_{t=1}^{N} \mathcal{N}_C(\mathbf{x}_t;\mathbf{A}(\mathbf{z}_t\odot\mathbf{s}_t),\sigma_\varepsilon^2\mathbf{I}) \\ &= \frac{1}{(\pi\sigma_\varepsilon^2)^{ND}}\exp\left(-\frac{\mathrm{tr}(\mathbf{E}^{\mathrm{H}}\mathbf{E})}{\sigma_\varepsilon^2}\right), \end{aligned} \quad (12)$$

ここで，

$$\mathbf{E}=\mathbf{X}-\mathbf{A}(\mathbf{Z}\odot\mathbf{S})$$

であり，各時刻でのデータは独立同分布であると仮定している．

3.3 事後分布

ここではこれまでに示した事前分布と尤度関数を用いて，ベイズの定理に基づいた事後分布の推論について述べる．ここで推論された事後分布からのサンプリングによって分離信号，各信号のアクティビティ，混合行列の推定を行う．

3.3.1 音源信号

z_{kt} が active であるとき，s_{kt} の事後分布は式 (5) と尤度関数から以下のようになる．

$$\begin{aligned} P(s_{kt}|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t\mathbf{z}_t) &\propto P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_t,\mathbf{z}_t,\sigma_\varepsilon^2)P(s_{kt}) \\ &= \mathcal{N}_C\left(s_{kt};\mu_s,\sigma_s^2\right), \end{aligned} \quad (13)$$

ここで，

$$\sigma_s^2=\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2+\mathbf{a}_k^{\mathrm{H}}\mathbf{a}_k},\quad \mu_s=\frac{\mathbf{a}_k^{\mathrm{H}}\varepsilon_{-kt}}{\sigma_\varepsilon^2+\mathbf{a}_k^{\mathrm{H}}\mathbf{a}_k}$$

である．$\mathbf{s}_{-kt}$ は $\mathbf{s}_t$ から s_{kt} を除いたもの，ε_{-kt} は $\varepsilon|_{z_{kt}=0}$ を意味する．

3.3.2 音源のアクティビティ

z_{kt} が active になる事後確率と inactive と事後確率の比は式 (14) によって計算される．この事後確率の比 r は二つの部分に分けられ，片方は事前確率の比 r_p，もう片方は尤度の比 r_l である．

$$\begin{aligned} r &= \frac{P(z_{kt}=1|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt})}{P(z_{kt}=0|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt})} \\ &= \frac{P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt},z_{kt}=1,\sigma_\varepsilon^2)}{P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt},z_{kt}=0,\sigma_\varepsilon^2)}\frac{P(z_{kt}=1|\mathbf{z}_{kt})}{P(z_{kt}=0|\mathbf{z}_{kt})} \\ &= r_l r_p. \end{aligned} \quad (14)$$

事前確率の比 r_p は以下のように計算される．

$$r_p = \frac{P(z_{kt}=1|\mathbf{z}_{-kt})}{P(z_{kt}=0|\mathbf{z}_{-kt})} = \frac{m_{k,-t}}{N-m_{k,-t}}. \tag{15}$$

これは式 (11) の IBP に基づく音源のアクティビティの事前分布から導かれる[Griffiths and Ghahramani, 2006]．

尤度の比は式 (16) から計算される．

$$\begin{aligned} r_l &= \frac{P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt},z_{kt}=1,\sigma_\varepsilon^2)}{P(\mathbf{x}_t|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt},z_{kt}=0,\sigma_\varepsilon^2)} \\ &= \sigma^2 \exp\left(\frac{|\mu_s|^2}{\sigma_s^2}\right), \end{aligned} \tag{16}$$

これらを掛け合わせることで事後確率の比 r が得られ，$z_{kt}=1$ となる事後確率はこの比 r から計算される．

$$P(z_{kt}=1|\mathbf{A},\mathbf{s}_{-kt},\mathbf{x}_t,\mathbf{z}_{-kt}) = \frac{r}{1+r}. \tag{17}$$

z_{kt} が active かどうかを決定するために，一様分布 Uniform$(0,1)$ から u をサンプルし，それを $r/(1+r)$ と比較する．もし，$u \le r/(1+r)$ なら z_{kt} は active となり，そうでなければ active でないということになる．

3.3.3 新たに現れる音源の数

初めからは存在しておらず，時刻 t になって初めて出現する音源について考える．κ_t をそのような音源の数とすると，この κ_t は Metropolis-Hastings アルゴリズムによってサンプルされる．

まず，κ_t の事前分布は以下のようになる．

$$P(\kappa_t|\alpha) = \text{Poisson}\left(\frac{\alpha}{N}\right). \tag{18}$$

κ_t をサンプルしたのち，新しい音源とそのアク r ティビティを初期化する．

次に，この更新を受理するかどうかを決定する．現状態 ξ から，新しく κ_t 個の音源が加わった次状態 ξ^* への遷移確率 $J(\xi^*|\xi)$ は，Meeds ら[Meeds *et al.*, 2007]や Knowles ら[Knowles and Ghahramani, 2007]によると，次状態 ξ^* の事前分布と等しくなる．したがって，この遷移が採用される確率は $\min(1, r_{\xi\to\xi^*})$ となる．ただし，$r_{\xi\to\xi^*}$ は次式の通りである．

$$\begin{aligned} r_{\xi\to\xi^*} &= \frac{P(\xi^*|rest)J(\xi|\xi^*)}{P(\xi|rest)J(\xi^*|\xi)} \\ &= \frac{P(rest|\xi^*)P(\xi^*)P(\xi)}{P(rest|\xi)P(\xi)P(\xi^*)} \\ &= \frac{P(rest|\xi^*)}{P(rest|\xi)} \end{aligned} \tag{19}$$

$rest$ は ξ や ξ^* 以外のパラメータすべてをまとめたもの表す．つまり，$r_{\xi\to\xi^*}$ は更新前の状態と更新後の状態の尤度の比となる．この比を計算すると以下のようになる．

$$r_{\xi\to\xi^*} = (\det\Lambda_\xi)^{-1} \exp\left(\mu_\xi^{\mathrm{H}}\Lambda_\xi\mu_\xi\right), \tag{20}$$

ここで，

$$\Lambda_\xi = \mathbf{I} + \frac{\mathbf{A}^{*\mathrm{H}}\mathbf{A}^*}{\sigma_\varepsilon^2}, \quad \Lambda_\xi\mu_\xi = \frac{1}{\sigma_\varepsilon^2}\mathbf{A}^{*\mathrm{H}}\varepsilon_t.$$

である．また，$\mathbf{A}^*$ は $D\times\kappa_t$ の行列で，$\mathbf{A}$ の追加された部分を表している．

3.3.4 混合行列

混合行列は各列ごとに推定する．式 (6) で示した事前分布と尤度関数を用いると，事後分布は以下のようになる．

$$\begin{aligned} & P(\mathbf{a}_k|\mathbf{A}_{-k},\mathbf{S},\mathbf{X},\mathbf{Z},\sigma_\varepsilon^2,\sigma_{\mathbf{A}}^2) \\ \propto\ & P(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{Z},\sigma_\varepsilon^2)P(\mathbf{a_k}|\sigma_{\mathbf{A}}^2) \\ =\ & \mathcal{N}_C(\mathbf{a}_k;\mu_{\mathbf{A}},\Lambda_{\mathbf{A}}^{-1}), \end{aligned} \tag{21}$$

ここで，

$$\begin{aligned} \Lambda_{\mathbf{A}} &= \left(\frac{\mathbf{s}_k^{\mathrm{H}}\mathbf{s}_k}{\sigma_\varepsilon^2} + \frac{1}{\sigma_{\mathbf{A}}^2}\right)\mathbf{I}_{D\times D}, \\ \mu_{\mathbf{A}} &= \frac{\sigma_{\mathbf{A}}^2}{\mathbf{s}_k^{\mathrm{H}}\mathbf{s}_k\sigma_{\mathbf{A}}^2+\sigma_\varepsilon^2}\mathbf{E}|_{\mathbf{a}_k=0}\mathbf{s}_k \end{aligned}$$

である．

3.3.5 雑音と混合行列の分散

雑音の分散は推定された信号の雑音のレベルに，混合行列の分散は推定された信号の振幅のスケールに対応している．それぞれの事後分布は以下のようになる．

$$\begin{aligned} P(\sigma_\varepsilon^2|\mathbf{E}) &\propto P(\mathbf{E}|\sigma_\varepsilon^2)P(\sigma_\varepsilon^2|p_1,p_2) \\ &= \mathcal{IG}\left(\sigma_\varepsilon^2; p_1+ND, \tfrac{p_2}{1+p_2\,\mathrm{tr}(\mathbf{E}^{\mathrm{H}}\mathbf{E})}\right). \end{aligned} \tag{22}$$

$$\begin{aligned} P(\sigma_{\mathbf{A}}^2|\mathbf{A}) &\propto P(\mathbf{A}|\sigma_{\mathbf{A}}^2)P(\sigma_{\mathbf{A}}^2|p_3,p_4) \\ &= \mathcal{IG}\left(\sigma_{\mathbf{A}}^2; p_3+DK, \tfrac{p_4}{1+p_4\,\mathrm{tr}(\mathbf{A}^{\mathrm{H}}\mathbf{A})}\right). \end{aligned} \tag{23}$$

3.3.6 IBP のパラメータ

IBP のパラメータ α の事後分布は以下のようになる．

$$\begin{aligned} p(\alpha|\mathbf{Z}) &\propto P(\mathbf{Z}|\alpha)P(\alpha|p_5,p_6) \\ &= \mathcal{G}\left(\alpha; K_+ + p_5, \frac{p_6}{1+p_6H_N}\right). \end{aligned} \tag{24}$$

ここで，K_+ は active となっている音源の数，$H_n = \sum_{j-1}^{N}\frac{1}{j}$ は N 番目の調和級数である．

3.4 後処理

周波数領域の ICA と同様に，本手法でもパーミュテーション問題とスケーリング問題について考えなければならない．これらの問題は，本手法では各周波数帯域で独立に分離を行うために，各帯域での出力信号の振幅および出力順序を揃える必要があるというものである．

ここで，スケーリング問題は projection back [Murata *et al.*, 2001]という方法で解決する．この方法は，分離信号に対して推定された混合行列の要素をかけあわせるこ

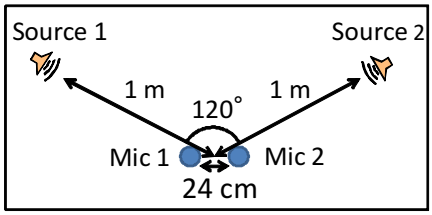


Figure 1: Locations of microphones and sources

Table 2: Experimental conditions

音源数 K	2
マイク数 D	2
サンプリング周波数	16 [kHz]
STFT 窓幅	64 [msec]
STFT シフト幅	32 [msec]

とによって，各帯域で揃っていなかった振幅をマイクに入力される信号の振幅に合わせられるというものである．

パーミュテーション問題は本稿では混合前の原信号を用いて，分離信号との相関をとることで解決する．これは ISFA の複素拡張自身の分離性能を評価するためである．このパーミュテーション問題に対する解法は Sawada ら [Sawada *et al.*, 2004]などによって提案されているが，いまだ画期的な解法は開発されていないため，この問題の解法については今もなお活発に議論されている．

4 実験結果

本手法の分離性能の評価のために音声信号を用いた分離実験を行った．まず，本手法とベースラインである実数領域の ISFA とを比較する．実験は、瞬時混合・無響室録音のインパルス応答の畳み込み混合・会議室録音のインパルス応答の畳み込み混合の 3 種類の設定を用いた．Table 2 は実験状況をまとめたもので，Fig. 1 はマイクと音源の配置を示している．ATR 音素バランス単語データベース中の 32 単語の発話を用いた．反復回数は 150 回である．

Figures 2–5 のスペクトログラムはそれぞれ元音源，入力の混合信号，本手法による分離信号，ベースラインによる分離信号を表している．SDR (Signal to Distortion Ratio)，ISR (Image to Spatial distortion Ratio)，SIR (Source to Interference Ratio)，SAR (Source to Artifacts Ratio) [Vincent *et al.*, 2007]を用いた定量的な評価も行った．

結果は Table 3 の通りである．Baseline は時間領域の ISFA を表している．時間領域の ISFA については，瞬時混合の音声の分離実験では大変よい分離性能となっているが，畳み込み混合音声の分離実験では無響室程度のかなり短い残響時間の畳み込み混合でさえほぼ分離できていないのに対し，本手法は無響室，会議室などの畳み込み混合音声でも分離可能である．

無響室環境の場合，本手法はベースラインの手法と比較して SDR の平均で 2.91[dB] の改善がみられ，会議室環境においても本手法はベースラインに勝る分離性能となることが確認された．特に畳み込み混合の音声に対しては，SIR において本手法がベースライン手法に対し大きな改善が見られる．本手法の SAR の結果がベース手法と比較して悪化しているが，これは分離の際に用いるアクティビティ推定のミスにより生じるノイズによるものであると思われる．

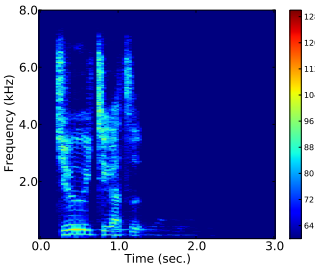


Figure 2: Spectrogram of source signal

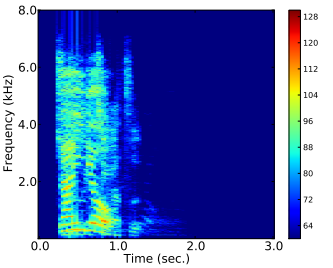


Figure 3: Spectrogram of mixed signal

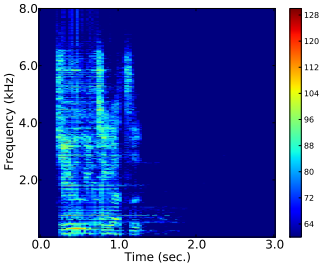


Figure 4: Separated signal with ours

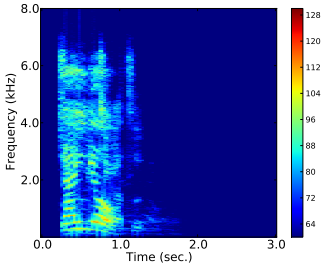


Figure 5: Baseline separated signal

また、我々の従来手法[柳楽ら, 2011]との性能比較も行った．実験条件は先ほどと同じであり，データは JNAS データベース中の 30 文を用いた．反復回数は 100 回である．その結果，無響室の場合，本手法が従来手法と比較して SDR で 0.27[dB]，SIR で 1.00[dB] の改善を確認した．

5 結論

本稿では実環境での反射音，残響，音源のマイクへの到達時間差などを考慮した畳み込み混合音声に対するブラインド音源分離と各音原のアクティビティの同時推定手法について述べた．本手法はノンパラメトリックベイズに基づいており，各周波数帯域ごとに ISFA の複素拡張を用いて複素混合信号を分離する．無響室環境での畳み込み混合音声の分離実験において，本手法によってベースラインの時間領域 ISFA と比較して平均 SDR で 2.91[dB] の改善がみられ，さらに会議室環境の畳み込み混合音声の分離実験でも分離性能の改善が見られた．また，我々の従来手法との比較においても改善が確認された．

今後の課題として，今回は音源のアクティビティについての評価を行い，これを発話区間検出（Voice Activity Detection）やパーミュテーション問題の解法に応用する事を考えている．そして，ロボット等への応用を考慮すると，リアルタイム処理を目指して本手法の処理速度の向

Table 3: Average separation performance from experimental results [dB]

	Instantaneous		
	Before	Baseline	Proposed
SDR	-1.19	**25.07**	2.27
ISR	2.35	**30.90**	4.06
SIR	1.17	**35.57**	10.45
SAR	75.77	**35.23**	2.83
	Anechoic chamber		
	Before	Baseline	Proposed
SDR	-1.01	-0.83	**2.08**
ISR	1.51	2.57	**3.86**
SIR	0.91	1.54	**8.91**
SAR	59.24	**36.25**	2.80
	Meeting room		
	Before	Baseline	Proposed
SDR	-1.96	-1.86	**0.60**
ISR	1.02	1.93	**2.98**
SIR	1.65	2.23	**4.90**
SAR	58.93	**36.08**	3.09

上について考える必要がある．

謝辞

本研究の一部は，科研費基盤 (S), JST-ANR BINAAHR, GCOE の支援を受けた．また，数多くの有益な助言をいただいた武田龍博士，平澤恭治氏に感謝の意を表する．

参考文献

[Belouchrani *et al.*, 1997] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *Signal Processing, IEEE Transactions on*, 45(2):434–444, 1997.

[Griffiths and Ghahramani, 2006] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, 18:475–482, 2006.

[Hyvärinen *et al.*, 2001] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. Wiley-Interscience, 2001.

[Knowles and Ghahramani, 2007] D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. *Independent Component Analysis and Signal Separation*, pages 381–388, 2007.

[Meeds *et al.*, 2007] E. Meeds, Z. Ghahramani, R.M. Neal, and S.T. Roweis. Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems*, 19:977–984, 2007.

[Murata *et al.*, 2001] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.

[Nakadai *et al.*, 2010] 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 Speakers. *Advanced Robotics*, 24(5–6):739–761, 2010.

[Sawada *et al.*, 2002] H. Sawada, R. Mukai, S. Araki, and S. Makino. Polar coordinate based nonlinear function for frequency-domain blind source separation. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'02)*, pages 1001–1004, 2002.

[Sawada *et al.*, 2004] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Trans. on Speech and Audio Processing*, 12(5):530–538, 2004.

[Seltzer *et al.*, 2004] M.L. Seltzer, B. Raj, and R.M. Stern. Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Trans. on Speech and Audio Processing*, 12(5):489–498, 2004.

[Valin *et al.*, 2004] 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*, volume 3, pages 2123–2128. IEEE, 2004.

[Vincent *et al.*, 2007] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First stereo audio source separation evaluation campaign: data, algorithms and results. *Independent Component Analysis and Signal Separation*, pages 552–559, 2007.

[Wölfel and McDonough, 2009] M. Wölfel and J. McDonough. *Distant Speech Recognition*. Wiley, 2009.

[柳楽ら, 2011] 柳楽 浩平, 高橋 徹, 尾形 哲也, 奥乃 博. ノンパラメトリックベイズによる時間周波数領域における音声信号のブラインド音源分離. **第 *29* 回日本ロボット学会学術講演会**, 3A2–5, 2011.

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-B102-10

マルチロボットによるKinectを用いた同期合奏

Multi-robot synchronized ensemble with Kinect

糸原達彦 † **Tatsuhiko Itohara**
水本武志 † **Takeshi Mizumoto**
Angelica Lim† **Angelica Lim**
大塚琢馬 † **Takuma Otsuka**
中村圭佑 ‡ **Keisuke Nakamura**
長谷川雄二 ‡ **Yuji Hasegawa**
中臺一博 ‡ **Kazuhiro Nakadai**
尾形哲也 † **Tetsuya Ogata**
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**
{itohara, mizumoto, angelica, ohtsuka, ogata, okuno}@kuis.kyoto-u.ac.jp
‡HRI-JP **Honda Research Institute Japan**
{keisuke, yuji.hasegawa, nakadai}@jp.honda-ri.com

Abstract

New issues will arise when plural robots participate in an ensemble with human players to attain the "unification" at three aspects of music, rhythm, melody, and harmony. We assumes that every interaction should be explicitly expressed and observable among all participants, even between robots and human. In this paper, we identify the issues for the ensemble between multiple-robot and multiple-human and we report our approach to the rhythmical unification. We focus on audio-visual integration for beat-tracking by using a Kinect with its four microphones, a stereo camera, and an infrared sensor. The resulting system provides a highly accurate beat-estimation for multi-player situations even if two robots use different beat-tracking methods.

1 はじめに

近年, 人の生活にホームロボットが密接に関わるようになっている. 人とロボットの共生において, ホームロボットが自らに付属したセンサで周囲の状況を認識し, 人と同期して動作を行うことは重要な要件である. なぜなら, 人の生活環境内で活動するロボットは, ロボット周辺の環境に対して, 適応的に人と協調作業を行う必要があるからである. この要件を満たす課題の一つに音楽合奏があげられる. 音楽合奏の課題達成は, 同期にズレを許容する表現力豊かなインタラクションの実現に重要である. また, 複数のロボット環境下でのロボット同士のコミュニケーションを, コンピュータ上の電子的なやりとりでなく人が知覚できる範囲に制限することは, 人との同期タスクを達成する上で, 必要不可欠な要素となる.

音楽は, リズム, メロディー, ハーモニーの 3 要素で定義される. 我々は, これら 3 つの要素が同期したとき合奏が成立したと定義する (3.1 節参照). 従来, 人とロボットの合奏に用いられるセンサ入力は音響信号のみであった[Otsuka, 2010; Murata, 2008]. 3 要素のうち, メロディーとハーモニーの同期は, ピッチや和音構成によって決定される. これらの要素を他のモダリティから得ることは難しく, 例えば, トロンボーンやテルミンのような手の位置によりピッチコントロールする楽器であっても, 動作が緻密であるため, 視覚から正確なピッチを得ることはできない. リズムの同期はタイミングにより決定される. タイミング検出手法として, 音楽の拍時刻とテンポを推定するビートトラッキングが盛んに研究されている. しかし従来手法の多くは, ピッチや和音同様, 入力は音響情報のみであった[Goto, 2001; Hainsworth, 2003]. その一方で人の感覚器官を考えると、発音タイミングは聴覚以外でも知覚できる。例えば, バスドラムやベースの出す低周波音の発音タイミングは触覚を用いて振動を肌で感知できる. また, 楽器演奏における発音タイミングに相関のある動作や, 演奏者同士のアイコンタクトなど, 視覚情報によるタイミング知覚は演奏者同士の同期において重要であり, 研究も盛んに行われている[Fredrickson, 1994].

我々は合奏に必要なマルチモーダルセンサとして Kinect を用いる. Kinect とは, 2010 年に Microsoft 社が発売した, RGB カメラと深度センサ, 4 チャネルのマイクアレイを搭載したデバイスである. 本来はゲーム用に発売されているが, 視覚, 聴覚, 深度のマルチモーダル情報を得られる安価なデバイスとして注目されている (2.3 節参照).

本研究では, 複数ロボット, 複数人での合奏おける, リズムの同期の部分に着目したマルチモーダル情報による合奏実現を目標とする. 複数ロボット間での同期における

課題を定義し, Kinect を用いることで, この様な複雑な実験環境下において, よりロバストなビートトラッキングを達成し, 合奏の達成を向上する.

第 2 章では音楽ロボット, ビートトラッキング, Kinect の関連研究について議論し, 本稿の立場を明らかにする. 第 3 章で複数ロボットを用いた合奏における問題を定義し, それぞれの解決に関する議論を行う. 第 4 章で上記の問題の解決の一つである視聴覚統合ビートトラッキングの Kinect を用いた視覚トラッキングについて述べ, 第 5 章で視覚トラッキングの簡単な評価, 及び複数ロボット合奏についての考察を述べる.

2 従来研究

2.1 音楽ロボットに関する研究

音楽ロボットの演奏技術の発展は近年目覚しく，発音タイミング精度や音量コントロール等による演奏表現の自由度が高いものが開発されている[Solis, 2008]．また, ロボットそのものだけではなく, 制御手法をアプローチとする研究も行われている. 水本らのロボットに依存しない汎用テルミン演奏モデル[Mizumoto, 2010a]もその一つである.

共演者ロボットの実現には，いかに協調演奏をするかという課題も存在する．Weinberg らは，ロボット 2 体と人 2 名の 4 楽器でのジャムセッションを報告している[Weinberg, 2009]．用いられたロボットは pow-wow ドラム演奏の Haile とマリンバ演奏の Shimon で，Haile はパーカッショニストとの音量主体の演奏主導権の移動を，Shimon はキーボーディストの演奏の模倣演奏と, ディスプレイを用いた疑似的なアイコンタクトとを行う．Petersen らは, フルート演奏ロボットと人のサックスとの協調演奏を報告した[Petersen, 2008]．ロボットはサックスの位置に応じて演奏パターンの変更を行う.

これらの協調演奏の対象はジャムセッションであり，タイミングに関する根本的な取り組みは行われていない．一方, ビートトラッキングという拍時刻検出手法を用いることで, 音楽に同期した足踏み[Murata, 2008]や楽譜に従った協調演奏[Mizumoto, 2010b]を行うロボット実演が報告されている. 我々もリズムに焦点をおいた合奏実現を行うため, 同様にビートトラッキングを利用する.

2.2 リズム同期に関する研究

協調演奏における要素技術の一つとして, ビートトラッキングの関連研究を示す．後藤らは, 多数のエージェントによるビートの複雑さに頑健なマルチエージェント手法を報告した[Goto, 2001]. 多数のエージェントが独立に拍時刻を推定し, 信頼度に従い分裂と消滅を繰り返す. 最終的に楽曲に一致する推定値を持つエージェントだけが残るので, 正しい拍時刻及びテンポが推定できる．村田らは, STPM(Spectro-temporal Pattern Matching) によるビートトラッキングを報告している[Murata, 2008]．STPM の利点は，定常雑音に対する頑健さ, テンポ変化に対する鋭敏さ, 実時間処理に適した動作遅延が小ささである.

パーティクルフィルタのような確率的手法を用いたビートトラッキングも報告されている．入力の特徴量として, Hainsworth らは音響信号のパワー変化[Hainsworth, 2003]を，大塚らはスペクトログラムの相互相関と楽譜情報[Otsuka, 2010]を用いている．これらの手法が音響情報のみを用いるのに対し, 我々は従来研究において, 音響情報に加え, 手のストローク動作の画像情報を用いたマルチモーダルビートトラッキングを報告した. これにより, ギター演奏という, 音がまばらで拍検出が難しい状況下でのビートトラッキングが可能になった.

画像情報を利用した他のリズム同期として, 開始及び終了タイミングの取得, 演奏主導権交代があげられる。Lim らは, Hough 変換により検出されたフルートの傾き変化に応じて, 演奏開始, テンポ変化, 演奏終了キューを検知するジェスチャー認識を報告した[Lim, 2010]. Pan らは, オプティカルフローにより顔の向きの変化を取得し, これを主導権の交代をキューの 1 つとして使用した[Pan, 2010].

2.3 Kinect, 深度情報を用いた研究

従来，深度情報を得るために，ステレオカメラや TOF(Time of Flight) カメラが用いられてきた. しかし, ステレオカメラはカメラ校正や計算コストの大きさが, TOF カメラは価格や色情報との同期の難しさが問題があった. 一方, Kinect は色情報と深度情報が紐付けされた状態で取得できる上に, 安価であるという利点がある.

Kinect の色情報と深度情報の両方を利用した研究は盛んに行われており, Saenko らは物体の高精度なラベル付け[Saenko, 2011]を, Oikonomidis らは手の関節の姿勢のトラッキング[Oikonomidis, 2011]を報告している．一方で, 音響情報を同時に利用した研究は少ない. 本稿では, 視聴覚の両方の情報を同期して用いることで, 演奏タイミングに対するより高精度な同期を行う.

3 本稿におけるロボット合奏

3.1 合奏の定義

音楽合奏は, リズム, メロディー, ハーモニーの 3 要素で構成される. リズムとは発音のタイミング, 長短, 強弱, 及びその組み合わせ, メロディーとは音の高低 (ピッチ), およびその順列 (旋律), ハーモニーはメロディーの組み合わせによる和音である.

合奏の成立を 3 つの要素が同期することと定義する. リズムの同期を各演奏者の発音タイミングの時間ズレが十分に小さい状態とする. しかし実際の合奏においては, 単に発音時間が近ければいいとは限らない. 文献[Friberg, 2002]では, ジャズ楽曲においてドラムパート, ソロパート

のスウィング, つまり基準となる時間とのズレがテンポに比例する形で現れると述べられている. 同文献によると, ソロパートにおいて, 特に表拍 (1 小節を偶数個等分したときのの奇数番目の拍) において少し遅いタイミングで演奏を行うケースが多数観測された. 同様のスウィングの研究はジャズ楽曲を中心に広く行われており, このスウィングが豊かな音楽表現につながると考えられる. メロディー及びハーモニーが同期するということはピッチ, 和音構成が同期することとみなせる. ピッチの同期とは, 2 つの音の基準音, 例えば A4 の音が十分に近いことであると定義できる. 演奏においてピッチ同期は重要であり, 多くの楽器は演奏前のピッチチューニングにより同期を行う. 一方, 管楽器のように温度の変化などでピッチが変わる場合は, 時変なピッチコントロールを行う必要がある. 単独演奏の場合, 音の相対ピッチ差が保たれていれば絶対ピッチはそれほど重要ではない. しかし, 複数による合奏演奏の場合は, パート間のピッチが近い必要がある. 一説には, 人のピッチ分解能は 5-6[cent] とされている[Loeffler, 2006]. しかし, 合奏においてピッチのズレが認知できることと, 合奏として不快と感じることとは必ずしも一致せず, この点に関する研究は不足している. 本稿では, ピッチに関する同期は使用するロボットの動作モデルに依存するため, 扱わないものとする. また, 和音の同期は主に和音の “進行”, モード (調) により決定される. セッション合奏のようなメロディーの生成が必要な場合は, 生成されたメロディーと伴奏和音の同期が特に問題となる. 本稿では演奏楽曲の和音進行とメロディーは既知であるとし, 考慮しない.

以上のことから, 合奏タスクをある程度のズレを許容したリズムの同期と定義する必要がある. 本稿では, 合奏タスクの失敗を 4 分音符間隔以上のズレが生じること, 成功を上記のような失敗が演奏中に発生しないことと定義し, 議論を進めていく.

3.2 合奏の構成

本稿の合奏の構成を, ロボット 2 体と人 2 名であるとする.

ロボットのうちは 1 台は VOCALOID による歌唱と, 手の振りによるビートタイミングに合わせたダンス[Oliveira, 2010b]を行う (以後, “ダンスロボット” と呼ぶ). もう 1 台はテルミンを演奏する (以後, “テルミンロボット” と呼ぶ). テルミンは音量とピッチの二種類のアンテナを持った非接触性の楽器である. ロボットの演奏動作には, テルミン演奏モデルに基づいた, ハードに依存する部分を分離した動作生成モジュール[Mizumoto, 2010a]を用いる. テルミンロボットのビートトラッキングの音響・画像情報の入力に Kinect を用いる. 詳しくは次節で述べるが, 音響・画像情報に加え, 深度情報を利用することで, 複数音源環境でもロバストな動作を可能とした.

人のうち一人はギターを担当する. 初期テンポ共有のために, 演奏開始時に 4 分音符間隔のギター打撃音を鳴らす. その後は楽曲に応じたストローク動作で演奏を行う.

もう一人はフルートを担当する. フルート奏者の正面に USB カメラを設置し, Ready, Start, Fermata-End の 3 つのジェスチャを検出することで, ロボットの演奏との同期を行う[Lim, 2010].

以下に合奏の構成をを示す.

合奏の構成

- ダンスロボット：歌唱 (VOCALOID)&ダンス [Oliveira, 2010b]
- テルミンロボット：テルミン演奏[Mizumoto, 2010a]
- 人：フルート (ジェスチャー認識[Lim, 2010])
- 人：ギター (ビートトラッキング)

3.3 複数ロボット合奏の問題と解決

前節で示した条件下でのマルチロボット同期合奏では, 以下のような問題が生じる.

1. 複数のビートトラッキング手法を用いた合奏遂行
2. 音源が増えたことによる検出拍候補の増加

以下でこれらの問題の解決について議論する.

3.3.1 複数のビートトラッキング手法を用いた合奏遂行

今回, 2 体のロボットそれぞれに対し, 異なるビートトラッキング手法を用いている. その理由は, ロボットのビートトラッキングに対する要求がそれぞれ異なるからである. ダンスロボットには, ダンスにおいてロボットの動作制約があるので, 大きなテンポ変動には対応できない. 一方テルミンロボットの演奏動作は, ダンスロボットに比べて比較的小さく, テンポの変動に対し機敏に対応できる. また, 担当パートがベースのような伴奏パートに当たるので, 同じく伴奏であるギターの演奏を正確に追従する必要がある. 以上より, ダンスロボットの動作タイミングは IBT[Oliveira, 2010a]による比較的人によるテンポの流動性を吸収した拍時刻を, テルミンロボットの動作タイミングは視聴覚ビートトラッキング[Itohara, 2011]によるテンポ変動に鋭敏な拍時刻を与えることとした. 以下, 前者のビートトラッキングを “ハードビートトラッキング”, 後者を “ソフトビートトラッキング” と呼ぶ.

この様な二つの異なるビートトラッキング手法間での同期を解決する方法は二つある. 一つはロボット間で電子的な通信を行うことである. しかし, これは人にはその同期の様子が伝わらず, 人との同期の要件を満たすことはできない.

もう一つの方法は, 人同士の合奏同期と同様に “リズムリーダー” を決めることである. オーケストラで言えば指揮者が, ロックバンドで言えばドラム奏者がそれにあた

る．リーダーとの同期は視覚や聴覚と言った様々なモダリティで行われている．本稿における合奏では，ギター奏者がリーダーにあたる．しかし，人の，特にアマチュア奏者の演奏の場合，テンポの流動性は回避しきれない．その一方でダンスロボットは，ハードビートトラッキングを用いており，また，ダンスや歌声という視聴覚からリズムのとりやすい動作をしている．よって，ダンスロボットを相補的なリズムリーダーとすることで，アマチュア奏者でもロボットとのテンポの安定した同期合奏が実現できる．

3.3.2　音源の増加による検出拍候補の増加

テルミンロボットはギター奏者に追従するビートトラッキング手法を用いている．しかし，テルミンロボットの入力デバイスである Kinect には，ギター以外の多数の音が混合された状態で入力される．各演奏者の発音タイミングにはズレが生じているので，ソフトビートトラッキングがギター以外のズレに引きずられ，その誤差が蓄積することで，合奏タスクが失敗する可能性がある．これに対する一般的な解決法は，(1) 対象楽器に対するトラッキングのモダリティを増やして精度を高めること，(2) 方向による音源分離をしてギターの音だけを強調すること，である．(1) は，リズムリーダーの音だけでなく，動きなど別の要素に注目することで同期を図ることに一致する．(2) はカクテルパーティー効果のような，多数ある音の中から目的音だけに着目することと同義である．本稿では (1) を採用する．具体的には，Kinect の深度情報を使い，ギターの視聴覚統合ビートトラッキングの追従性能を向上させることで上記の問題を解決する．次章にてその詳細を示す．

4　Kinect を用いた視聴覚統合ビートトラッキング

4.1　視聴覚統合ビートトラッキング概要

本稿では，ロボットがギター演奏との同期を行うために，ギターの演奏音と手の動作との相関性を用いた視聴覚統合ビートトラッキング[Itohara, 2011]を用い，ロボットの演奏タイミング検出を行う．出力は入力演奏のテンポと 1 小節を 4 等分した拍の位置であり，ロボットはこれで示されるタイミングに基づいて演奏する．ロボットの演奏動作生成に時間がかかるので，それにしたがって拍推定は少し時間を遡って行われる．本実験では 500[msec] とした．

従来手法において，視覚情報処理，つまり手のトラッキングの部分で，手とギターの色が似ているために起こる手の誤検出の問題があった．本章では，Kinect の深度情報を用いたギター平面の検出，及び画像マスキングによる，色の類似に頑健な手のトラッキングを報告する．これにより，ギタービートトラッキングの性能向上が期待される．音響情報・視覚情報処理，及びパーティクルフィルタの実装に関する詳細は文献[Itohara, 2011]に譲り，以下ではギターのマスキングについて述べる．

4.2　深度情報によるギターマスキング

Kinect による入力は，サイズ 640×480[pixel] の RGB と深度画像である．また，深度画像として各ピクセル座標における x, y, z 方向の値 (単位:[m]) が得られる．x, y, z 正の方向はそれぞれ水平 Kinect からむかって左，鉛直下向き，カメラ方向である．

以下にギターのマスキングの過程を示す．

1. 背景閾値以上の奥行き (z) を持つ座標を，RGB 画像，深度画像においてマスクをかける
2. 深度画像を縮小．以下，非マスク部を “特徴点” とする．
3. 特徴点からギターの平面パラメータを導出
4. 深度画像の各座標と 3 の平面の距離を計算し，閾値以下なら対応する RGB 画像上の点にマスクをかける

本稿では，背景閾値を 3[m]，画像の量子化は 16×12，4. の平面との距離閾値は 5[cm] とした．以下で 3. における，3 次元空間における Hough 変換を用いた，ギターの表板を表す平面パラメータの推定について述べる．

Hough 変換では，画像中の各特徴点を通るすべての平面のパラメータを算出，パラメータ空間に対して投票を行い，最大票を獲得したパラメータを推定平面のパラメータであるとする．平面パラメータは，球座標における原点を始点とする平面の法線ベクトル (ρ, θ, ϕ) である．ρ は原点と平面の距離を表す．それぞれのパラメータの定義域，各パラメータの空間の分割幅は以下のように定めた．

$$0.7 \leq \rho \leq 1.4, \qquad \rho_{bin} = 0.05[\mathrm{m}]; \tag{1}$$

$$0 \leq \theta < \pi/4, \qquad \theta_{bin} = \pi/16[\mathrm{rad}]; \tag{2}$$

$$0 \leq \phi < 2\pi, \qquad \phi_{bin} = \pi/6[\mathrm{rad}]; \tag{3}$$

5　実験検証

本章では，4.2 節での深度情報を用いた手のトラッキングの簡単な評価を行う．また，それらを実際に適応した複数ロボット，複数人数合奏実験の検証を行い，解決された課題，今後に残された課題についての議論を行う．

5.1　ギターマスキングによる手のトラッキング

本節では，深度情報を用いた手のトラッキングを RGB 入力のみのものと比較し，性能比較を行う．二つの手のトラッキング手法は，入力がマスクされた画像か否か以外は同じで，オプティカルフローにより変位ベクトルをとった後，その平均を中心とした矩形と色相カーネルを用いた平均値シフト法により手の位置を取得する．詳しくは，文献[Itohara, 2011]を参照されたい．

図 1 に手のトラッキングの結果の一部を示す．既存の RGB 画像のみの入力では，手を指し示すカーソルがギター

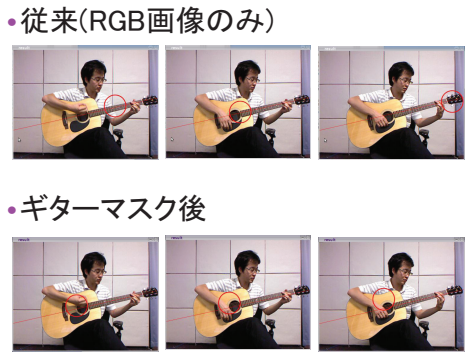


Figure 1: 手のトラッキング結果の比較．赤い丸が手の位置に対応する．上段と下段は同じフレームに対応している．

Figure 2: 合奏デモの様子．右側が Hearbo(1 号機) で，左側が Hearbo(2 号機)．

に吸い寄せられ手を避けるかのような挙動が見られることがあった．これは色相カーネルの選択において，色相の近いギターの色に引き寄せられたことが原因だと考えられる．また，ギターのヘッド (ギター左手側の先) 部分も手同様演奏中に動くことがある．これによりオプティカルフローベクトルがそちらに現れ，ヘッド部分にトラッキングカーソルが動く場合が見られた．一方，マスク後は上記のような誤検出は一切見られなかった．ただし，計算コストがフレームレートに比べ大きすぎるため，手のトラッキング結果が表示されるころには実際の手の位置が異なってしまっていることが確認できた．

5.2 複数ロボットによる同期合奏

3.2 節で示した構成で合奏デモを行なった．使用したロボットは，HRI-JP のヒューマノイドロボット，Hearbo(1 号機，2 号機) で，肩，肘，手首，指 (片手)，首にそれぞれ 2,1,2,4,3 の自由度を持つ．Hearbo(1 号機) にダンスロボットを，Hearbo(2 号機) にテルミンロボットを割り当てた．楽曲はイングランド民謡のグリーンスリーブスを用いた．図 2 に実際の合奏の写真を示す．前節の手のトラッキング精度向上の結果，テルミンロボットのギターへの追従性は大きくあがったと言える．また，ダンスロボットのハードビートトラッキングにより，人のテンポ流動性が抑えられ，ズレの誤差の蓄積が減り，合奏タスクの成功率が向上した．今後はこれらのタスク成功率を定量的に評価する必要がある．

今回の合奏実験の成功率は 30%程度に留まっている．この一番の原因は，2 つのビートトラッキング手法間の同期が失敗することにある．これを解決するためには，ロボット間の明示的な同期が必要である．人同士の場合は，リズムリーダーを最も信頼をした拍推定を行なって同期を行なうが，同時に全体の同期も考慮する．この行動は，例えば誰かがフレーズを誤った場合にそちらを注視するといった行動に現れる．現在のビートトラッキング手法では，どちらも一つの音源のみに頼ったビートトラッキングを行なっており，ロボット同士はお互いの演奏や動作を一切考慮していない．この状況が解決されることで，合奏の成功率は大きく上がると考えられる．

6 おわりに

本稿では，複数ロボットと複数人で構成されるリズム同期に焦点をあてた合奏について議論し，ロボット 2 体と人 2 名による合奏システムについて報告した．その際，リズム同期で必要なビートトラッキングにおいて，ハードビートトラッキングを用いることで，人の流動的な演奏と合わせて相補的なテンポの維持が可能になった．また，ソフトビートトラッキングにおいて，Kinect の深度情報を合わせたマルチモーダルビートトラッキングを用いることで，伴奏演奏により合った同期演奏が達成できた．

今後の課題として，ロボット合奏タスク達成率の向上があげられる．そのためには，ビートトラッキング自体の精度向上，複数のビートトラッキング手法間の協調のようなロボット間の同期の実現が必要となる．本稿では，リズムのみを考慮した合奏を行なったが，ピッチなどの他の要素が同期ことに注目した合奏考察は，豊かなインタラクションという点で重要である．例えばピッチにおいては，声楽における長音のビブラート，ギターなどの弦楽器で弦を押し上げることによるピッチ変化などに現れる．また，テルミンのような無音階楽器の演奏における他の楽器とのピッチ同期は特に重要である．これに対し，実際の演奏音を聞いて，適応的に動作を変化させるロボット演奏の研究も行なわれている[水本他, 2011]．また，従来の音楽合奏タスクの評価は，本稿同様，リズムのみに着目したものが多かった．上記のようなピッチ変動，またはセッションなどのメロディー生成における和音同期や，合奏としての楽しさといった主観的な要素などを盛り込んだ，新しい合奏の評価尺度の考察が必要である．

謝辞 本研究の一部は科研費 (S)，新学術領域，JST-ANR BINAAHR，GCOE の支援を受けた．また，Hearbo の使用許可をいただいた HRI-JP に感謝します．

参考文献

[Fredrickson, 1994] W.E. Fredrickson. Band musicians' performance and eye contact as influenced by loss of a visual and/or aural stimulus. *Journal of Research in Music Education*, 42(4):306, 1994.

[Friberg, 2002] A. Friberg and A. Sundström. Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern. *Music Perception*, 19(3):333–349, 2002.

[Goto, 2001] M. Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *J. of New Music Research*, pages 159–171, 2001.

[Hainsworth, 2003] S. Hainsworth and M. Macleod. Beat tracking with particle filtering algorithms. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,*, pages 91–94. IEEE, 2003.

[Itohara, 2011] T. Itohara et al. Particle-filter based audio-visual beat-tracking for music robot ensemble with human guitarist. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*. IEEE, 2011.

[Lim, 2010] A. Lim et al. Robot musical accompaniment: integrating audio and visual cues for real-time synchronization with a human flutist. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, pages 1964–1969, 2010.

[Loeffler, 2006] B.D. Loeffler. *Instrument Timbres and Pitch Estimation in Polyphonic Music*. PhD thesis, Citeseer, 2006.

[Mizumoto, 2010a] T. Mizumoto et al. Human-robot ensemble between robot thereminist and human percussionist using coupled oscillator model. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, pages 1957–1963. IEEE, 2010.

[Mizumoto, 2010b] T. Mizumoto et al. Integration of flutist gesture recognition and beat tracking for human-robot ensemble. In *Proc. of IEEE/RSJ-2010 Workshop on Robots and Musical Expression*, pages 159–171, 2010.

[Murata, 2008] K. Murata et al. A beat-tracking robot for human-robot interaction and its evaluation. In *Proc. of 8th IEEE-RAS Int'l Conf. on Humanoids*, pages 79–84. IEEE, 2008.

[Oikonomidis, 2011] I. Oikonomidis et al. Efficient model-based 3d tracking of hand articulations using kinect. *Procs. of BMVC, Dundee, UK (August 29–September 10 2011)[547]*, 2011.

[Oliveira, 2010a] J.L. Oliveira et al. Ibt: A real-time tempo and beat tracking system. In *Proc. of Int'l Society for Musical Information Retrieval Conference*. IEEE, 2010.

[Oliveira, 2010b] J.L. Oliveira et al. Synthesis of dancing motions based on a compact topological representation of dance styles. In *Proc. of IEEE/RSJ-2010 Workshop on Robots and Musical Expression*. IEEE/RSJ, 2010.

[Otsuka, 2010] T. Otsuka et al. Design and Implementation of Two-level Synchronization for Interactive Music Robot. In *Proc. of Association for the Advancement of Artificial Intelligence*, pages 1238–1244, 2010.

[Pan, 2010] Y. Pan et al. A robot musician interacting with a human partner through initiative exchange. In *Proc. of Int'l Conf. on New Interfaces of Musical Expression*, pages 166–169, 2010.

[Petersen, 2008] K. Petersen et al. Development of a real-time instrument tracking system for enabling the musical interaction with the waseda flutist robot. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, pages 313–318, 2008.

[Saenko, 2011] K. Saenko et al. Practical 3-d object detection using category and instance-level appearance models. In *Proc. of IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*. IEEE, 2011.

[Solis, 2008] J. Solis et al. Understanding the mechanisms of the human motor control by imitating flute playing with the Waseda Flutist Robot WF-4RIV. *Mechanism and Machine Theory*, 44(3):527–540, 2008.

[Weinberg, 2009] G. Weinberg et al. The Creation of a Multi-Human, Multi-Robot Interactive Jam Session. In *Proc. of Int'l Conf. on New Interfaces of Musical Expression*, pages 70–73, 2009.

[水本他, 2011] 水本 武志 他. テルミン演奏ロボットのための unscented kalman filter による適応的音高制御. In **日本ロボット学会第 *29* 回学術講演会**, 2011.

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-B102-11

耳介を持つバイノーラル聴覚ロボットの音源方向推定の検討

On sound direction estimation by binaural auditory robots with pinnae

公文誠, 木元大輔
Makoto KUMON and Daisuke KIMOTO
熊本大学
Kumamoto University
kumon@gpo.kumamoto-u.ac.jp

Abstract

Binaural auditory systems which use two microphones to percieve auditory signals are considered as the minimal configuration for practical sound localization since animals are able to achieve this ability only with their two ears. Interaural Time Difference(ITD) of two signals measured by two ears can be used to estimate the direction of the sound source in the plane where two ears locate. However, the deviation perpendicular to the plane does not make any difference in ITD, which implies additional features but ITD are necessary to estimate the direction of the sound source. It is known that human and animals with two ears utilize their pinnae, or external ears, to localize sound sources since irregular shapes of pinnae encode the direction of the sound source as frequency domain cues.

Because the relationship between the angle and the cue is complicated, this paper considers the method to extract such frequecy cues precisely by introducing a linear transformation of the cue space. In order to validate the proposed approach, experiments with a real binaural auditory robot with a pinna were conducted, and results show the improvement of the obtained estimates.

1 はじめに

ロボットにとって音信号を利用することは, 周辺環境の認識や柔軟なマンマシンインターフェイスを実現する上で不可欠で, ロボット聴覚として盛んに研究されている[奥乃, 2001]. このように音信号を利用する聴覚ロボットにとって, 音の到来方向あるいは音源の方向を正確に推定することは, 特に重要な基礎機能である. また, 人間や動物は 2 つの耳のみで, 現実的な音源定位を実現しているので, バイノーラル聴覚は音源定位のための必要最小限の構成だと考えられる. ロボットでも 2 つのマイクロホンだけで音源定位能を実現することは, 聴覚システムの簡素化や音源定位の原理の解明など, 興味深い課題を含んでいると言えよう. そこで, 本研究では 2 つのマイクロホンからなるバイノーラル聴覚ロボットにおいて音源定位を実現するための方法を検討することを考える.

観測された音信号から音源方向を推定するには, MUSIC 法[Shimidt, 1986]やビームフォーマ[佐々木, 2010]などマイクロホンアレイを用いた方法が良く知られている. バイノーラル聴覚にあっても, マイクロホンを含む面内での音源の変位については, マイクロホンで収録される音信号の間の到達時間差を測定することで, 音源の方向を推定することが出来る (このような音源方向のことを以下では方位角と呼ぶ). しかし, マイクロホンを含む平面に対して垂直な方向の変位 (このような音源方向を仰伏角と呼ぶ) は, 音源と 2 つのマイクロホンまでの行路差が不変なため, 両耳間時間差では音源方向を推定することが出来ず, 音源方向の推定には他の特徴量が必要となる.

人間や動物では, 耳に耳介と呼ばれる音の反射・集音を果たす器官が存在する. 耳介の形状は一般に複雑なため, 耳介の音響特性が音の到来方向に応じて異るという性質がある. 特に耳介のゲイン特性が顕著に抑制されている帯域を耳介ノッチと呼び, 耳介ノッチの周波数が音源の仰伏角の関数になっていることが知られている[Shaw, 1968]. 従って, 耳介ノッチの周波数を検出すれば音源方向を推定することが可能である. 実際, このような考えに基づき, Shimoda ら[Shimoda, 2006]は耳介ノッチの周波数と仰伏角の間を線形モデルで近似し, 音源の上下を推定する方法を提案している. また, Hörnstein ら[Hornstein, 2006]は周波数領域での特徴量とロボット頭部の運動パターンをニューラルネットワークで学習し, 音源方向へロボット頭部をトラッキングする方法を示した. Finger ら[Finger, 2010]は, 耳介だけでなくロボット頭部の影響を考察し, 方位角と仰伏角の推定に関連があることを報告している. 章

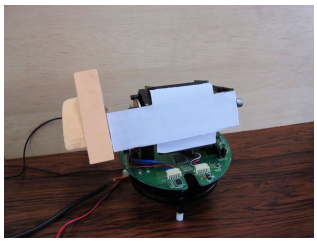
ら[章, 2008]は, 音源方向の推定のための特徴量として両耳間レベル差を用い, 事前に学習した特徴量との相関を求めることで, 仰伏角を推定している.

本研究ではこれらの手法の性能改善を目指し, 事前に求めた特徴量と音源方向の関係から, 観測した特徴量を求める手法の改良を試みる. 周波数領域での特徴量はベクトル量で表されるので, この問題は, 与えられたデータ点に対して, 既知の学習データの集合の中から最も「近い」ものを選び出す操作に対応する. しかし, 学習データ同士の差異が小さい領域と, 大きな領域では, この選び出す操作の精度が異なるため, 音源方向の推定にこのような学習データ群の歪みがあれば, 推定性能が音源方向に依存して変化してしまい問題である. そこで, 本研究では学習データである特徴量ベクトルを適当に写像し, このような歪みを取り除くことで, 音源方向推定の性能改善を考える.

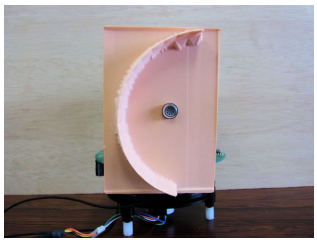
本稿の構成は次の通りである. まず, 対象とする耳介を有するバイノーラル聴覚ロボットを紹介し (第 2 節), 周波数特徴量から音源方向を推定する方法を説明する (第 3). この中で特徴量ベクトル同士の「距離」を与える計量について考察し, 音源方向推定の性能改善法を提案する. 提案法の性能を確認するため, 第 2 節で紹介したロボットを用いた実験結果を第 4 節で示し, 最後にまとめる.

2 耳介つきバイノーラル聴覚ロボット

本研究では図 1 に示すロボット頭部を用いる. このロボットには 2 つのマイクロホンが取り付けられたバイノーラル構成になっており, これらのマイクロホンの一つには人工耳介が取り付けられている. 耳介は対数螺旋状の形状を有し, その中心にマイクロフォンが埋め込まれる構造である. 後述するように本研究では左右のマイクロホンで受聴される信号がロボット形状の非対称性によって異った周波数伝達特性の影響を受けることを利用すると想定している. ここでは, 特に耳介の影響を際だたせるため, 片方にのみ耳介を取り付けることとした.



(a) ロボット頭部



(b) 耳介

Figure 1: バイノーラル聴覚ロボット

ロボット頭部は仰伏角, 方位角の 2 方向に動作可能で, 超音波モータ駆動のため, ロボットが動作中であっても可聴域でのエゴノイズはほとんど生じない. ロボットを制御する計算機は姿勢角および角速度の規範値を指定するとともに, マイクロホンからの信号を記録する.

3 音源方向推定

ロボットの受聴する音信号は環境やロボット自身の影響を受け原信号と異ったものとなる. 今, ロボットを基準とした音源までの距離, 仰伏角, 方位角を r, θ, ϕ とすると原信号から左右のマイクロホンへの音響特性を表す伝達特性のうち, ロボットの身体によるものを $H_l(\theta,\phi;\omega)$, $H_r(\theta,\phi;\omega)$, 環境の特性を $H_{le}(r,\theta,\phi;\omega)$, $H_{re}(r,\theta,\phi;\omega)$ と表すとする. また, 原信号 $s_O(\omega)$ に対してロボットの左右のマイクロホンで受聴する信号をそれぞれ $s_l(r,\theta,\phi;\omega)$, $s_r(r,\theta,\phi;\omega)$ と表せば,

$$\begin{aligned} s_l(r,\theta,\phi;\omega) &= H_l(\phi,\theta;\omega)H_{le}(r,\theta,\phi;\omega)s_O(\omega) \\ s_r(r,\theta,\phi;\omega) &= H_r(\theta,\phi;\omega)H_{re}(r,\theta,\phi;\omega)s_O(\omega) \end{aligned}$$

の関係がある. もし環境からの影響が $H_{le} \approx H_{re}$ と出来るのであれば, 両耳間レベル差 Δ_s は

$$\begin{aligned} \Delta_s &\equiv 20\log|s_l(r,\theta,\phi;\omega)| - 20\log|s_r(r,\theta,\phi;\omega)| \\ &= 20\log|H_l(\theta,\phi;\omega)H_{le}(r,\theta,\phi;\omega)s_O(\omega)| \\ &\quad -20\log|H_r(\theta,\phi;\omega)H_{re}(r,\theta,\phi;\omega)s_O| \\ &\approx 20\log|H_l(\theta,\phi;\omega)| - 20\log|H_r(\theta,\phi;\omega)| \end{aligned}$$

と近似でき, 両耳間レベル差 Δ_s が音源方向 θ, ϕ の関数として $\Delta_s(\theta,\phi,\omega)$ となるので, 両耳間レベル差はロボット身体の影響だけで特徴づけられることが分かる. 特に耳介はマイクロホン近傍にあって, 伝達特性を強く特徴づけると考えられ, 音源方向推定の情報を与えると期待される.

3.1 特徴量

両耳間レベル差が音源方向によって特徴づけられたものであるので, 本研究では対象とする周波数帯域の両耳間レベル差を特徴量ベクトルとする. なお, 原信号 s_O に含まれていない, あるいは非常に小さな周波数成分については Δ_s が正しく求まらないことが考えられるので, これを除外して考える必要がある. このため, 適当な正定数 ϵ に対して

$$f(x,a,b) = \begin{cases} 0 & \text{if } a<\epsilon \text{ or } b<\epsilon \\ x & \text{otherwise} \end{cases}$$

となる関数 f を用いて, 特徴量ベクトル X を

$$\begin{aligned} X &= [f(\Delta_s(\omega_1),|s_l(\omega_1)|,|s_r(\omega_1))|, \\ &\quad \cdots, f(\Delta_s(\omega_N),|s_l(\omega_N)|,|s_r(\omega_N)|)]^T \end{aligned}$$

と定める. ここで $\omega_1 \cdots \omega_N$ は対象とする周波数成分を表す.

3.2 音源方向推定

本研究では, 事前に音源方向と特徴量ベクトルとの間を適当な方法で学習し, この情報を利用して受聴した音信号から音源方向を推定する方法を考える.

3.2.1 学習ベクトルとの相関

音源方向 θ, ϕ から周波数成分を十分に含んだ試験信号を与え, 規範となる特徴量ベクトル $X_d(\theta, \phi)$ を計測する. 推定対象となる方向を $\theta_1, \cdots, \theta_{N_\theta}, \phi_1, \cdots, \phi_{N_\phi}$ とすれば, $N_\theta \times N_\phi$ 点について全て特徴量ベクトルを計測し, これらを学習データとして保存する.

次に, 方向を推定したい音信号が与えられ, この信号の特徴量ベクトル X が得られたとする. この時, この方向の学習データは, 特徴量ベクトル X との間で高い相関を示す. 今, 相関は以下のようにベクトル間の適当な計量の下での規格化された内積として表せる.

$$S(X, X_d) = \frac{< X, X_d >_M}{|X||X_d|} \tag{1}$$

なお, S は 特徴量ベクトル X が与えられた時 θ, ϕ の関数として求まるので, $S(X) = S(X, \theta, \phi)$ である.

3.2.2 推定方向の算出

S の観測には雑音などの影響によって, 望ましい領域以外でも大きな相関値を持つ可能性がある. そこで, 複数のフレーム (N_F をフレーム数とする) で観測された S (それぞれ $S^1 \cdots S^{N_F}$ と表記) について, 要素毎に

$$\bar{S}(\theta, \phi) = \alpha \prod_{k=1,\cdots N_F} \left\{ S^k(\theta, \phi) - \min_{\xi,\eta} \left(S^k(\xi, \eta) \right) \right\}$$

として, 雑音の影響を抑制する. ただし, α は

$$\sum_{\theta,\phi} \bar{S}(\theta, \phi) = 1$$

となるような正規化係数である. この操作は, S を確率分布の離散的な表現と見做した時の複数観測を統合する操作あるいは際だったピークを強調し, 点在する低いピークを抑制する操作と考えることが出来る. 以下では 対象とするピークを強調することを期待して, N_F 個の積の中にあるフレームの S を複数回適用する処理も含めて考えることとする.

また, この $\bar{S}$ は学習した方向については正しい値が得られると期待されるが, それ以外の点の情報は適当な方法で補完するなど汎化の必要がある. 相関を重みとした重心を求めることも考えらえるが, 学習領域の辺縁部では不正確になるおそれがある. そこで, この分布を適当なモデルにあてはめて, モデルのパラメータによって音源方向を推定することとした. 具体的には本研究ではモデルとしてガウス分布を用い, この平均値と分散によって, 音源方向の推定情報とした. このため, 以下で定める評価関数 E を最小化する平均 μ と共分散 Σ を求めている.

$$E = \sum_{\substack{\theta_1, \cdots, \theta_{N_\theta} \\ \phi_1, \cdots, \phi_{N_\phi}}} \left| \bar{S}(X, \theta, \phi) - \rho p(\theta, \phi, \mu, \Sigma) \right|^2 \tag{2}$$

ここで, $p(\theta, \phi, \mu, \Sigma)$ は平均 μ, 共分散 Σ の二次元ガウス分布を与え, ρ は $\bar{S}$ に合わせて分布の頂点を揃えるためのパラメータを表すものとする.

3.2.3 提案法

文献[章, 2008] では, 相関にユークリッド計量を選び, 音源方向の推定値として, S を最大化するもの (ガウス分布の分散をデルタピークとしたもの) を与えている.

今, 適当な単位列ベクトルの列 $x_1, \cdots x_m$ を考え, 計量 M の相関を考える. (ここでベクトルの次元 n は m よりも大きいとする) ここで

$$A = [x_1, x_2, \cdots, x_m]$$

となる行列 A を考えれば,

$$x_i = A\,[0, 0, \cdots, 1, \cdots, 0]^T$$

である. A の擬似逆行列 $A^\dagger \equiv A^\dagger = A^T(AA^T)^{-1}$ を考えると

$$A^\dagger x_i = [0, 0, \cdots, 1, \cdots, 0]^T$$

なので, 計量 M に $A^\dagger$, j 番目のみが 1 の列ベクトルを y_j とすれば,

$$< x_i, y_j >_M = \delta_{ij}$$

となり, 与えられた学習ベクトルを区別する最適な判別関数を得ることが出来る.

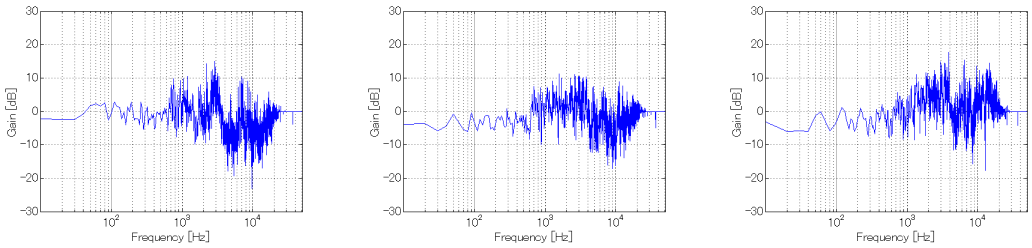
このことから, 本研究では, 特徴量ベクトルの判別にこの擬似逆行列を計量とし, 規範ベクトルには適当な単位ベクトルから成る基底群を用いる方法を提案する. 各学習ベクトル間の距離を最大にする計量となっており, 従来法に比べ音源方向に依らず方向推定性能を均質化する効果がある. 実装上は, 擬似逆行列と単位ベクトルを乗じて得られるベクトル列, つまり擬似逆行列の行ベクトルの列を学習ベクトルとして記憶しておけば十分なので, 方向推定における演算量は, 従来法 (文献[章, 2008]) と同一であることに注意されたい.

4 実験

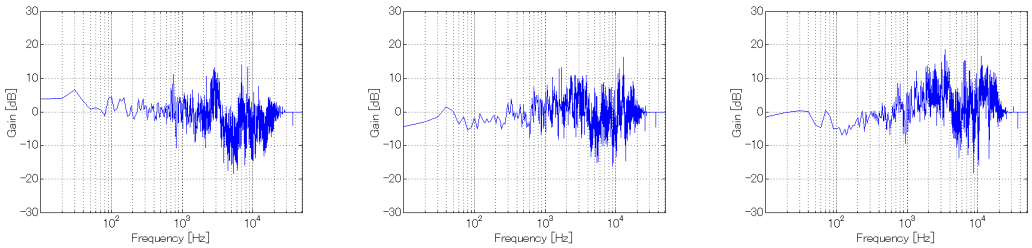
4.1 実験環境

学習データの収録および実験は, 図 2 に示す居室内で行った.

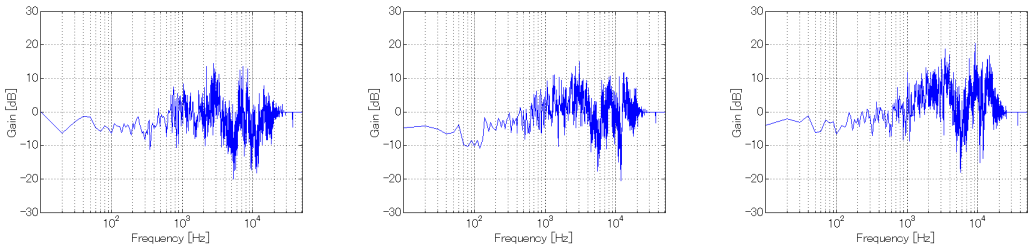
学習におけるロボット頭部の方向は θ, ϕ それぞれを 180 度を 10 度, 90 度を 10 度の刻み幅の格子点上とし, ロ



(a) 仰伏角 0(deg) 方位角-30(deg)　(b) 仰伏角 0(deg) 方位角 0(deg)　(c) 仰伏角 0(deg) 方位角 30(deg)



(d) 仰伏角 30(deg) 方位角-30(deg)　(e) 仰伏角 30(deg) 方位角 0(deg)　(f) 仰伏角 30(deg) 方位角 30(deg)



(g) 仰伏角 60(deg) 方位角-30(deg)　(h) 仰伏角 60(deg) 方位角 0(deg)　(i) 仰伏角 60(deg) 方位角 30(deg)

Figure 3: 両耳間レベル差

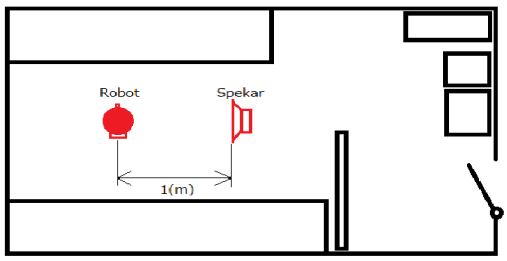


Figure 2: 実験環境

ボットの前方 1m に設置したスピーカから白色雑音を印加し音源とした. アンプで適当に増幅した信号をサンプリング周波数は 100kHz でサンプリングした信号を AD 変換して取り込んだ.

なお, この学習データの計測では, 特徴量ベクトルを複数回測定し, 不適切な測定を除外した後, 平均操作を施している.

4.2 両耳間レベル差

本研究で基礎とする両耳間レベル差が音源方向の関数になっていることを確かめるため, まず特徴的な姿勢での両耳間レベル差を示す.

聴覚ロボットの頭部角の変化の下で測定した両耳間レベル差の例を図 3 に示す. 高周波数域で 0dB になっているのは, 十分な成分が検出されなかったため閾値 ϵ によって計算対象から除外されたためである. 頭部方向に応じて数 kHz から 10kHz 付近の帯域が顕著に変化していることが分かる.

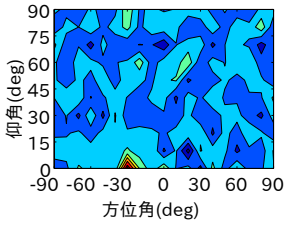
4.3 方向推定

上述の通り, 両耳間レベル差が音源方向の特性を与えることが分かったので, 前節の方法に従って, 音源の方向を推定する実験を行った. この実験では, 学習データとは別に, 新たに収録した音信号によって, 音源方向の推定を行うこととし, 学習データに含まれない方向の音源でも検証する.

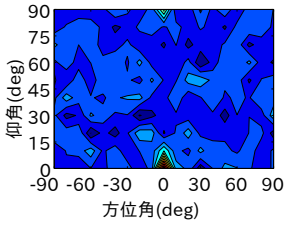
収録した音信号から (1) に従って求めた S を図 4 に示す. 対象とする音源の方向は学習に用いた方向に含まれているもので, 学習された結果を適切に想起できるかを判別する. (a) から (c) は計量に擬似逆行列を用いる提案法に

よる相関, (d) から (e) はユークリッド計量による内積によるものである.

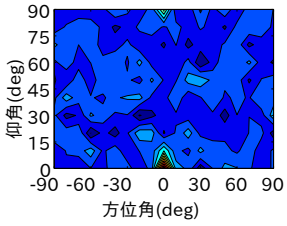
図より, いずれも音源方向に最大値を持つことから, 学習データを正しく想起したことを示している. 提案法は音源方向近傍にのみ鋭いピークを持つのに比べ, 内積では広くなだらかなパターンを示しているので, 提案法が単なる内積に比べ, 弁別能が高いと言える.



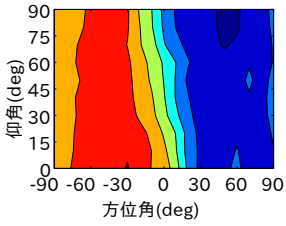
(a) 提案法 仰伏角 0(deg) 方位角-30(deg)



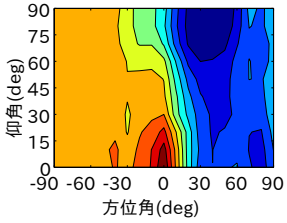
(b) 提案法 仰伏角 0(deg) 方位角 0(deg)



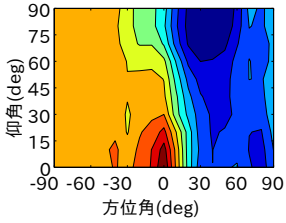
(c) 提案法 仰伏角 0(deg) 方位角-30(deg)



(d) 内積 仰伏角 0(deg) 方位角-30(deg)



(e) 内積 仰伏角 0(deg) 方位角 0(deg)



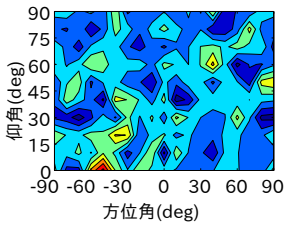
(f) 内積 仰伏角 0(deg) 方位角 30(deg)

Figure 4: 推定結果 (学習ベクトルの想起)

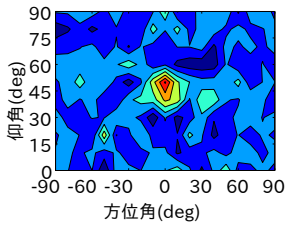
次に学習データに含まれない方向からの音信号に対する推定結果を図 5 に示す. この場合も先の結果と同様, 提案法および内積による方法の両方とも音源方向付近にピークを与えており, 汎化により方向推定が可能であることが分かる. 提案法が内積の場合に比べて鋭いピークを与えている点も, 先の例と同じであり, 推定性能に改善が見られる.

最後に式 (2) を最適化し, 上で求めた S を用いて音源方向を推定した結果を表 1, 2 にまとめる. なお, E の最適化には $\bar{S}$ を確率分布と見做した時の ϕ, θ の期待値および共分散を初期値とし, Mathworks 社 MATLAB の繰り返し最適化法 ($fminsearch$) を適用した. また $\bar{S}$ のためのフレーム数 N_F には 10 と 20 の場合を考え, ここでは単一の S に対する冪を用いた.

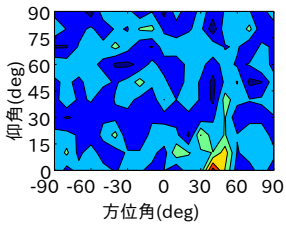
表 1 より, $N_F = 20$ の場合, 学習データに対応するもの



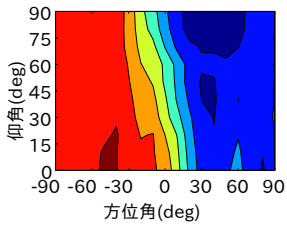
(a) 提案法 仰伏角 0(deg) 方位角-45(deg)



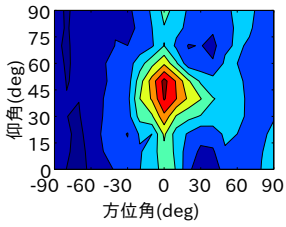
(b) 提案法 仰伏角 45(deg) 方位角 0(deg)



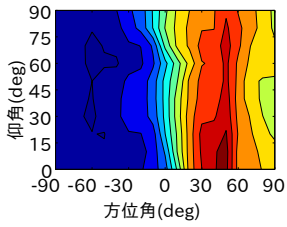
(c) 提案法 仰伏角 0(deg) 方位角 -45(deg)



(d) 内積 仰伏角 0(deg) 方位角-45(deg)



(e) 内積 仰伏角 45(deg) 方位角 0(deg)



(f) 内積 仰伏角 0(deg) 方位角 45(deg)

Figure 5: 推定結果 (汎化能)

は完全に想起出来ていること, 学習データに含まれないものでも 5 度 程度の誤差の範囲で推定出来ていることが分かる. また, 共分散行列の各要素は非常に小さく, ほぼ目的とする方位の信号のみに値を持つ結果となった.

表 2 は内積による従来法の結果を示しているが, 提案法に比べると推定値 (平均) が真値から外れたものが多く, 共分散行列の要素が大きいことからばらつきの大きな結果になっていることが示されている一方, 未学習データに対しては比較的正確な結果が得られた.

以上をまとめると, 提案法は未学習データに対しては従来法と同等程度であったが, 学習した方向に対しては提案が優れた結果を与えることが分かった.

5 おわりに

本研究では, 耳介を用いた音源方向の推定の方法として, 両耳間レベル差を用いる方法を検証し, 文献[章, 2008]を一般化した推定方法の効果を調べた. 学習された特徴量ベクトルの張る空間の計量として, 音源方向に対して均一になるよう学習データから擬似逆行列を用いた計量を導入したところ, 十分に周波数成分の豊かな試験信号に対して音源方向の推定性能が向上することが実験によって示さ

Table 1: 推定結果

真値 (仰伏角, 方位角)	推定結果 (提案法)	
	$N_F = 10$	$N_F = 20$
	平均, 共分散	平均, 共分散
$(0°, -30°)$	(-2.1,-30.9) $\begin{pmatrix} 29.9 & 13.2 \\ \star & 31.82 \end{pmatrix}$	(0.0,-30.0) $\begin{pmatrix} 0.0189 & 0.0077 \\ \star & 0.0128 \end{pmatrix}$
$(0°, 0°)$	(-0.5,0.0) $\begin{pmatrix} 14.2 & 0.0013 \\ \star & 0.319 \end{pmatrix}$	(0.0,0.0) $\begin{pmatrix} 0.0243 & 0.000 \\ \star & 0.000 \end{pmatrix}$
$(0°, 30°)$	(-2.0,30.3) $\begin{pmatrix} 35.3 & -8.38 \\ \star & 19.0 \end{pmatrix}$	(0.0,30.0) $\begin{pmatrix} 0.198 & -0.0047 \\ \star & 0.0067 \end{pmatrix}$
$(0^\star, -45^\star)$	(-29.6,-90.8) $\begin{pmatrix} 340 & 464 \\ \star & 962 \end{pmatrix}$	(0.0,-50.1) $\begin{pmatrix} 11.9 & 18.2 \\ \star & 31.7 \end{pmatrix}$
$(45°, 0°)$	(49.5,-0.2) $\begin{pmatrix} 24.5 & 10.6 \\ \star & 46.0 \end{pmatrix}$	(50.0,0.0) $\begin{pmatrix} 0.315 & 0.184 \\ \star & 0.374 \end{pmatrix}$
$(0°, 45°)$	(-4.0,42.0) $\begin{pmatrix} 35.2 & -17.9 \\ \star & 54.9 \end{pmatrix}$	(0.0,40.0) $\begin{pmatrix} 0.0887 & 0.0309 \\ \star & 0.0907 \end{pmatrix}$

Table 2: 推定結果

真値 (仰伏角, 方位角)	推定結果 (内積)	
	$N_F = 10$	$N_F = 20$
	平均, 共分散	平均, 共分散
$(0°, -30°)$	(-16.6,-32.7) $\begin{pmatrix} 816 & -143 \\ \star & 418 \end{pmatrix}$	(-28.6,-25.5) $\begin{pmatrix} 719 & -152 \\ \star & 241 \end{pmatrix}$
$(0°, 0°)$	(-41.1,29.7) $\begin{pmatrix} 687 & -521 \\ \star & 1060 \end{pmatrix}$	(-10.1,8.7) $\begin{pmatrix} 104 & -85.5 \\ \star & 161 \end{pmatrix}$
$(0°, 30°)$	(-23.9,27.5) $\begin{pmatrix} 722 & 70.1 \\ \star & 121 \end{pmatrix}$	(0.8,30.0) $\begin{pmatrix} 126 & 5.11 \\ \star & 6.09 \end{pmatrix}$
$(0^\star, -45^\star)$	(30.3,-53.5) $\begin{pmatrix} 799 & -121 \\ \star & 416 \end{pmatrix}$	(3.6,-46.9) $\begin{pmatrix} 752 & -108 \\ \star & 283 \end{pmatrix}$
$(45°, 0°)$	(46.1,0.5) $\begin{pmatrix} 39.8 & -2.13 \\ \star & 23.0 \end{pmatrix}$	(47.0,0.0) $\begin{pmatrix} 24.2 & -0.636 \\ \star & 2.66 \end{pmatrix}$
$(0°, 45°)$	(-5.8,44.1) $\begin{pmatrix} 797 & 33.9 \\ \star & 149 \end{pmatrix}$	(-3.9,47.2) $\begin{pmatrix} 531 & 20.9 \\ \star & 49.5 \end{pmatrix}$

れた. 紙面の都合で掲載しなかったが, 周波数成分が疎な信号では性能が劣化することが実験で判明しており, 今後は文献[章, 2008]の方法と組み合わせるなど, ロバストな推定法を考察しなければならない.

本研究では簡単のため仰伏角と方位角の両方を単一の特徴量である両耳間レベル差で求めたが, 方位角については両耳間時間差を用いて求める方法が一般的であり, 複数の特徴量を統合した方向推定は今後の課題である. また, 2 つのマイクロホン両方に耳介を取り付けることで, 両耳間レベル差を強調する方法も考えられ, 方向推定に適した耳介形状や取り付け方法についても検討する必要がある.

参考文献

[奥乃, 2001] 奥乃博, 中臺一博:ロボット聴覚の課題と現状, 情報処理学会研究報告, 音声言語情報処理 pp.69–74, 2001.

[Shimidt, 1986] Schmidt, R.O.: Multiple Emitter Location and Signal Parameter Estimation, in IEEE Trans. Antennas Propagation, Vol. AP-34 pp.276-280, 1986.

[佐々木, 2010] 佐々木洋子, 椛澤光隆, S. Thompson, 加賀美聡, 尾路京一: 低サイドローブ設計 64ch 球形マイクロホンアレイの開発, 人工知能学会研究会資料 SIG-Challenge 研究会, pp.3-8, 2010

[Shaw, 1968] Shaw, E. A., and Teranishi, R.: Sound pressure generated in an external-ear replica and real human ears by a nearby point source, Journal of the Acoustical Society of America, Vol 44-1, pp.240-249, 1968

[Shimoda, 2006] T. Shimoda, T. Nakashima, M. Kumon, R. Kohzawa, I. Mizumoto and Z. Iwai: Spectral Cues for Robust Sound Localization with Pinnae, in Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.386–391, 2006

[Hornstein, 2006] Hörnstein, J., Lopes, M., Santos-Victor, J. and Lacerda, F.: Sound Localization for Humanoid Robots - Building Audio-Motor Maps based on the HRTF, in Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.1170–1176, 2006.

[Finger, 2010] Finger, H., Ruvolo, P., Liu, S.C., Movellan,J.: Approaches and Databases for Online Calibration of Binaural Sound Localization for Robotic Heads, Proceedings of 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.4340-4345, 2010.

[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定位, 日本機械学会論文集 C 編, Vol.74-739, pp.642–649, 2008

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

SIG-AI-Challenges home page (WWW): `http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/`