人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-055 (11/22)

# ＡＩチャレンジ研究会(第 55 回)

*Proceedings of the 55th Meeting of Special Interest Group on AI Challenges*

CONTENTS



日時：2019年11月22日，場所：慶応義塾大学 矢上キャンパス 12棟102室
*Keio University, Tokyo, Nov. 22, 2019*

一般社団法人　人工知能学会
Japanese Society for Artificial Intelligence

一般社団法人 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-055-1
(11/22)

# 音源定位技術が切り開くサルの生態と会話における未解決問題

# Sound localizations reveal the unsolved issues of primate vocal communications, conversations and conservations

香田啓貴 [1*]　豊田有 [2]　Suchinda Malaivijitnond[34]　鈴木麗璽 [5]
Hiroki Koda[1]　Aru Toyoda[2]　Suchinda Malaivijitnond[34]　Reiji Suzuki[5]

[1] 京都大学霊長類研究所
[1] Primate Research Institute, Kyoto University
[2] 中部大学創発学術院
[2] Chubu University Academy of Emerging Science
[3] チュラロンコン大学理学部
[3] Faculty of Science, Chulalongkorn University
[4] タイ国立霊長類センター
[4] National Primate Research Center of Thailand
[5] 名古屋大学大学院情報学研究科
[5] Graduate School of Informatics, Nagoya University

**Abstract:** サル類は，群れを形成しその中でコミュニケーションする点で，ヒトとの類似性が高い．音声によるコミュニケーションを分析すると，発声者間のやり取りの時間規則性は，会話と類似する傾向が確認され，個体間の「会話」頻度の計測可能性は，音声コミュニケーションに基づいた社会ネットワーク分析の良い研究対象となる．本稿では，テナガザルの歌を例題としてマイクロフォンアレイとロボット聴覚技術を利用することで，サルの群れ社会のなかで運用される音声の社会生態学的な役割を明らかにする試みと将来性について論じる．

## 1　ヒトの会話とサルの「会話」

分類学上，ヒトは霊長類 (サル類) の一種に分類される哺乳類である．脳の構造という点も含め身体設計において高い相同性を保持しながら多くの共通点が見られる．行動や心理現象，さらには群れと言った社会は，基本的に身体と生理などの生物基盤の上で実現されている現象である．そのため，行動や社会などの化石に残らない「人らしさ」については，サルの行動や脳機能，社会や生態の適応状況などの証拠と比較して，その進化と起源を推定する方法が有力となる．中でも，ヒトの会話は，個体間で言語活動を相互作用させる行為と言える．ここでいう言語活動とは，表象や意図を記号化する思考だとすれば，そのような記号変換行為はサルでは困難であるため，サルではヒトと同じ会話行為は見つからない．こうしたヒト固有な言語活動を排除すれば，共通成分としては，個体間の信号のやり取りが残る．そして，その信号のやり取りの規則性を客観的に解析すると，ヒトの会話と共通する規則が発見できることが知られてきた [1].

ヒトとサルの発声のやり取りのなかで，共通する規則性として客観的に定義できるのは，時間的規則性である [1]．古くは，Sacks と Schegloff, Jefferson が，ヒトの会話における話者交代現象を単純な時間的規則によって定義した [2]．二者間の発話のやり取り，時間の流れに着目すると，その発話の時間間隔 (inter-call intervals) の生起頻度の分布は，1 秒以内の短時間に最頻値を示す．誰かが話したのち,「返事」や「相槌」は一定時間以内に応答する，という規則である．サルの発声にも同じ傾向が発見できる．サルは群れを形成し，空間的なまとまりを維持して，森の中を移動する．特定の意味を記号的に符号化し，情報をやり取りする発話とは異なり，サルの発声は個体間の空間的な位置確認を行い，凝集性を保つ生態学的な役割があると考えられている．二個体間の発声のやり取りを分析すると，類似した時間規則性が確認できる (図 1)．実際のところ，これらの現象は飼育室でくらすサルでも，声のやり取りが確認できる，同様の時間規則が確認できる [1]．こうして，サルの発声を検出し，時間間隔を評価することで，コミュニケーションを頻繁にとりあう関係性が客観的に評価できる．

## 2　サル研究での様々な問題

霊長類研究者は，発声のやり取りの時間的規則の会話との類似性に着目し，個体間の会話関係についての研究を実施してきた．例えば，飼育下でのリスザルの研究によると，親和的な個体（個体間の距離が近いなど）どうしほど，音声コミュニケーションの頻度が高いという [3]．一方で，野生ニホンザルの会話分析では，普段毛づくろいなどの身体接触が少ない個体同士でむしろ音声コミュニケーションが頻繁であるという，逆の傾向も報告されている [4]．飼育下のピグミーマーモセットの研究では，鳴き交わす個体間で，音声の特徴が似てくるという報告がある [5]．母子間のやり取りという，特別な関係性に着目した会話分析も研究例として存在

*連絡先：京都大学霊長類研究所
〒 484-8506 愛知県犬山市官林 41-2
E-mail: koda.hiroki.7a@kyoto-u.ac.jp

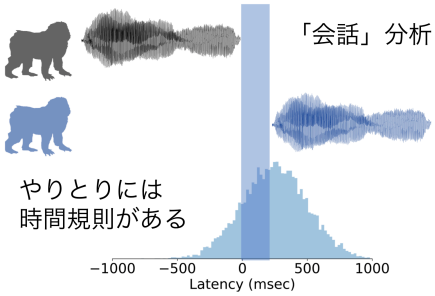



図 1: サルの発声の時間間隔には一定の最頻値が存在する．この値を境界として，音声に対する反応かどうかを客観的に定義することができる．

する．飼育下のコモンマーモセットの母子間の音声コミュニケーションの研究では，会話頻度とコドモの音声発達との間に相関関係が見られる (より「会話」する母子では，コドモの音声が素早く発達する)[6]．母子が同時に発声する時間的規則が，発達に伴い変化するという報告も，複数のサル種で存在する [7, 8]．このように，発声者間のやり取りの頻度，すなわち,「会話頻度」，は，霊長類の社会性を評価するような良い指標の一つであると言える．一方で，このような「会話頻度」が十分に調べられてきたとは言えないのも事実である．一連の先行研究の背景には，発声者情報の特定が必要不可欠な技術となる．しかし，これには方法的に極めて困難な作業であった．例えば，野生下の研究では，発声者の特定は観察者の技術と判断に依存する．典型的でわかりやすく，目視で確認もできる状況でのみ，２者間での発声やり取りを記載できるが，多くの観察データは発声者特定が困難であり，分析に利用できるデータは限られる．飼育下でのやり取りについては，マイクロフォンを首輪などの形で装着させる方法で特定することが手段の一つとして考えられるが，動物への負担などの面で，必ず採用できるとは限らない．音声コミュニケーションに基づく，社会ネットワークの評価は，サルの社会維持や音声発達という面で，重要性が認識されているが，方法の困難さから十分な評価がされていない．

# 3　応用問題としての類人猿歌の音源位置特定

発声者を特定し会話を可視化することは，重要な問題でありながら，未だ十分な形で取り組まれているとは言えない．サルの群れ社会での音声が，彼らの空間移動や集団凝集性と関わる媒体として生態学的に機能しているのであれば，群れ内と群れ間での音声コミュニケーションの全体像を把握することは，彼らの集団行動を理解する上で極めて重要であろう．

そこで，テナガザル類の歌に着目した我々の近年の取り組みについて紹介する．主に，マイクロフォンアレイによる録音とロボット聴覚技術 HARK(Honda Research Institute Japan Audition for Robots with Kyoto University)[9, 10] を用いた，テナガザル歌の音源定位問題についての予備的な報告をする．

## 3.1　生物学的背景

テナガザルが会話可視化の良い例題として選んだ理由となった，いくつかの行動学的な特徴について述べる．テナガザルは，主に東南アジアに生息するサルの仲間である．ヒトに近縁なサル類として類人猿が知られているが，テナガザルも類人猿に分類される [11]．ただし，チンパンジーやゴリラ，オランウータンといった大型類人猿とは異なり，体重は 5kg から 10kg 程度と小柄なことから，小型類人猿と言われる．小柄な体格は，樹上生活によく適応している．ほとんどの時間を熱帯雨林の樹上 30m の場所で暮らし，枝から枝に腕渡りしながら落下せず暮らしている．社会も特徴的である．群れの中心は，オトナメスとオトナオスの夫婦であり，その夫婦で 1 $km^2$ 程度の広さの縄張りを防衛する．その夫婦には，２年から３年おきにコドモを１頭が生まれる．コドモは８歳ごろまでは，その家族で暮らし，性成熟後群れを移出し，新しい家族を形成していく．このように，夫婦に数頭のコドモが共にある，核家族を形成する．家族関係は，敵対的とされる．つまり，縄張りを家族間で防衛する．

彼らは,「歌」を歌うことで知られている [12]．ここでいうテナガザルの「歌」とは，連続的な発声のことである．多くのサル類の発声は,「単発」であり，連続的な発声はあまり見られない．しかし，テナガザルは，系列的な規則性を伴った発声をする (図 2)．この系列規則性は，種に特異的とされる．すなわち，種に応じて歌の系列的な特長があることが知られている．さらに，この歌の系列的な特長は，雌雄差があることがわかっている．すなわち，オスの歌とメスの歌が存在する．核家族を形成するテナガザルの家族の中で，夫婦ではデュエットとしてやり取りをすることが知られている．夫婦間のデュエットの生態学的な機能としては，二つ提唱されている．一つは，夫婦間の絆の強化とされる．よく唱和することで，夫婦の関係性を強化すると考えれている．もう一つは，縄張り防衛機能である．森林内では，複数の家族が連続的に暮らしている状況となる．家族は縄張りを持つが，お互いの家族デュエットを歌い合い，縄張りの防衛に役立っていると考えらえている．このデュエットは，多くのテナガザルで毎朝早朝に見られる．大変興味深いことに，歌は一旦開始されると２時間以上継続することが多い．さらに，歌は群れの縄張り防衛の役割を果たすことからも，音は 2 $km$ 程度と，かなり長距離まで到達する．まとめると，以下のような特徴があると言える．

- 歌は家族単位で長時間連続的に鳴く．
- 歌はオスとメスで特徴があり音響的に区別できる．
- 歌は種によって特徴が異なり歌の特徴から種を区別できる．
- 歌は２キロ程度届くため，遠くからの歌による観測が可能である．

こうした，生物学的な特徴は音源定位を扱う問題として，大変有望な可能性を有していると考えられた．そこで，我々は，まずテナガザルが飼育されている集団ケージを対象に，歌の可視化可能性について検討した．

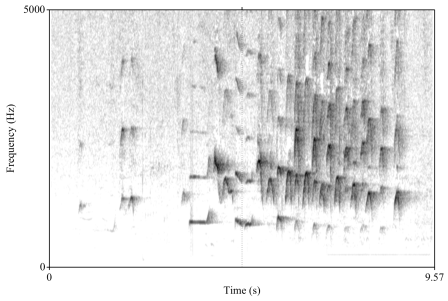



図 2: テナガザルの歌のサウンドスペクトログラム．図は，テナガザルの一種であるボウシテナガザルのオスの歌．こうした歌声が，長時間にわたって観察できる．

### 3.2 対象

対象は，タイ国 Krabok Koo Wildlife Breeding Center(以下，センターとする) で飼育されているテナガザルを対象に実施した．録音は，2019 年 1 月に実施した．

センターでは，3m x 10m x 5m 程度のケージに，一つの家族が飼育されるケージを配置してテナガザルを飼育していた．そして，そのケージは，10 個が連続した形で配置され，それぞれに家族 (1-4 頭が暮らしている) が飼育されていた．そこで，10 連ケージを一つを対象し，歌の発声時の音源定位を試みた (図 3)．

図 3: 10 連ケージを正面から撮影．正面に，マイクロフォンアレイが三脚にのせて設置されている．

### 3.3 装置

音源定位に利用する録音には市販されている 8ch マイクロフォンアレイ (TAMAGO-03, System-In-Frontier Inc. Tokyo, Japan) を利用した．図のように，10 連ケージに対し，3 台を挟み込むように配置した (図 4)．録音には Raspberry Pi 3 を利用し，スマートフォンの Wi-Fi テザリングを利用して簡易ネットワーク環境を構築した上で，時刻同期と録音制御を ssh を経由して制御・実施した．

### 3.4 予備的分析

図 5 は，10 連ケージの個体の鳴き声を定位するために設置した，同じ側にある 2 つのマイクロフォンアレイの録音を用いて予備的分析を行った例である．2019 年 1 月に行った録音のうちの一日分，早朝からの約 8 時間の録音を採用した．

図 5(a) は片方のマイクロフォンアレイ（図中印のついたもの）を用いて音源到来方向を推定した例であり，図 5(b) は，そのうち約 2 分のスペクトログラム，

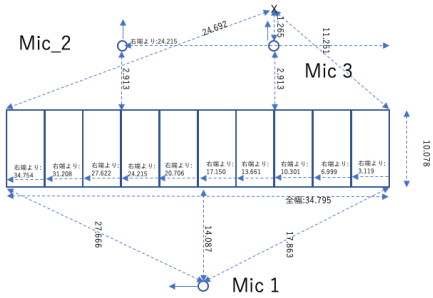



図 4: 10 連ケージとマイク配置の位置関係の模式図．レーサー測位により，マイク位置とケージのは位置関係を測位した．単位は $m$.

MUSIC スペクトル，定位音源情報（矩形）を示している．MUSIC スペクトルの算出には鳥類の鳴き声音源定位ツール HARKBird[13] を利用しテナガザルの鳴き声のみを抽出しやすいように 500-1800Hz の周波数域を対象とした．音源定位情報は，各時刻におけるスペクトルのピーク（図中の点）をクラスタリングして算出した．これは，現環境では MUSIC スペクトルのノイズや揺れが大きいため，時間と方位の空間においてばらつきながらも集中して存在するピークを音源として抽出することを狙った．その結果，同図下段のように，異なる方位で交互に定位される音源が抽出され，そのパターンが上段のスペクトログラムにおける鳴き声の周波数パターンと概ね対応する状況を確認した．これは，前述の個体間における会話を抽出できることを示唆している．同時に，図 5(a) からは，時刻によって複数の方向で音源が繰り返し定位され，上記のような会話が断続的に生じている可能性示している．

図 5(c) は，両マイクロフォンアレイが同時に定位した音源の方向から三点測量の要領で位置を定位できた音源を，時間経過に伴って色を変えながらプロットしたものである．ケージ内で音源が集中する領域が複数見られ，一部のケージ内の個体が頻繁に鳴き声を発したことが示唆される．その他，ケージを挟んで反対側に設置したマイクロフォンアレイからは，10 連ケージの個体に加え，隣接するケージの個体の鳴き声も観測され，ケージ間での相互作用の分析の可能性も示唆された．また，音源定位結果の重複状況から一個体による単独の鳴き声の抽出が可能であることも示唆された．しかし，ノイズの大きい環境である上に鳴き声が複雑であり，個体間の間隔もごく小さいため，高い精度で個体・集団レベルの鳴き声相互作用を抽出するには更なる工夫が必要な状況にあるといえる．

## 4 将来に向けた展望

飼育下とはいえ，東南アジア地域の不安定なネットワーク環境および電源環境でのマイクロフォンアレイによる録音は，様々な困難に直面する．短期間での実装には，安価なリチウムイオン電池やラズベリーパイに代表される簡便な録音装置開発環境の登場，またスマートフォンでも実装可能な簡易ネットワークなど技術面での運用のしやすさの向上が関与している．今回の取り組みでは，3 台の録音装置を用い，同期録音を試みたが，10 台程度の環境構築も不可能ではない．実際，筆

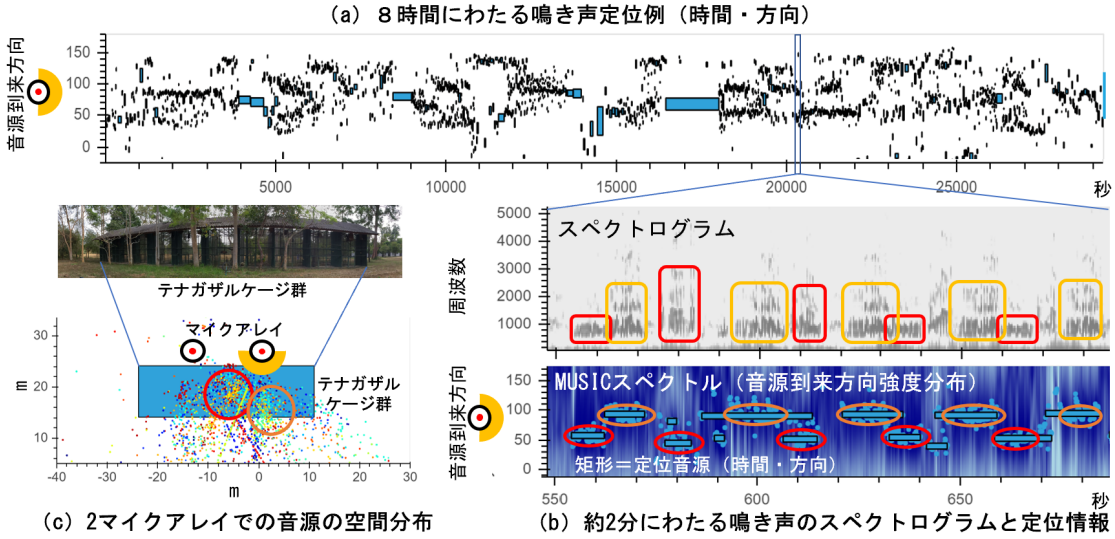



図 5: HARK で予測した音源到来方位を利用して 10 連ケージ内の個体の鳴き声を抽出した一例．予備的段階で多くの雑音等が含まれる状況に基づく結果だが，明瞭な歌信号の特定も数多く存在する．

者たちの研究グループは，10 台以上のマイクロフォンアレイによる同期録音を試み，一部データを得ている．ケージという場所の情報が明らかである環境でもあり，環境計測を合わせて進めることで，音源定位の精度向上，最終的にテナガザル家族間の会話の可視化の実現が期待できる．

さらに，テナガザル固有の問題にも，こうした音源定位技術は将来有望な技術の一つと認識している．筆者は 10 年ほど前に，インドネシア・スマトラ島にて，長期間に渡り，生態調査をおこなっていた．その調査では，毎朝早朝に，森に調査仲間とともに出かけ，尾根に登り歌の聞こえる方位と時刻を，GPS と録音機を用いて，ひたすら地図上に記録していた．観察者の計測位置を GPS により取得し，歌の聞こえる時刻・方位を時計とコンパスによって記録した．複数人の調査データ (歌の時刻・方位データ) を持ち合わせ，地図上でその音源を定位し，日々テナガザル家族の「歌地図」を手で作成していた (図 6)．歌の一記録から縄張り面積の推定などを実施したが，以前，こうした作業の自動化は遠い将来と考えていた．技術向上とりわけ環境構築上の簡易化と安価は，テナガザルという絶滅危惧種の社会と生態を解き明かす手段として有望であろうと期待される．

## 謝辞

本稿執筆にあたり，文部科学省・新学術領域研究「共創言語進化」(#4903, 17H06380, 17H06383)，および JST 戦略的創造研究推進事業 (CREST)「人間と情報環境の共生インタラクション基盤技術の創出と展開」(「脳領域／個体／集団間のインタラクション創発原理の解明と適用」17941861 #JPMJCR17A4) からの支援を一部いただいた．また，京都大学の森田尭博士と中部大学松田一希博士から様々な助言をいただいた．

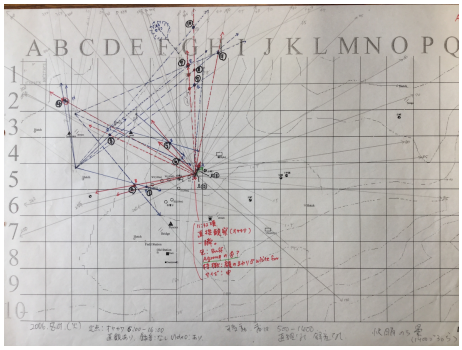


図 6: 筆者らの研究チームで自作した地図の上に，音の定位の記録を描き，交点を求め音源場所を推定している．当時は，自由に利用可能な衛星地図も少なく，位置計測などは研究者自身で実施した．図の中で見られるたくさんの線分は，音が聞こえた「方位」を書き込んだものである．

## 参考文献

[1] Stephen C Levinson. Turn-taking in human communication - origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14, jan 2016.

[2] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, dec 1974.

[3] Nobuo Masataka and Maxeen Biben. Temporal rules regulating affiliative vocal rxchanges of squirrel monkeys. *Behaviour*, 101(4):311–319, jan 1987.

[4] Masazumi Mitani. Voiceprint identification and its application to sociological studies of wild

Japanese monkeys (Macaca fuscata yakui). *Primates*, 27(4):397–412, oct 1986.

[5] A.Margaret Elowson and Charles T. Snowdon. Pygmy marmosets, Cebuella pygmaea, modify vocal structure in response to changed social environment. *Animal Behaviour*, 47(6):1267–1277, jun 1994.

[6] D Y Takahashi, A R Fenley, Y Teramoto, D Z Narayanan, J I Borjon, P Holmes, and A A Ghazanfar. The developmental dynamics of marmoset monkey vocal production. *Science*, 349(6249):734–738, 2015.

[7] Hiroki Koda, Alban Lemasson, Chisako Oyakawa, Rizaldi, Joko Pamungkas, and Nobuo Masataka. Possible role of mother-daughter vocal interactions on the development of species-specific song in gibbons. *PLoS ONE*, 8(8):e71432, aug 2013.

[8] A. Lemasson, L. Glas, S. Barbu, A. Lacroix, M. Guilloux, K. Remeuf, and H. Koda. Youngsters do not pay attention to conversational rules: is this so for nonhuman primates? *Scientific Reports*, 1(1):22, dec 2011.

[9] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Design and implementation of robot audition system 'HARK' - Open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, jan 2010.

[10] Kazuhiro Nakadai, Hiroshi G. Okuno and Takeshi Mizumoto. Development, deployment and applications of robot audition open source software HARK. *Journal of Robotics and Mechatronics*, 27:16–25, Feb 2017.

[11] Susan Lappan and Danielle June Whittaker. *The gibbons: new perspectives on small ape socioecology and population biology*. Springer, 2009.

[12] Hiroki Koda. Gibbon songs: Understanding the evolution and development of this unique form of vocal communication. In Ulrich H. Reichard, Hirohisa Hirai, and Claudia Barelli, editors, *Evolution of Gibbons and Siamang*, pages 347–357. Springer, 2016.

[13] Shinji Sumitani, Reiji Suzuki, Naoaki Chiba, Shiho Matsubayashi, Takaya Arita, Kazuhiro Nakadai and Hiroshi G. Okuno. An integrated framework for field recording, localization, classification and annotation of birdsongs using robot audition techniques – HARKBird 2.0. In *Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019)*, pages 8246-8250. IEEE, 2019.

一般社団法人 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-055-2
(11/22)

# 生成モデルに基づく鳴き声を用いた 鳥類に対するプレイバック実験の試行

# A playback experiment on songbirds using simulated vocalizations based on a generative model

炭谷晋司[1*]　鈴木麗璽[1]　松林志保[2]　有田隆也[1]　中臺一博[3,4]　奥乃博[5]
Shinji Sumitani[1]　Reiji Suzuki[1]　Shiho Matsubayashi[2]　Takaya Arita[3]
Kazuhiro Nakadai[3,4]　Hiroshi G. Okuno[5]

[1] 名古屋大学，Nagoya University
[2] 大阪大学, Osaka University
[3] 東京工業大学，Tokyo Institute of Technology
[4] （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
[5] 早稲田大学，Waseda University

**Abstract:** 本稿は，鳥類の音声コミュニケーションにおける，音響特性が与える影響をより理解するための実験の枠組みの構築を目的とする．予備実験として，変分オートエンコーダ（Variational Autoencoder, VAE）で学習したウグイスの 2 種類（H・L 型）の歌の特徴空間を利用して，両者および中間の歌を生成し，それらの歌を用いたプレイバック実験を行い，複数のマイクアレイによる 2 次元音源定位で行動パターンを分析した．その結果，生成音に対しても個体は反応を示し，歌として認識することが示唆された．

## 1　はじめに

鳥類にとって歌（さえずり）は，縄張りの主張や求愛などに用いられる重要な意思伝達手段である [1]．歌行動に基づく種間・個体間相互作用は，歌うタイミングなどの時間的相互作用，なわばり関係等の空間的相互作用，音の種類や周波数の特徴などの音響的相互作用といった様々な次元を持つ．そのため，この理解は，生物由来の音と周囲の環境との関係を様々な次元で理解を試みる生態音響学 [2] では重要な主題とされており，我々は特に歌う鳥の集団を歌を介して相互作用する複雑系とみなして理解を試みている [14].

我々は，鳥類の歌行動理解に活用することを目的として，ロボット聴覚オープンソースソフトウェア HARK[8] と市販のマイクアレイで構築される，録音分析システム HARKBird[15] を構築し，鳥類の歌行動のタイミング，方向（位置），および，音源の自動抽出の試行や．その応用や発展の可能性について検討してきた [16, 17, 12]. 現在では，機械学習を利用した音源の分類ツールが搭載され，より容易に音源の分類が可能となっている [13].

その中でも，鳴き声の種類に基づく音響的相互作用により焦点を合わせた研究として，スピーカから鳥類の歌を再生し，再生音が鳥類個体のに与える影響を調査するプレイバック実験を行ってきた．特に，ウグイスに対して，同種の持つ 2 種の歌（H型：求愛やなわばりの宣言，L型：警戒）をタイミングを変更して様々な条件でプレイバック実験を行い，その時の注目個体が歌を発した方向を単一のマイクアレイを用いて分析を行った [17]．その結果，再生条件に応じて注目個体の歌の頻度やH型・L型の割合，マイクから見た個体の方向の変化の傾向に違いがあることが定量的に示された．これは，従来の人手による観測では容易でなかった，多様な次元の詳細な歌行動傾向を把握可能なことを示唆している．

一方，近年では機械学習を用いて鳥類の歌行動を理解する試みがあり，次元削減アルゴリズムを用いた鳥類の歌のシーケンスに関する調査手法や，生成モデルに基づいた歌コミュニケーションにおける知覚や行動に関する調査手法が提案されるなど，機械学習を用いた鳥類の音響的相互作用の理解が注目されている [10]. 特に，生成モデルに関しては，潜在空間から新しい歌構造，音素構造を生成できるため，プレイバック実験に盛り込むことで歌構造の役割や，その適応性に関してもさらに深く理解できることが期待できる．

*連絡先： 名古屋大学大学院情報学研究科
〒 464-8601 愛知県名古屋市千種区不老町
E-mail:sumitani@alife.cs.is.nagoya-u.ac.jp

本研究は，鳥類の音声コミュニケーションにおける時間的・空間的・音響的関係を詳細に抽出可能な観測実験分析フレームワークをロボット聴覚技術と機械学習を融合して実現することを目的とする．本稿では，その中でも，上記の生成モデルを利用した歌の音響特性が与える影響をより理解するための実験の枠組みの構築を検討する．具体的には，生成モデルの 1 つである，変分オートエンコーダ（Variational Autoencoder, VAE）[6] を用いて，ウグイス *Horornis diphone* の 2 種類（H・L 型）の歌を学習し，その潜在空間から人工的な歌の生成を行った．その歌を利用してプレイバック実験を行い，まずウグイス個体は生成音を歌として認識しうるか，認識した場合，その中でも役割に違いのある H 型と L 型の中間の合成（混合音）にどのような反応を示すかについて調査した．それぞれのプレイバック実験では，複数のマイクアレイによる録音を行い，2 次元での音源定位によりプレイバックの影響を調査した．

# 2 手法

## 2.1 対象種

ウグイス *Horornis diphone* は,「ホーホケキョ」と聞こえる高いピッチで歌う H 型と「ホーホホホケキョ」などとホーの部分が断続する低いピッチで歌う L 型の 2 種の歌を持つ [3]．スピーカから同種の歌のプレイバックを行い，その間に行われたウグイス個体のさえずり回数と種類の計測を行った百瀬の実験では，プレイバックのある間は再生を行っていない期間と比較して H 型の頻度が減少し，L 型の割合が増加することを人手による観測結果から示した．このことから，H 型には縄張りの主張や求愛，L 型には近隣個体への威嚇の意味があるとされている [7]．前述の我々のウグイスに対するプレイバック実験 [17] は，百瀬の実験結果と一致した上に，移動に関しては，L 型の割合が高いほど移動の頻度が高く，スピーカ上を通過するような飛び方を頻繁に行うこと，距離の大きな移動後には L 型を歌う傾向があることを示した．さらに，人工物と生物の相互作用の調査のために行ったボーカロイド初音ミクによって作成した L 型の真似音をプレイバックした結果，ウグイス個体は再生音に興味を示し，スピーカ上を飛び回るような挙動を示した [20]．

上記のように，ウグイスは 2 種の歌をうまく使い分けることで，意思伝達を行っている．従来，歌の構造の地域差等に関する議論はなされてきた [4, 5] が，音響特徴から見た 2 種の歌の間の関係やその適応的意義に関しては議論の余地があると考えられる，生成モデルによって H 型，L 型の中間音を作成することで，両者の関係や役割の理解に貢献しうると期待し実験を行った．

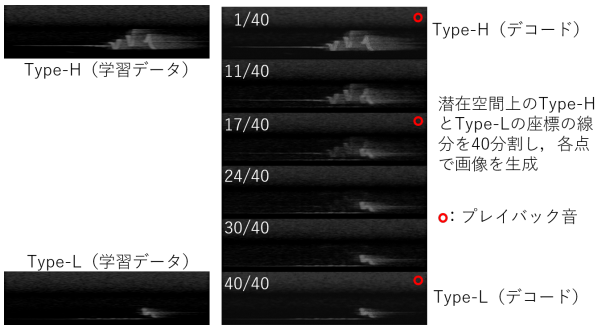



図 1: VAE 潜在空間から生成した歌

## 2.2 VAE による歌の生成

VAE[6] は，潜在変数が正規分布に従うように学習させるオートエンコーダである．その特徴から，デコーダ部分で生成される出力は潜在空間上で連続的な変化を伴うため，イメージの合成手法として応用される．今回は，ウグイスの H 型，L 型の歌それぞれの録音から $492 \times 128$ のグレースケールのスペクトログラム画像を作成し，それをデータセットとして学習を行った．得られた特徴空間から H 型，L 型および複合音の画像を生成し，その画像を再度音声信号に変換することで人工的な歌を生成した．

学習に用いた歌は，今回の実験対象となるウグイスとは異なるウグイス個体の H 型，L 型である．今回は，先行研究 [17] との知見の比較を念頭に，そこで実際にプレイバックに使用した H 型，L 型の歌それぞれ 1 個ずつからなる最小のデータセットを用いた．これは，生成モデルとしてはやや例外な手法であるが，予備的知見として，デコード・生成処理を経た上記のプレイバック音をウグイス個体が歌として認識するかを確かめたいということと，使用した歌の特徴空間において 2 種の中間の歌を容易に選択し生成可能ということによる．

ネットワーク構造には，エンコード側は 7 層のフィルタサイズ $3 \times 3$ の畳み込み層と 2 層の全結合層，デコード側はエンコード側と計算が逆となる同じ数の層を用いた．図 1 に学習に用いた 2 種のデータと，生成した画像を示す．生成した画像は H 型から L 型にかけて徐々に変化する様子が確認できる．

## 2.3 プレイバック実験

実験は，2019 年 5 月 4 日の午前中，名古屋大学大学院生命農学研究科付属フィールド科学教育研究センター稲武フィールド（愛知県豊田市稲武町）の森林で行った．予めウグイス 1 個体が鳴く場所を調査し，その場所にノート PC（TOUGHBOOK CF-C2; Panasonic）に

USB 接続した 2 つのマイクロホンアレイ（TAMAGO-03; System in Frontier Inc.）を三脚に固定し，約 7m ほど離して配置した．また，スピーカ（Mega; Tronsmart）は各マイクアレイからそれぞれ 7m，5m ほど離れた位置にある地上 2m 付近の枝の上に配置した（図 2）．


図 2: 実験風景

プレイバック音には，前節の VAE より生成した H 型，L 型および H 型と混合音 1 種を用いた（図 1 参照）．混合音に関しては，学習に用いた H 型，L 型のスペクトログラム画像をエンコードしたときに得られる H 型，L 型それぞれの潜在空間の座標間の線分を 40 分割し，H 型から 17 番目の座標を用いて生成した画像を音声に変換したものを用いた．これは実際に聞くことにより体感的に中間らしい音を選んだ．プレイバックの方法としては，各再生音を 2 度再生し，2 分間のインターバルを繰り返す方法を採用し，混合音（M），H 型，L 型，混合音，プレイバックなしの録音の順でそれぞれ 30 分間の実験を連続して行った．前のプレイバック実験や人の接近によるウグイス個体への影響を考慮するために，実験の間で数分程度間隔を開け，また，プレイバックを 30 分よりやや長めに行った．分析データは録音開始後 5 分からの 30 分間のデータを採用した

### 2.4 分析

まず，HARKBird を用いてウグイス個体の歌をうまく定位するように定位のパラメータを適宜調整し，音源の定位・分離を行った．短時間フーリエ変換によって得られた各チャネルのスペクトログラムから MUSIC 法 [11] を用いて音源定位を行い，その定位結果に基づいて GHDSS 法 [9] を用いて対応する音源を抽出した．その後，抽出した分離音源を 100×64 のグレースケール画像に変換し，t-SNE（t-distributed Stochastic Neighbor Embedding）[18] により 2 次元まで次元の圧縮を行い，得られた特徴空間を用いてプレイバック音およびウグイス個体の歌を分類した．分類した結果をもとに，同じラベルで分類された各マイクアレイの定位音源のペアで Sumitani らの MUSIC スペクトルに基づく三角測

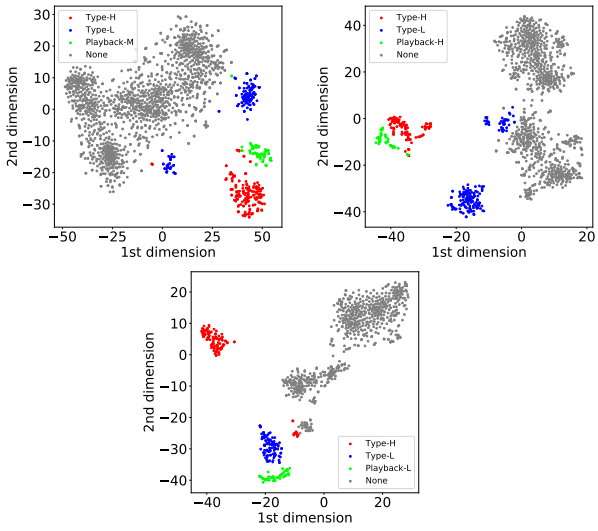



図 3: t-SNE 特徴空間上での音源分離（左上：混合音再生時，右上：H 型再生時，下：L 型再生時

量の手法 [12] で二次元定位を行い，移動距離や歌の頻度等の定量的観測を試みた．2 回目の混合音を用いたプレイバック実験では，ウグイス個体の歌う位置が，2 つのマイクアレイの直線状付近で二次元定位が困難であり，またプレイバックなしの録音ではほとんどウグイス個体が鳴いていなかったため，今回はそれらの結果の報告を割愛する．

## 3 結果

まず，各実験において定位（到来方向推定）された音源の音響的特徴とその分布について述べる．図 3 は各実験条件における，音源定位・分離結果に基づいて t-SNE[18] を実行し，その特徴空間上でウグイス個体の歌（H 型，L 型）とプレイバック音を抽出した結果である．多くの注目個体の歌やプレイバック音は，他の定位音源とは独立してクラスタを形成し同じ音同士で集まり分布した．また，プレイバック音に着目してみると，H 型のプレイバック音は個体の H 型近傍に，L 型のプレイバック音は個体の L 型近傍に，複合音は個体の H 型，L 型の中央付近に分布しており，構造間の関係をゆるやかに反映した分布であることが確認できる．“None” としてラベル付けされている音源の多くは，システム配置場所付近にあった 2 つの小川を流れる水音が占めていた．

次に，分類結果を用いて 2 次元定位を行った．図 4 は，各再生音をプレイバックしたときのウグイス個体の歌とプレイバック音の 2 次元定位結果を示す．概して，定位結果はシステム周りに集中しており，ウグイス個体はシステム周りでスピーカ音を警戒して動いて

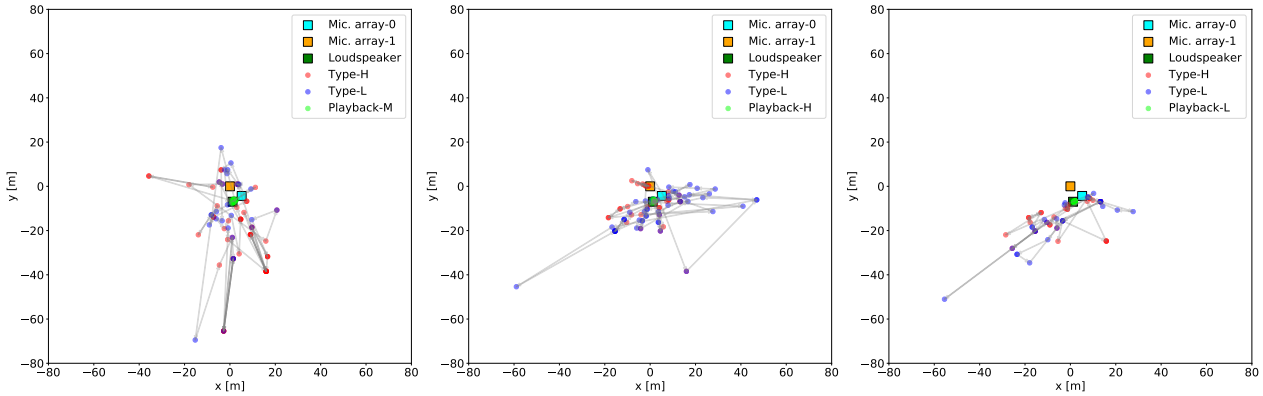



図 4: 2 次元定位結果（左：混合音再生時，中央：H 型再生時，右：L 型再生時）ウグイス個体の歌（赤点：H 型，青点：L 型）を繋ぐ矢印は，定位できたウグイス個体の歌を時系列順に繋いだもの

表 1: 各実験条件でのウグイス個体の歌行動．下線はプレイバックに対して最も警戒した反応（最高 L 型頻度・最低 H 型割合・最近平均距離）を示したと考えられる挙動．

| 再生音 | Mic. array | H 型 | L 型 | 合計 | H 型の割合 H/(H+L) | 2 次元定位数 All (H, L) | スピーカから歌定位位置までの平均距離 All（H，L）[m] |
|---|---|---|---|---|---|---|---|
| 混合歌 | 0 | 83 | 60 | 143 | 0.580 | 88 (60, 28) | 19.737（21.233，16.531） |
| | 1 | 80 | 67 | 147 | 0.544 | | |
| H 型 | 0 | 56 | 81 | 137 | 0.409 | 98 (37, 61) | 12.907（10.313，14.481） |
| | 1 | 49 | 93 | 142 | 0.345 | | |
| L 型 | 0 | 60 | 49 | 109 | 0.550 | 65 (34, 31) | 16.098（17.928，14.430） |
| | 1 | 51 | 49 | 100 | 0.510 | | |

いる様子が確認できる．先行研究 [17] において，プレイバックのない録音のみの分析では，ウグイス個体は樹上等のいくつかの場所で留まりくりかえし歌行動を行う傾向があり，プレイバックのある場合は頻繁にスピーカ周りを移動する傾向があった．このことを考慮すれば，ウグイス個体は VAE により生成された歌を種の歌として認識していると考えられる．

なお，全ての実験条件において，スピーカから大きく離れて定位されている定位音がいくつかあるが，これらの音源は全て注目したウグイス個体のものであった．それぞれのマイクアレイの定位結果を確認したところ，分離音源ははっきりとした波形であり，システムからそう遠くない場所で歌われた歌であることが考えられる．そのため，これらの結果は定位誤差により生じた例外である可能性がある．また，それらは図中下方に多く存在するため，その方面の環境の音響特性の影響を受けた可能性もある．

表 1 に，各実験条件でのウグイス個体の歌を定位した回数をまとめた．各マイクアレイで定位した歌の総数は，混合音の再生実験ではマイクアレイ 0，1 でそれぞれ 143 回，147 回の定位数，H 型再生時には，137 回，142 回，L 型再生時には，109 回，100 回と，実験が後になっていくにつれて減少傾向にあった．これは，実験を連続して行ったために，いわゆる学習効果によって，ウグイス個体がプレイバックに慣れていったことが考えられる．あるいは鳥類は朝方に頻繁に鳴き，次第に静かになる傾向があるため，実験対象としたウグイス個体に関しても歌う頻度が減少したことも考えられる．

各条件で歌った H 型と L 型の割合に着目みてみると，混合歌を再生した場合には，各マイクアレイの結果としてそれぞれ 0.580，0.544，H 型を再生した場合は 0.409，0.345，L 型を再生した場合は 0.550，0.510 であった．これらの結果は，我々の先行研究におけるプレイバックのない録音でのウグイス個体が歌った H 型の割合と比較しても小さい値であり，また，プレイバック実験時の割合に比較的似ていた．以上のことから，歌の割合からもウグイス個体がプレイバック音に対して反応を示しているといえる．

一方で，時間的学習効果を考慮すれば，冒頭に行った混合歌再生時の方がより警戒を示し，H 型の割合はより小さくなることが期待されるが，H 型，もしくは，L 型再生時における H 型の割合の方が小さかった．これは，混合歌は，H・L 型と比べて反応が低いことを示していると考えられる．

2 次元定位数は，各マイクアレイで定位できた音源数に比較すると数を減らした．それぞれのマイクアレイの音源定位結果を詳細に調べた結果，以下の原因によることが確認された：1．どちらか片側のマイクアレイで歌が定位できておらず 2 次元定位に至らなかった．

2．ペアとなる音源はあるが，各マイクアレイの定位方向が平行か、あるいは定位方向の延長線が広がる向きで定位しているため，交点ができなかった．これらは，実験場所が木々の多く生い茂る地点であり，かつ近隣に 2 つの河川が流れる地点であったため，音源定位がそれらの反響や雑音の影響を大きく受けたため起きてしまったと考えられる．また，二次元定位が困難な場所である，マイクアレイを繋ぐ直線上付近にウグイス個体のよく鳴く位置があったことも原因として挙げられる．

さらに，ウグイス個体が空間的にどのように歌っていたかを調査するために，各条件においてウグイス個体が歌った H 型，L 型あるいは双方における，スピーカから歌の定位位置までの平均距離を求めた．3 つの条件を比較すると，H 型，L 型合計での平均距離は，混合音再生時（19.737 m），L 型再生時（16.098 m），H 型再生時（12.907 m）の順であった．これは，H 型の割合が大きい順と一致しており，混合歌再生時のプレイバック音に対する警戒の度合いの小ささを反映することが示唆される．また，これらの距離の違いは特に H 型を歌う場合に顕著みられ，L 型には差が大きくない．H 型は縄張りの宣言や求愛の意味を持つため，不特定多数に対する歌であり，一方で L 型は近隣個体への威嚇の意味を持ち，その相手個体に対して発する歌であるとされている．そのため，固定されたスピーカに対しては L 型は同等の距離を保ちがちである一方，H 型の歌はより広い範囲で歌われた結果，歌う位置に差がみられたことが考えられる．

以上から，今回再生した生成音 3 種に対してウグイスはいずれも明らかな反応を示したが，H 型と L 型の混合音に対しては最も反応が弱く，遠くで H 型の歌を歌いがちであることが推測された．混合音は対象個体にとっては同種に近いが聞きなれない歌であるため，警戒しつつも戸惑っている状況であった可能性もあると考えている．

## 4　おわりに

音響特性が与える影響をより理解するための実験の枠組みの構築検討のために，ウグイスの 2 種類（H・L 型）の歌を VAE で学習し，その潜在空間から人工的な歌の生成を試み，その歌を利用してプレイバック実験を行った．その結果，ウグイス個体は，H 型，L 型の混合音も歌として認識することがわかった．その影響の理解にはより詳細な分析が必要であるが，生成モデルから生成した歌がプレイバック実験等の鳥類の歌コミュニケーションにおける音響的相互作用理解に利用可能であることを示した．今回は，2 種の歌の生成と，その中間音の生成のみの検討であったが，VAE の潜在空間ではある特徴を強調して生成することも可能であるため [19]，例えばウグイスに関していえば，H 型，L 型の特徴を強調した歌の生成も可能であると考える．現在，大きなデータセットを用いた学習も進めており，より VAE の特性を活かした，歌構造の役割やその適応性に関する知見を得たいと考えている．

## 謝辞

高部直紀氏（名古屋大学）のフィールド調査への協力と，Zhao Hao 氏（名古屋大学）のデータ分析への協力に謝意を表す．本研究の一部は JSPS 科研費 JP18K11467，JP19KK0260，JP17H06383（#4903）の助成を受けた．

## 参考文献

[1] C. K. Catchpole and P. J. B. Slater, *Bird Song: Biological Themes and Variations*. Cambridge University Press, 2008.

[2] A. Farina and S. H. Gage, *Ecoacoustics: The Ecological Role of Sounds*. John Wiley and Sons, 2017.

[3] S. Hamao, “Japanese bush warbler,” *Bird Research News*, vol. 4, no. 2, 2007.

[4] S. Hamao, “Acoustic structure of songs in island populations of the japanese bush warbler, cettia diphone, in relation to sexual selection,” *Journal of Ethology*, vol. 31, no. 1, pp. 9–15, Jan 2013. [Online]. Available: https://doi.org/10.1007/s10164-012-0341-1

[5] S. Hamao, “Rapid change in song structure in introduced japanese bush-warblers (cettia diphone) in hawai‘i,” *Pacific Science*, vol. 69, no. 1, pp. 59 – 66, 2015. [Online]. Available: https://doi.org/10.2984/69.1.4

[6] D. P. Kingma and M. Welling, “An introduction to variational autoencoders,” *CoRR*, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691

[7] H. Momose, *Mechanism of maintaining territories by acoustic communication. Reproductive strategies of birds*. Toukaidaigaku-shuppankai, Tokyo, Japan, 2016.

[8] K. Nakadai, H. G. Okuno, and T. Mizumoto, “Development, Deployment and Applications of Robot Audition Open Source Software HARK,” *Journal of Robotics and Mechatronics*, vol. 27, pp. 16–25, 2017.

[9] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, “Blind source separation with parameter-free adaptive step-size method for robot audition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1476–1485, 2010.

[10] T. Sainburg, M. Thielk, and T. Gentner, “Animal vocalization generative network (AVGN): A method for visualizing, understanding, and sampling from animal communicative repertoires,” in *Proceedings of the 41th Annual Meeting of the Cognitive Science Society*, 2019, p. 3563.

[11] R. Schmidt, “Bayesian nonparametrics for microphone array processing,” *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[12] S. Sumitani, R. Suzuki, S. Matsubayashi, K. Arita, T. Nakadai, and H. G. Okuno, “Extracting the relationship between the spatial distribution and types of bird vocalizations using robot audition system hark,” in *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 2485–2490.

[13] S. Sumitani, R. Suzuki, S. Matsubayashi, K. Arita, T. Nakadai, and H. G. Okuno, “An integrated framework for field recording, localization, classification and annotation of birdsongs using robot audition techniques - harkbird 2.0,” in *Proceedings of ICASSP 2019*, 2019, pp. 8246–8250.

[14] R. Suzuki and M. Cody, “Complex systems approaches to temporal soundspace partitioning in bird communities as a self-organizing phenomenon based on behavioral plasticity,” *Artificial Life and Robotics*, pp. 1–6, 09 2019.

[15] R. Suzuki, S. Matsubayashi, K. Nakadai, and H. G. Okuno, “HARKBird: Exploring acoustic interactions in bird communities using a microphone array,” *Journal of Robotics and Mechatronics*, vol. 27, pp. 213–223, 2017.

[16] R. Suzuki, S. Matsubayashi, F. Saito, T. Murate, T. Masuda, Y. Yamamoto, R. Kojima, K. Nakadai, and H. G. Okuno, “A spatiotemporal analysis of acoustic interactions between great reed warblers (acrocephalus arundinaceus) using microphone arrays and robot audition software hark,” *Ecology and Evolution*, vol. 8, pp. 812–825.

[17] R. Suzuki, S. Sumitani, Naren, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno, “Field observations of ecoacoustic dynamics of a japanese bush warbler using an open-source software for robot audition hark,” *Journal of Ecoacoustics*, vol. 2, p. EYAJ46.

[18] L. van der Maaten and G. Hinton, “Viualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[19] T. White, “Sampling generative networks: Notes on a few effective techniques,” *CoRR*, vol. abs/1609.04468, 2016. [Online]. Available: http://arxiv.org/abs/1609.04468

[20] 炭谷晋司，松林志保，鈴木麗璽，“マイクロホンアレイを用いたプレイバック実験に基づくウグイスのさえずりの方向分布分析”，**日本鳥学会** *2016* **年度大会講演要旨集**，p. 124，2016.

一般社団法人 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-055-3
(11/22)

# 複数マイクロホンアレイにおける音源方向尤度に基づく三次元音源追跡

# Sound Source Tracking Based on Source Direction Likelihood from Multiple Microphone Arrays

山田 泰基[1*]　糸山 克寿[1]　西田 健次[1]　中臺 一博[1,2]
Taiki Yamada[1], Katsutoshi Itoyama[1], Kenji Nishida[1], Kazuhiro Nakadai[1,2]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[2] Honda Research Institute Japan Co.,Ltd.

**Abstract:** 本稿は, 複数のマイクロホンアレイを用いた三次元音源位置の推定および追跡手法を提案する. マイクロホンアレイによる信号処理は, 音源方向推定に用いられる技術として確立されている. しかし, マイクロホンアレイをロボットやドローンに搭載して災害救助する際など, 音源方向に併せて音源との距離の取得, 言い換えれば音源位置の取得が必要になる場合が存在する. そこで, 本稿では各マイクロホンアレイが算出した音源方向尤度を統合することで得られる音源位置尤度を定義し, 求めた音源位置尤度から音源位置の推定および追跡を行う. また, 数値シミュレーションを通じて提案手法の有効性を検証し, 三角測量的アプローチで音源位置を追跡する他手法との性能比較を行った. 提案推定手法は 5° 刻みの伝達関数を用いたとき, 20 m 離れた音源に対し, 平均推定誤差 2.45 m と, 三角測量を用いた音源追跡手法の平均推定誤差より 3.25 m 小さく追跡することができることが確認された.

## 1　はじめに

音環境理解の分野において, マイクロホンアレイを用いた音源定位は盛んに研究が行われており, 例えば, 災害現場で瓦礫に埋もれた要救助者探索などに応用が期待できる有用な技術である. 一般に, 単独マイクロホンアレイは音源方向の推定に用いられるが, 単独で用いる代わりに複数マイクロホンアレイを用いることで音源位置を推定する研究が盛んに行われている. 複数マイクロホンアレイを用いて音源を観測することで, 音源を多角的に観測することができ, 音源方向以上の情報が得られる上, 移動音源をリアルタイムで追跡することができるというメリットがある. 複数マイクロホンアレイで音源位置を推定する手法として, 複数マイクロホンアレイから得た音源方向を元に三角測量する Brandstein らの手法 [1] や Daniel らの手法 [2] が挙げられる. 三角測量的アプローチは音源が遠方にあるほど, 外れ値となるような三角測量点が現れ, 結果的に位置推定誤差が大きく現れてしまう問題がある. そこで, 山田らはこれまで, 三角測量を用いた音源位置推定に加え, Gaussian Sum Filter を用いて外れ値となるような三角測量点を無視するような音源追跡手法を提案し, 数値シミュレーションにおいて, Brandstein らの手法より低い位置推定誤差で推定することを確認した [3]. しかし, 音源方向推定段階で大きな誤差が発生した際に, 三角測量点が距離方向に大きなずれて発生し, 音源位置の追跡が困難になってしまうことが屋外実験を通じて分かった [4]. そこで, 本稿では, 音源方向推定を介さず, 各マイクロホンアレイで求められる音源方向尤度をもとに直接音源位置の尤度を求め, 音源位置推定を行う手法を提案する. これによって, 音源位置が遠方にあったり, 音源方向指定の分解能が低かったりしても, 音源位置の推定誤差, 特にマイクロホンアレイから音源への距離方向の推定誤差を抑えられることが期待できる.

本稿は, 2 節で提案手法を説明し, 3 節では数値シミュレーションによる提案手法の有効性の評価を行う. 最後に, 4 節にまとめ, 今後の課題を記す.

*連絡先： 東京工業大学
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: yamada@ra.sc.e.titech.ac.jp

## 2　音源尤度からの三次元音源追跡

本稿は, 複数マイクロホンアレイから算出した音源方向の尤度を元に, 音源位置の三次元位置の推定および追跡を行う. 一般に単独マイクロホンアレイは音源方向の推定に用いられるが, 複数マイクロホンアレイを用いれば, 各マイクロホンアレイの方向推定戦略を統合することで, 音源位置推定が可能となる. これまで山田らは, 複数のマイクロホンアレイが推定した方向より三角測量を行うことで音源位置を推定する手法を報告しているが, 三角測量を用いるアプローチは音源位置が遠方にある場合やマイクロホンアレイの伝達関数の分解能が低い場合に, 著しい推定誤差を起こしたり三角測量が出来なかったりする問題がある. そこで本稿では, 各マイクロホンアレイにて音源方向を推定するかわりに, 音源方向推定に用いられる音源方向尤度を三次元音源位置尤度に変換する処理を施すことで, 三角測量を介さない音源位置推定を実現し, 遠方音源に対する位置推定誤差の抑制を試みる.

### 2.1　問題設定

本稿では, 複数のマイクロホンアレイを用いて単音源を追跡することについて考える. マイクロホンアレイは $N$ 個存在し, それぞれ

$$\mathrm{MA}_1, \ldots, \mathrm{MA}_N$$

と定義する. 各マイクロホンアレイは三次元空間上を移動・回転をすることができ, あるマイクロホンアレイ $\mathrm{MA}_n$ の時刻 $t$ における状態を

$$\boldsymbol{m}_n(t) = [\boldsymbol{m}_{n,xyz}^T, \boldsymbol{m}_{n,\phi\theta\psi}^T]^T \tag{1}$$
$$\boldsymbol{m}_{n,xyz} = [x_n(t), y_n(t), z_n(t)]^T \tag{2}$$
$$\boldsymbol{m}_{n,\phi\theta\psi} = [\phi_n(t), \theta_n(t), \psi_n(t)]^T \tag{3}$$

とおき, 既知であるとする. $x_n(t), y_n(t), z_n(t)$ は $\mathrm{MA}_n$ の中心の 3 次元位置座標を指し, $\phi_n(t), \theta_n(t), \psi_n(t)$ はそれぞれ $\mathrm{MA}_n$ のロール, ピッチ, ヨー角を指す. また, 各マイクロホンアレイは $M$ 個のマイクロホンから構成されており, $\mathrm{MA}_n$ に収録される音響信号は時間領域で $\boldsymbol{s}_n(t) \in \mathbb{R}^M$ と記述する. 音源は一つのみ存在し, 点音源であると仮定する. この音源の三次元座標は,

$$\boldsymbol{e}(t) = [x_e(t), y_e(t), z_e(t)]^T \tag{4}$$

とする. 本稿で取り組む問題は, 各マイクロホンアレイ状態 $\boldsymbol{m}_n(t)$ と収録信号 $\boldsymbol{s}_n(t)$ から音源位置 $\boldsymbol{e}(t)$ を一定時間おきに推定することで, 音源軌跡を推定することである.

### 2.2　追跡手法説明

これまで, マイクロホンアレイを用いて音源方向を推定する際は, 各方向において音源が存在する尤もらしさ $P(\phi, \theta)$ ($\phi$は方位角, $\theta$は仰角) を算出し, 最も $P(\phi, \theta)$ が大きくなる方向が音源方向であるとしている. この各マイクロホンアレイで得られる方向尤度 $P(\phi, \theta)$ を三次元位置に対する尤度に変換することで音源位置を推定することが提案手法のアプローチである. 本手法は $T_k$ おきに音源位置を推定する. つまり, $T_k$ 秒おきに推定音源軌跡を更新する. 三次元空間上にグリッド点を設け, 各グリッド点にて各マイクロホンアレイが算出する方向尤度 $P(\phi, \theta)$ を考慮に入れた評価関数を求めることで, 音源位置に対する尤度を表現する. 各行程の詳細は以下に記述する.

#### 2.2.1　音源方向尤度の算出

音源方向に対する尤度と見なせる指標は多数報告されている.　2 つのマイクロホンで TDOA (Time Difference Of Arrival) を推定する手法の一つである CSP 法 [5] で用いられる CSP 係数や, Delay-and-Sum ビームフォーマから求める空間スペクトルは, マイクロホンアレイから見た方向をパラメータに持つスカラー量であり, 一般に音源が存在する方向にピークが立つ性質を持つ [6]. 本稿では, 3 つ以上のマイクロホンで構成されるマイクロホンアレイを想定し, Delay-and-Sum ビームフォーマによる空間スペクトルより鋭いピークを音源方向に出す MUSIC スペクトルを音源方向尤度として用いることを考える. MUSIC 法 [7] とは, 空間相関行列が張る固有空間を解析手法であり, 目的音源の部分空間と雑音部分空間の直交性を用いて音源の方位・仰角を推定する手法である. 角周波数 $\omega$, 方位角 $\phi$, 仰角 $\theta$ の音源からマイクロホンアレイへの伝達関数を $\boldsymbol{a}(\omega, \phi, \theta)$ とすると, $(\phi, \theta)$ における空間スペクトル $P(\phi, \theta)$ は

$$P(\phi, \theta) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} \frac{\boldsymbol{a}(\phi, \theta)^H \boldsymbol{a}(\phi, \theta)}{\boldsymbol{a}(\phi, \theta)^H \boldsymbol{E}(\omega) \boldsymbol{E}(\omega)^H \boldsymbol{a}(\phi, \theta)} \tag{5}$$

で表せる. ただし, $\boldsymbol{E}$ は空間相関行列において雑音部分空間が張る固有ベクトル行列であり, $\omega_L, \omega_H$ はそれぞれ空間スペクトルの評価に用いる角周波数の下限と上限である. 空間スペクトル $P(\phi, \theta)$ は MUSIC スペクトルとも呼ばれ, 一般に方向推定をする際は, MUSIC スペクトルがピークを取る方向を推定方向とする. 本稿では $\mathrm{MA}_n$ で求めた MUSIC スペクトルを $P_n(\phi, \theta)$ と記述し, $\mathrm{MA}_n$ における音源方向に対する尤度であると見なす.

#### 2.2.2 音源位置尤度への変換

各マイクロホンアレイで算出した音源方向尤度 $P_n(\phi,\theta)$ を用いて, 音源位置尤度を表現することで, 音源位置を推定する. 具体的には, 想定される三次元空間をグリッド状に分割し, 各グリッド点ごとに音源位置尤度を求め, 最も音源位置尤度が大きいグリッド点を推定音源位置とする. あるグリッド点 $i$ の座標を

$$\boldsymbol{g}_i = [x_i, y_i, z_i]^T \tag{6}$$

とし, あるマイクロホンアレイ $\mathrm{MA}_n$ の中心からグリッド点 $i$ までの方位角, 仰角をそれぞれ $\phi_{ni}$, $\theta_{ni}$ とおく. 方向 $(\phi_{ni},\ \theta_{ni})$ は, ベクトル $\boldsymbol{m}_{n,xyz} - \boldsymbol{g}_i$ を極座標変換することで得られる. あるグリッド点 $i$ における音源位置尤度 $L(\boldsymbol{g}_i)$ は以下の式のように定義する.

$$L(\boldsymbol{g}_i) = \sum_n P(\tilde{\phi}_{ni}^{\mathrm{round}}, \tilde{\theta}_{ni}^{\mathrm{round}}) \tag{7}$$

$$\tilde{\phi}_{ni}^{\mathrm{round}} = \mathrm{round}(\tilde{\phi}_{ni}), \quad \tilde{\theta}_{ni}^{\mathrm{round}} = \mathrm{round}(\tilde{\theta}_{ni}) \tag{8}$$

$$\begin{bmatrix} \cos\tilde{\phi}_{ni}\cos\tilde{\theta}_{ni} \\ \sin\tilde{\phi}_{ni}\cos\tilde{\theta}_{ni} \\ \sin\tilde{\theta}_{ni} \end{bmatrix} = R_n^{-1} \begin{bmatrix} \cos\phi_{ni}\cos\theta_{ni} \\ \sin\phi_{ni}\cos\theta_{ni} \\ \sin\theta_{ni} \end{bmatrix} \tag{9}$$

ここで, round$(\cdot)$ は伝達関数 $\boldsymbol{a}(\omega,\phi,\theta)$ の方位角・仰角の分解能に合わせて方向 $(\phi_{ni},\theta_{ni})$ を丸める関数であり, $R_n$ は $\mathrm{MA}_n$ の姿勢を表す回転行列である. つまり, 各マイクロホンアレイから見たグリッド点 $i$ へのの方向を算出し, その各方向に対応する音源方向尤度を足し合わせた値を, グリッド点 $i$ の音源位置尤度としている.

#### 2.2.3 音源位置推定と追跡

全てのグリッド点において音源位置尤度を求めたあと, 最大値を取るグリッド点を推定音源位置とする. このように推定音源位置を求めると, 推定音源位置は必ずどれかのグリッド点上に求められる. よって, グリッド点間隔が大きい場合, 最終的に推定される音源軌跡は大きく振動するような軌跡になる. そこで, 移動平均フィルタやカルマンフィルタを用いてスムージングすることで, スムージングしない場合と比べてより良く音源軌跡を推定することができる. この効果は, 次節『数値シミュレーション』で確認できる.

## 3 数値シミュレーション

提案手法の有効性と性能を検証するために, MATLAB® を用いて数値シミュレーションを行った. また, 他の三次元追跡手法も MATLAB® で実装し, シミュレーション結果を比較することで提案手法の性能を評価する.

### 3.1 シミュレーション内容

図 1 のような三次元空間上を点音源が $z = 0$ で等速円運動をしているシナリオを考える. 点音源は以下のダイナミクスに従って等速円運動を行い, ホワイトノイズを出力している.

$$\boldsymbol{e}(0) = [5, 0, 0]^T \tag{10}$$

$$\dot{\boldsymbol{e}}(t) = \left[-\frac{\pi}{5}y_e(t),\ \frac{\pi}{5}x_e(t),\ 0\right]^T \tag{11}$$

音源位置は $T_k = 0.1$ 秒 おきに推定し, 計 10 秒間シミュレーションを行った. マイクロホンアレイは 4 つ用意し, 各マイクロホンアレイの状態は表 3 に記されている. マイクロホンアレイは 16 ch の球形マイクロホンアレイであり, 44.1 kHz, 24 bit で収録を行う. また, 三次元空間上の,

$$-6\mathrm{m} \le x \le 6\mathrm{m}$$
$$-6\mathrm{m} \le y \le 6\mathrm{m}$$
$$-5\mathrm{m} \le z \le 5\mathrm{m}$$

の範囲に, 0.1 m おきにグリッド点を定義し, 各グリッド点で音源位置尤度 $L(\boldsymbol{g}_i)$ を求めることで, 各時刻の音源位置を推定する. 本稿では, マイクロホンアレイ-音源間の距離, 伝達関数の刻みが各音源追跡手法に与える影響を考察するために, 音源との距離が 10 m の場合と 20 m の場合, 伝達関数が 1 ° 刻みの場合と 5 ° 刻みの場合と異なるシナリオを用意し, 計 4 種のシナリオに対してシミュレーションを行う.

また, 提案手法の性能を他の既存手法と比較するため, 複数マイクロホンアレイを用いて三角測量的に音源位置推定を行う LI (Linear Intersection) 法 [1] と, LI 法にガウス和フィルタを加えることで外れ値となるような三角測量点を無視する MT-GSFT (Multiple Triangulation and Gaussian Sum Filter Tracking) 法 [3] を実装し, シミュレーションを行った. また, 音源位置推定を行ったあと, スムージングをすることによる性能の変化を考察するため, 移動平均フィルタを用いて音源軌跡をスムージングする場合としない場合の両方についてシミュレーションを行った.

### 3.2 シミュレーション結果・考察

図 1,2 はそれぞれマイクロホンアレイ-音源間距離が 10 m かつ伝達関数が 1° 刻みの場合と, 距離が 20 m かつ伝達関数が 5° 刻みの場合シミュレーション結果である. 両図より, 提案手法は真の音源軌跡周辺を推定できていることが分かる. また, 推定音源位置と真値間のユークリッド距離を推定誤差として, 各シナリオにおける各手法の平均推定誤差と最大推定誤差を求め, 表 1,2

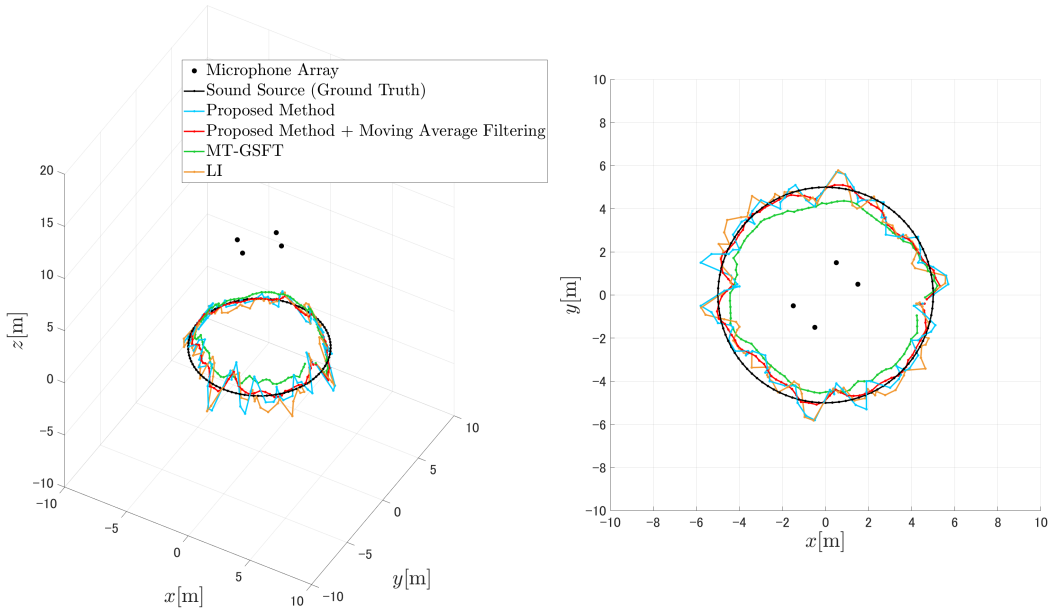



図 1: マイクロホンアレイの高さが 10 m, 伝達関数が 1° 刻みの場合の音源位置追跡結果（左図は俯瞰図, 右図は上から見た図）

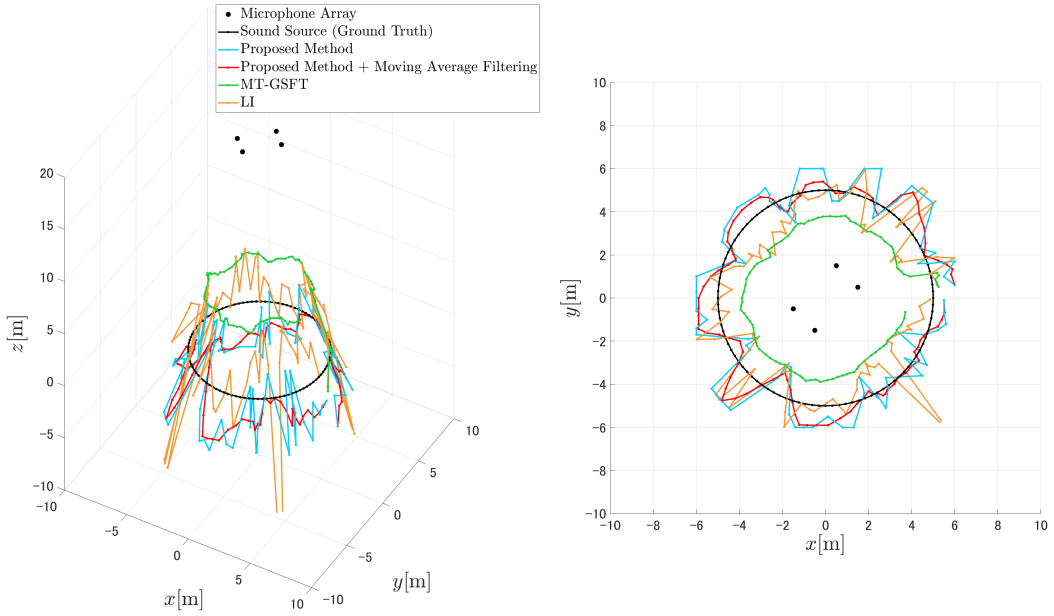



図 2: マイクロホンアレイの高さが 20 m, 伝達関数が 5° 刻みの場合の音源位置追跡結果（左図は俯瞰図, 右図は上から見た図）

表 1: 音源位置推定誤差（伝達関数が 1° 刻みの場合）

| | 高さ 10 m | | | 高さ 20 m | | |
|---|---|---|---|---|---|---|
| | 平均誤差 [m] | 最大誤差 [m] | 分散 [$m^2$] | 平均誤差 [m] | 最大誤差 [m] | 分散 [$m^2$] |
| 提案手法 | 0.99 | 2.41 | 0.32 | 2.12 | 5.16 | 1.37 |
| 提案手法 + 移動平均フィルタ | 0.53 | 1.57 | 0.13 | 1.25 | 4.16 | 0.61 |
| MT-GSFT | 1.45 | 2.03 | 0.11 | 1.88 | 3.66 | 0.23 |
| LI | 0.94 | 2.16 | 0.35 | 3.11 | 15.31 | 5.45 |

表 2: 音源位置推定誤差（伝達関数が 5° 刻みの場合）

| | 高さ 10 m | | | 高さ 20 m | | |
|---|---|---|---|---|---|---|
| | 平均誤差 [m] | 最大誤差 [m] | 分散 [$m^2$] | 平均誤差 [m] | 最大誤差 [m] | 分散 [$m^2$] |
| 提案手法 | 2.04 | 5.64 | 1.92 | 3.21 | 5.33 | 2.39 |
| 提案手法 + 移動平均フィルタ | 1.69 | 3.45 | 0.65 | 2.45 | 5.31 | 1.72 |
| MT-GSFT | 1.47 | 2.32 | 0.17 | 5.70 | 6.56 | 1.19 |
| LI | 1.36 | 4.78 | 0.91 | 3.67 | 22.83 | 8.76 |

表 3: 各マイクロホンアレイの状態. 高さ $z_n$ はシナリオによって異なる.（10 m or 20 m）

| マイクロホンアレイ | 状態 [m, m, m, rad, rad, rad] |
|---|---|
| $MA_1$ | $[\ 0.5,\quad 1.5, *, 0, 0,\quad \pi]$ |
| $MA_2$ | $[\ 1.5,\quad 0.5, *, 0, 0, 3\pi/2]$ |
| $MA_3$ | $[-0.5, -1.5, *, 0, 0,\quad 0]$ |
| $MA_4$ | $[-1.5, -0.5, *, 0, 0,\quad \pi/2]$ |

にまとめた. 表 1, 2 より, 提案手法に移動平均フィルタを組み合わせることで, 特に距離が 20 m の場合に推定誤差を他手法と比べて小さく抑えられることが分かる. また, 移動平均フィルタの導入より, 音源軌跡のバラつきが抑えられたことが, 図 1,2 の軌跡の形と, 表 3 の分散の減少から確認できる.

マイクロホンアレイ-音源間の距離が大きくなるにつれ, どの手法も特に距離方向の推定誤差が大きくなっていることが確認された. また, 伝達関数の刻みが大きくなると, 同様に距離方向の推定誤差が大きくなることが確認された. これはいずれの手法も, 音源が遠方にあるとき, 方向推定の微小な誤差が位置推定に大きく影響を与え, 距離方向にずれてしまう特性を持つからであると考えられる.

マイクロホンアレイ-音源間の距離が 20 m かつ伝達関数の刻みが 5° のとき, MT-GSFT は著しい推定誤差を示している. これは, MT-GSFT は前ステップで求めた三角測量点と近しい位置にある三角測量点付近を推定する傾向にあることと, 三角測量は伝達関数の分解能が低いほど観測が不得意な領域が大きく現れることに起因すると考えられる. 提案手法は, グリッド点が十分に三次元空間を網羅していれば, このような三角測量的なアプローチに見られる観測が難しい領域が現れることがないため, 伝達関数の刻みが大きくなっても, 伝達関数の刻みが小さい場合と比べた推定誤差の増分は小さく抑えることができたと思われる.

## 4 終わりに

本稿は, 複数マイクロホンアレイにおける音源方向尤度を元に, 三次元音源位置の推定および追跡を行う手法の提案をした. 提案手法の有効性を数値シミュレーションによって検証したところ, 三角測量的に音源位置を推定する他手法に比べて, マイクロホンアレイ-音源間の距離の増大や, 伝達関数の刻みの増大による推定誤差の増加を抑えられる効果があることが確認できた. 提案手法の複数音源に対する拡張や, 音源位置尤度の分布を応用した音源ダイナミクスの推定が今後の課題である.

## 謝辞

本研究は JSPS 科研費 16H02884, 17K00365 および 19K12017 の助成をうけた.

## 参考文献

[1] Brandstein, M. S., and Silverman, H. F., A practical methodology for speech source localization with microphone arrays, *Computer Speech & Language*, 11(2), pp. 91–126, (1997)

[2] Gabriel, D., *et al*, 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system, *Advanced Robotics*, 30(7–8), pp. 403–414, (2019)

[3] Yamada, T., *et al*, Sound Source Tracking by Drones with Microphone Arrays, *SII2020* (Submitted)

[4] 山田他, 複数マイクロホンアレイを搭載した複数の UAV による移動音源の三次元追跡手法の実収録音評価, 第 *37* 回日本ロボット学会学術講演会 *(RSJ2019)*, (2019)

[5] Knapp, C. and Carter, G., The generalized correlation method for estimation of time delay, *IEEE transactions on acoustics, speech, and signal processing*, 24(4), pp. 320–327, (1976)

[6] 浅野 太, 音のアレイ信号処理, コロナ社, (2009)

[7] Schmidt, R., Multiple emitter location and signal parameter estimation, *IEEE Trans. on Antennas and Propagation*, 34(3), pp. 276—280, 1986.

一般社団法人 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-055-4
(11/22)

# マイクロフォンアレイおよびデプスセンサーのオンラインキャリブレーションに関する考察

## Online calibration of microphone array and depth sensors

劉 超然[1*]　石井 カルロス[1]
Chaoran Liu[1]　Carlos Ishi[1]

[1] 国際電気通信基礎技術研究所 石黒浩特別研究所
[1] ATR Hiroshi Ishiguro Laboratories

**Abstract:** RGB-D sensor and microphone array are widly used for providing an instantaneous representation of the current visual and auditory environment. Sensor pose is needed for sharing and combining sensing results together. However, manual calibration of different type of sensors is tedious and time consuming. In this paper, we propose an online calibration framework that can estimate sensors' 3D pose and works with RGB-D sensor and microphone array. In the proposed framework, the calibration problem is described as a factor graph inference problem and solved with a Graph Neural Network (GNN). Instead of frequently used visual markers, we use multiple moving people as reference objects to achieve automatic calibration.

## 1 Introduction

In a sensor network, the observations from each sensor are measured on sensor's own 3D coordinate and need to be combined together to yield a collective observation. This process requires each sensor's 3D pose in the world coordinate to conduct the coordinate conversion. The calibration of sensors is a crucial but often tedious problem in a sensor network. Especially when the network includes different type of sensors, the observations themself are difficult to be used as references. This fact motivate us to propose a calibration framework that is able to calibrate different type of sensors without human intervention.

Calibration of cameras and RGB-D sensors is a well-studied topic. For cameras, researchers used wand [1], plane [2] or orthogonal planes [3] as calibration objects to calculate intrinsic and extrinsic parameters. Regarding RGB-D sensors, point cloud of a calibration plane was used to generate virtual points and calibrate sensors [4]. Conventional calibration object such as checkerboard is also used for calibrating multiple RGB-D sensors [5]. Other than calibration objects, human bodies are also used in calibration process. In [6], skeleton-based viewpoint invariant transformation (SVIT) is proposed to derive the transformation from human body to RGB-D sensor. A commonly observed human body (skeleton) by two neighboring sensors is used to calculate the relative position and orientation between two sensors. Similarly, an algorithm that calibrate and automatically re-calibrate RBG-D sensors using joints of observed skeleton is proposed in [7].

Microphone arrays are widely used for auditory environment sensing and improving robot audition [8, 9]. In [10], 3 Ω D sound maps are created by a moving 3D microphone array taking into account the prior probability of sound emitting. In [11], multiple calibrated microphone arrays are used to reproduce and/or manipulate auditory environment for people at a remote location. Multiple 3D microphone arrays are also used for hearing support system with the ability of emphasizing the target sound and depressing undesired ones [12]. In [13], a pair of linear placed microphone arrays (Kinect) are used together for sound source localization. Note that all above works using multiple microphone arrays are calibrated manually. There are few works focus on the calibration of multiple microphone arrays [14].

In this paper, we propose a auto-calibration framework works with different type of sensors simultaneously including RGB-D sensors and microphone ar-

*連絡先：株式会社　国際電気通信基礎技術研究所
〒 619-0288 京都府相楽郡精華町光台二丁目 2 番地 2
E-mail: chaoran.liu@atr.jp

rays. We used a factor graph to describe the calibration process. GNN is employed for the parameter inference.

# 2 Background

In this section, we briefly introduce 3D rotation and translation on Special Euclidian group SE(3), factor graph and Graph Neural network.

## 2.1 3D rotation & translation

A transformation in a 3D space is usually described as:

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{R}^{3\times3} & \boldsymbol{t}^{3\times1} \\ \boldsymbol{0}^{1\times3} & 1 \end{pmatrix}$$

with the top left matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ is a rotation matrix with 3 degrees of freedom, the top right vector $\boldsymbol{t} \in \mathbb{R}^{3\times1}$ is a translation vector. The set of rotation matrix $\boldsymbol{R}$s form a 3D Spatial Orthogonal group $SO(3)$ with group product the standard matrix product.

$$SO(3) = \left\{ \boldsymbol{R} \in \mathbb{R}^{3\times3} | \boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{I} | \det(\boldsymbol{R}) = 1 \right\}$$

Most gradient-based optimization algorithms such as gradient descent, Gauss-Newton and Levenberg-Marquart are designed to work on Euclidian space but not on a $SO(3)$ since addition is not defined on this manifold. The associated Lie algebra $\mathfrak{so}(3)$ of group $SO(3)$ is used instead of the matrix form $\boldsymbol{R}$ for calibrating the Jacobians on $SO(3)$ manifold. Following exponential and logarithm functions are used as $\mathfrak{so}(3) \mapsto SO(3)$ mapping function and its inverse.

$$\begin{aligned} \boldsymbol{R} &= \exp(\xi^\wedge) = \boldsymbol{I} + \frac{\sin|\xi|}{|\xi|}\xi^\wedge + \frac{1-\cos|\xi|}{|\xi|^2}(\xi^\wedge)^2 \\ \xi &= \log(\boldsymbol{R}) = \left(\frac{\theta}{2\sin(\theta)}(\boldsymbol{R} - \boldsymbol{R}^\top)\right)^\vee \\ \theta &= \arccos((\mathrm{tr}(\boldsymbol{R}) - 1)/2) \end{aligned}$$

where $\boldsymbol{R} \in SO(3)$, $\xi \in \mathfrak{so}(3)$, $|\cdot|$ is the length of a vector, $^\wedge$ indicates conversion from vector to skew symmetric matrix, and vice versa.

$$\boldsymbol{a}^\wedge = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}^\wedge = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix} = \boldsymbol{A}$$
$$\boldsymbol{A}^\vee = \boldsymbol{a}$$

In order to avoid complex calculation of Jacobian, derivative of $\mathfrak{so}(3)$ could be used as an approximation in many on-manifold optimization algorithms. Give a rotation $\boldsymbol{R} = \exp(\psi^\wedge)$ and an initial position $\boldsymbol{p}$, the derivative with respect to the increment $\Delta\boldsymbol{R} = \exp(\phi^\wedge)$ can be written as:

$$\begin{aligned} \frac{\partial \boldsymbol{R}\boldsymbol{p}}{\partial\phi} &= \lim_{\phi\to0} \frac{\exp(\phi^\wedge)\exp(\psi^\wedge)\boldsymbol{p} - \exp(\psi^\wedge)\boldsymbol{p}}{\phi} \\ &\approx \frac{(\boldsymbol{I} + \phi^\wedge)\exp(\psi^\wedge)\boldsymbol{p} - \exp(\psi^\wedge)\boldsymbol{p}}{\phi} = -(\boldsymbol{R}\boldsymbol{p}^\wedge) \end{aligned}$$

On the basis of $SO(3)$, Special Euclidian group $SE(3)$ and associated Lie algebra $\mathfrak{se}(3)$ are defined similarly:

$$SE(3) = \left\{ \boldsymbol{T} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^\top & 1 \end{bmatrix} \in \mathbb{R}^{4\times4} | \boldsymbol{R} \in SO(3), \boldsymbol{t} \in \mathbb{R}^3 \right\}$$

The exponential, logarithm and pseudo-derivative on $\mathfrak{se}(3)$ can be derived accordingly.

In this work, we use a c++ implementation named *Sophus* [1] for Lie algebra computation.

## 2.2 Calibration factor graph

In a real world problem, we cannot observe a true position of calibration objects with sensors due to measurement uncertainty. Instead, a probabilistic representation can be inferred from observed noisy data. Factor graph is a convenient graphical language for modeling such an inference problem. Assume we have two sensors $\boldsymbol{s} = (s_1, s_2)$ make measurements for a moving object $\boldsymbol{x} = (x_1, x_2, x_3)$. The observations are described as $\boldsymbol{z} = (z_{12}, \cdots, z_{23})$. Fig. 1 shows the factor graph for this sensor calibration problem.

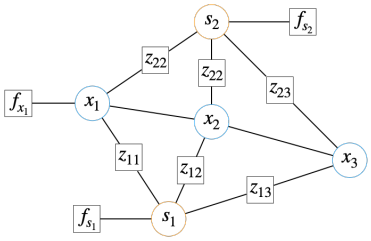



図 1: A factor graph for sensor calibration.

Fig. 1 defines a factor graph $F = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ where circles $\mathcal{V}$ describe variables, squares $\mathcal{U}$ describe factors and edges $\mathcal{E}$ are always between variables and factors. Like Bayesian networks, factor graph can describe joint probability as a product of factors. For

[1] https://github.com/strasdat/Sophus

the example in Fig. 1, the conditional probability $p(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{z})$ can be written as:

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{z}) \propto & p(x_1)p(x_2|x_1)p(x_3|x_2) \\ & \times p(s_1)p(s_2) \\ & \times l(s_1, x_1; z_{11})l(s_1, x_2; z_{12}) \cdots l(s_2, x_3; z_{23}) \end{aligned}$$

where $l(s_i, x_j; z_{ij})$ is the pseudo-likelihood factor of $s_i$ and $x_j$ given observation $z_{ij}$. Note that in this equation, $p$ and $l$ are denoted as those in Bayesian networks, but in factor graph, they are not necessarily probability distributions and can be replaced with other more generalized function $f$.

### 2.3 Graph neural network

Graph Neural Network (GNN) is a class of neural networks that process graph-structured data. Recently, it shows high ability in relation inference [15] and multi-agent interacting system [16, 17]. In [18], GNNs have been considered as performing local message passing on pairwise graphs. They generalized GNN to a Message Passing Neural Network (MPNN) architecture. In a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $v \in \mathcal{V}$ denote vertices and $e \in \mathcal{E}$ denote edges two adjacent vertices, the message pass operations in MPNN are defined as following:

$$\begin{aligned} v \to e: \quad & \boldsymbol{h}_{i,j}^{l} = \mathrm{NN}_{v2e}^{l}([\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, \boldsymbol{x}_{i,j}]) \\ e \to v: \quad & \boldsymbol{h}_{j}^{l+1} = \mathrm{NN}_{e2v}^{l}([\sum_{i \in \mathcal{N}_j} \boldsymbol{h}_{i,j}^l, \boldsymbol{x}_j]) \end{aligned}$$

where $v \to e$ and $e \to v$ denote vertex to edge and edge to vertex message passing, $\boldsymbol{h}_{i,j}^l$ and $\boldsymbol{h}_i^l$ are the embeddings (i.e. hidden layer) of edge $e_{i,j}$ and vertex $v_i$ in layer $l$ respectively, $\boldsymbol{x}_{i,j}$ and $\boldsymbol{x}_i$ are features for edge $e_{i,j}$ and $v_i$ in the initial layer, $\mathrm{NN}([\cdot])$ is a full connected neural network takes $[\cdot]$ as input, $[\cdot, \cdot]$ denotes concatenate of vectors. $\mathcal{N}_j$ denotes the set of all adjacent vertices of vertex $v_j$. These operations allow message passing between vertices and edges multiple rounds (depends on the depth of the neural network). Fig. 2 depicts these message passing neural networks.

## 3 Proposed method

In this work, we use human head positions as calibration objects to estimate position and orientation

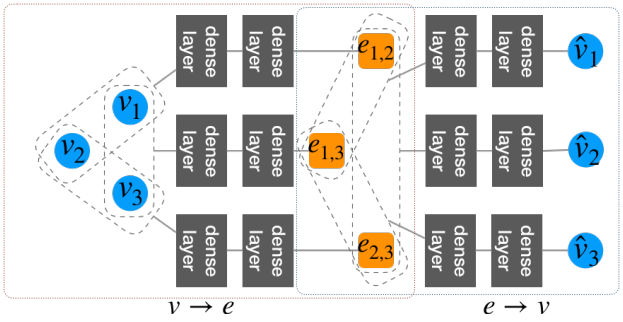



図 2: Vertex to edge and edge to vertex message passing.

of two *Intel RealSense*[2] cameras and one 16-channel microphone array. For cameras, head positions are extracted by *OpenPose* [19] and depth from RGB-D sensor. For microphone array, sound direction is calculated every 100ms.

One of our purpose in this work is to achieve online calibration need not human intervention. The calibration algorithm will run in the background and update the estimation of sensors continuously. That means the algorithm has to deal with the situation that multiple heads detected simultaneously. To this end, we first use a graph auto-encoder to perform unsupervised human identification, and then optimize the calibration factor graph on $\mathfrak{se}(3)$ manifold and estimate sensors' 3D position and orientation.

### 3.1 Human tracking in discontinuous periods

As shown in the extremely simplified Fig. 1, sensor id $i$ and object id $j$ are needed for each measurement $z_{i,j}$ to construct a proper factor graph. When there are multiple human observed simultaneously, we need to identify them before performing factor graph inference. In a continuous time period, this can easily achieved by a tracking system like Kalman filter. However, microphone array is not able to detect the direction of human all the times if he/she is not keep voicing. Consequently, a collection of audible time periods are used to perform graph inference. It is difficult to carry out human identification in discontinuous time periods. Fig. 3 shows a collection of sensor snapshots.

In Fig. 3, circles $x$ and $y$ denote two human detected simultaneously by sensors. Since we do not

[2]https://www.intelrealsense.com

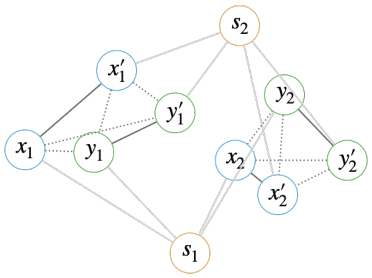



図 3: A complete graph that illustrate three snapshots of two people observed by two sensors.

have enough information about sensors, human are detected by each sensor independently. A complete graph (i.e. each pair of vertices is connected by an edge) is used as start point. However, some of the edges should not actually exist which we want to eliminate during the network training process.

A Variational Graph Auto-Encoder (VGAE) [20] is used for edges elimination. VGAE applies the idea of variational auto-encoder to graph structured data. The input of VGAE is a set of snapshots from each sensor. Sensor positions and orientations are randomly initiated. The measurements are converted to world coordinate using these randomly generated parameters. Fig. 4 shows the architecture of VGAE we used.

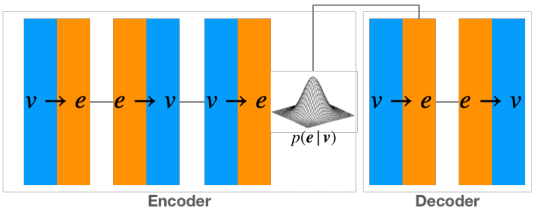



図 4: A variational graph auto-encoder used for edge elimination.

The target of encoder is to predict hidden distribution $p(\boldsymbol{z}|\boldsymbol{x})$ which is a discrete distribution with three states: non-existent, hard connected and loosely connected. The hard connected state indicates same object observed by different sensor, while the loosely connected state indicates objects share similar features (e.g. two human moved in same velocity side by side). In the decoder, the output of the first $v \to e$ neural network was multiplied by a sample extracted from $p(\boldsymbol{z}|\boldsymbol{x})$ and then used as the input of next $e \to v$ network. The decoder was trained to predict next snapshot on the timeline. The training process of VGAE is unsupervised as of a traditional VAE.

## 3.2 On-manifold 3D-2D calibration

Once a consistent human ID has been estimated, we can minimize the reprojected position error of same ID to calibrate sensors. Give RGB-D camera's pose, reprojected head position in world coordinate can be calculated easily:

$$\begin{bmatrix} x_i^{world} \\ y_i^{world} \\ z_i^{world} \end{bmatrix} = \exp(\xi_j^\wedge) \begin{bmatrix} x_i^{local} \\ y_i^{local} \\ z_i^{local} \\ 1 \end{bmatrix}$$

with $i$ and $j$ are human ID and camera ID respectively, $\xi_j$ is camera $j$'s pose in $\mathfrak{se}(3)$. The last dimension in the right hand side is omitted.

Assume that we use first RGB-D camera's coordinate as world coordinate, the pose of the second camera $\xi_2$ can be estimated by minimize:

$$\hat{\xi}_2 = \arg\min_{\xi_2} \frac{1}{2} \sum_i \left\| \boldsymbol{x}_i^1 - \exp(\xi_2^\wedge) \boldsymbol{x}_i^2 \right\|^2$$

where $\boldsymbol{x}_i^1$ is the head $i$'s position in camera 1's coordinate (world coordinate), $\boldsymbol{x}_i^2$ is the same head in camera 2's local coordinate.

Regarding the microphone array, since it only detect the direction of sound source with azimuth angle $\theta_1$ and elevation angle $\theta_2$, the optimization has to be performed in 2D. The cost function is the same one as in camera-camera calibration except we picked the second and third dimension as optimization target since it can be written as $\tan(\theta_1)$ and $\tan(\theta_2)$.

## 3.3 Experimental results

In order to test the proposed calibration algorithm, we modified an open dataset used in [21]. A Gaussian noise with standard deviation of 15cm was added into every camera measurements. Target angles to a randomly chosen microphone array position are also generated with 0 mean 2 standard deviation Gaussian noise. First, 5 sets of measurements with 100 time steps are used to test the VGAE. The training process started with fully connected graph. Experimental results show that the VGAE was able to eliminate 98.3% non-existent connections between different IDs. For the optimization on $\mathfrak{se}(3)$, we used *Ceres Solver*[3] to achieved a mean error of 22mm for RealSense and 57mm for microphone array.

[3] http://ceres-solver.org

# 4 Conclusion

In this paper, we proposed a sensor calibration framework that can automatically calibrate different type of sensors without any human intervention. It uses detected human head positions as calibration objects. The framework chooses suitable snapshot and concatenate them as time-discontinuous trunk of measurements used for calibration process. The human ID is predicted with a Variational Graph Auto-Encoder in an unsupervised manner. After corresponding human ID has been estimated, an optimization process is performed on Special Euclidian group with the associated Lie algebra $\mathfrak{se}(3)$. The experiment results on synthesized data with additional noise show that the proposed framework can predict human IDs with 100% accuracy and accurately estimate sensor positions and orientations simultaneously.

# Acknowledgment

This work was partly supported by the Tateishi Science and Technology Foundation, JST-Mirai Program Grant Number JPMJMI18C6, and JST, ERATO, Grant Number JPMJER1401.

# 参考文献

[1] Zhengyou Zhang, "Camera calibration with one-dimensional objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 892–899, July 2004.

[2] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, Nov 2000.

[3] O. Faugeras, *Three-dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA, USA: MIT Press, 1993.

[4] E. Auvinet, J. Meunier, and F. Multon, "Multiple depth cameras calibration and body volume reconstruction for gait analysis," in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, July 2012, pp. 478–483.

[5] R. Macknojia, A. Chávez-Aragón, P. Payeur, and R. Laganière, "Calibration of a network of kinect sensors for robotic inspection over a large workspace," in *2013 IEEE Workshop on Robot Vision (WORV)*, Jan 2013, pp. 184–190.

[6] Y. Han, S.-L. Chung, J.-S. Yeh, and Q.-J. Chen, "Localization of rgb-d camera networks by skeleton-based viewpoint invariance transformation," vol. 63, 10 2013, pp. 1525–1530.

[7] K. Desai, B. Prabhakaran, and S. Raghuraman, "Skeleton-based continuous extrinsic calibration of multiple rgb-d kinect cameras," in *Proceedings of the 9th ACM Multimedia Systems Conference*, ser. MMSys '18. New York, NY, USA: ACM, 2018, pp. 250–257. [Online]. Available: http://doi.acm.org/10.1145/3204949.3204969

[8] Kazuhiro Nakadai, Shunichi Yamamoto, H. G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino, "A robot referee for rock-paper-scissors sound games," in *2008 IEEE International Conference on Robotics and Automation*, May 2008, pp. 3469–3474.

[9] K. Nakamura, K. Nakadai, F. Asano, and G. Ince, "Intelligent sound source localization and its application to multimodal human tracking," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 2011, pp. 143–148.

[10] J. Even, Y. Morales, N. Kallakuri, J. Furrer, C. T. Ishi, and N. Hagita, "Mapping sound emitting structures in 3d," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 677–682.

[11] C. Liu, C. T. Ishi, and H. Ishiguro, "Bringing the scene back to the tele-operator: Auditory scene manipulation for tele-presence systems," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2015, pp. 279–286.

[12] C. T. Ishi, C. Liu, J. Even, and N. Hagita, "Hearing support system using environment sensor network," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 1275–1280.

[13] L. A. Seewald, L. Gonzaga, Jr., M. R. Veronez, V. P. Minotto, and C. R. Jung, "Combining

srp-phat and two kinects for 3d sound source localization," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7106–7113, Nov. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2014.05.033

[14] D. Su, T. Vidal-Calleja, and J. V. Miro, "Towards real-time 3d sound sources mapping with linear microphone arrays," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 1662–1668.

[15] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4967–4976.

[16] S. Sukhbaatar, a. szlam, and R. Fergus, "Learning multiagent communication with backpropagation," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2244–2252.

[17] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. kavukcuoglu, "Interaction networks for learning about objects, relations and physics," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. USA: Curran Associates Inc., 2016, pp. 4509–4517.

[18] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, pp. 1263–1272.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[20] T. N. Kipf and M. Welling, "Variational graph auto-encoders," 2016.

[21] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, "Bundle adjustment in the large," in *Proceedings of the 11th European Conference on Computer Vision: Part II*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 29–42.

一般社団法人 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-055-5
(11/22)

# スペクトル伸縮モデルと複素正規分布音源モデルに基づく複数マイクロホンの同期

# Synchronization of multiple microphones based on spectral warping and complex Gaussian source models

糸山 克寿 [1*]　中臺 一博 [1,2]
Katsutoshi Itoyama[1]　Kazuhiro Nakadai[1,2]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[2] Honda Research Institute Japan Co., Ltd.

**Abstract:** This paper describes a method of synchronizing microphones with independent A/D converters using audio signals which come from multiple sound sources observed by the microphones. The proposed model consisting a sound source model represented by a complex Gaussian distribution, a spatial transfer model represented by a steering vector, and an observation model represented by a spectrum warping matrix is constructed. Synchronization is realized as maximum a posteriori estimation of model parameters for the observed audio signals.

## 1　はじめに

マイクロホンアレイは，複数のマイクロホンと A/D コンバータ，およびコンバータを同期するためのクロックを生成する発振器からなる音響信号録音装置である．マイクロホンアレイを用いた音響信号処理では，録音された信号が全チャネルの全サンプルで同期されていることを前提に，音の伝わり方を表すステアリングベクトルや空間相関行列を用いて音源方向に基づく音源定位や音源分離などを実現する．独立した A/D コンバータをもつ複数のマイクロホンを用いた場合はマイクロホンアレイ音響信号処理を行うことはできない．同じメーカーの同じ製品であったとしても，録音された信号のサンプルレベルでの同期は仮定できないためである．

本稿では，独立した A/D コンバータをもつ複数のマイクロホンで録音された非同期音響信号の確率的生成モデルによるチャネル間の同期および音源定位・分離手法について述べる．確率的生成モデルは，音源スペクトルの生成過程を表す音源モデル，音源からマイクロホンへの伝達過程を表す空間伝達モデル，各マイクロホンのサンプリング周波数に基づいてスペクトルを変調させるスペクトル伸縮モデルの 3 つのモデルからなる．同一の音響信号が異なるサンプリング周波数で動作する複数のマイクロホンで録音された場合，その録音された音響信号は，サンプリング定理に基づいて各々のサンプリング周波数でリサンプリングされたとみなすことができる．リサンプリングは伸縮に相当する線形変換で近似できるため，音響信号のフーリエ変換で得られるスペクトルもまた，元のスペクトルの伸縮によって近似できる．この生成過程の逆問題を解くことにより，音響信号からサンプリング周波数と音源スペクトルが推定され，音源定位と分離が実現される．

*連絡先：東京工業大学 工学院 システム制御系
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: itoyama@ra.sc.e.titech.ac.jp

## 2　関連研究

非同期分散マイクロホンアレイ [1] もしくはアドホックマイクロホンアレイ [2] に関する研究における主要な課題は (1) マイクロホン位置推定，(2) 音源位置推定および分離，(3) チャネル間同期，である．課題 (1) はマイクロホンアレイのキャリブレーションとも呼ばれ，観測音エネルギーの差に基づく手法 [3, 4]，位相差に基づく手法 [5]，ビームフォーミングのクラスタリングに基づく手法 [6]，双線型写像の解に基づく手法 [7]，到達時間差 (time difference of arrival; TDOA) に基づく手法 [8]，距離行列の低ランク性を利用する手法 [9] などが報告されている．課題 (2) に関しては，TDOA に基づく手法 [10] やチャネル間同期と同時に解く手法 [11] が報告されている．(1) と (2) は類似の定式化が行える

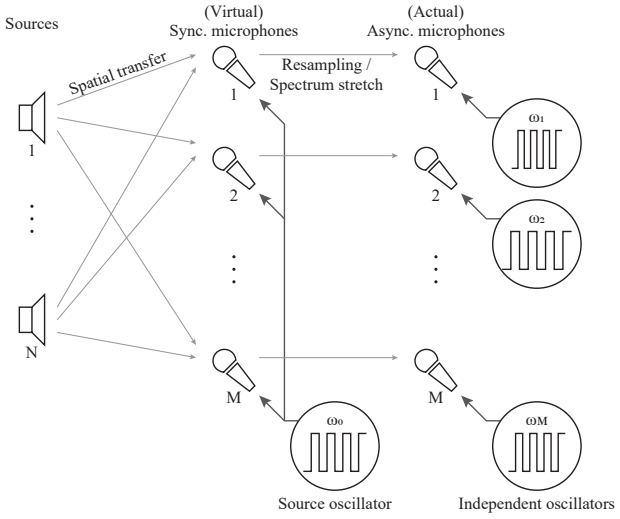



図 1: 非同期マイクロホンアレイで録音された音響信号の生成過程．複数音源からの音響信号もしくはスペクトルは空間特性にしたがった変調を受けながら各マイクロホンに到達し，サンプリング周波数と時間オフセットによってリサンプリングを受ける．

ことを利用してこれらを同時に解く手法 [3,4] や SLAM (simultaneous localization and mapping) の枠組みを用いた手法 [12,13] も報告されている．

課題 (3) に対しては，観測 TDOA に基づいて同期を行う手法 [14]，サンプリング周波数のずれが線型な位相変化を引き起こすことに基づく手法 [15] などが報告されている．また，SLAM の枠組みに基づいてサンプリング周波数とマイクロホン位置を推定する手法 [16]，サンプリング時刻のオフセットとマイクロホン位置を推定する手法 [17] が報告されている．一方で，センサネットワークの分野では，マイクロホン（センサ）間の無線通信による同期を実現する手法が報告されている [18–22]．

従来研究の多くは，環境からの雑音が十分に小さく無視できるとの仮定の下でアレイの周囲からの拍手音などを入力とすること，すなわち単独音が常に観測されることを前提として，同期が実現できることを報告している．ただし，環境からの雑音や外来音を制御できない実際の環境では，このような同期プロセスは実用的ではない．本研究では，複数同時音源の混合音を入力としたうえでマイクロホン間の同期を実現し，音源定位および分離を行うことを目的とする．

# 3 手法

本節では，本稿で提案する非同期分散マイクロホンの同期手法について述べる．提案する手法の入力と出力は以下の通りである．

**入力** 非同期分散マイクロホンで録音された音響信号のフーリエ変換で得られるマルチチャネルのスペクトログラム $X \in \mathbb{C}^{MTF}$

**出力** 各マイクロホンのサンプリング周波数 $\boldsymbol{\omega} \in \mathbb{R}^M$ とオフセット $\boldsymbol{\tau} \in \mathbb{R}^M$

**仮定** マイクロホンアレイから各方向に対するステアリングベクトル $\boldsymbol{g}_{lf} \in \mathbb{C}^M$ は既知

さらに，これに基づく音源定位・分離手法についても述べる．

## 3.1 音源モデル

音源スペクトルを $s_{ntf}$ で表現する．$n$ は音源のインデックス，$t$ と $f$ は時間フレームと周波数ビンを表す．$s_{ntf}$ は平均がゼロ，分散が $\sigma_s^2$ の複素正規分布から生成される [23] とする．

$$s_{ntf} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_s^2). \tag{1}$$

## 3.2 空間伝達モデル

$M$ をマイクロホン数，$y_{mtf}$ を $m$ 番目の仮想的な同期マイクロホンで観測されたスペクトルとする．マイクロホンの位置は既知であり，周波数領域での瞬時混合過程にしたがって複数音源が観測される．$y_{mtf}$ は音源スペクトル $y_{ntf}$ と $n$ 番目の音源のステアリングベクトル $\boldsymbol{a}_{nf} \in \mathbb{C}^M$ を用いて以下で表現される．

$$y_{mtf} = \sum_{n=1}^{N} a_{nfm} s_{ntf} + \epsilon_{mtf} \tag{2}$$

$\epsilon_{mtf}$ は観測ノイズであり，複素正規分布 $\mathcal{N}_{\mathbb{C}}(0, \sigma_y^2)$ から生成されるとすると，$y_{mtf}$ もまた以下の複素正規分布に従って生成される．

$$y_{mtf} \sim \mathcal{N}_{\mathbb{C}}\left(\sum_{n=1}^{N} a_{nfm} s_{ntf}, \sigma_y^2\right). \tag{3}$$

音源ステアリングベクトル $\boldsymbol{a}_{nf}$ について掘り下げる．空間は $L$ 個の部分空間に分割されており，各部分空間を代表する方向のステアリングベクトル $\boldsymbol{g}_{lf} = (g_{lf1}, \ldots, g_{lfM})^{\mathrm{T}}$ が計測もしくは幾何計算により与えられているとする．音源はこれらの部分空間のいくつかにまたがって存在し，その存在比を $\boldsymbol{r}_n = (r_{n1}, \ldots, r_{nL})^{\mathrm{T}}$ $(0 \leq r_{nl} \leq 1, \sum_l r_{nl} = 1)$ とする．この存在比 $\boldsymbol{r}_n$ を用いて，音源ステアリングベクトル $\boldsymbol{a}_{nf}$ を以下で定義する．

$$\boldsymbol{a}_{nf} = \sum_{l=1}^{L} r_{nl} \boldsymbol{g}_{lf} \tag{4}$$

存在比 $\boldsymbol{r}_n$ の事前分布はディリクレ分布 $\mathcal{D}(\alpha)$ とする．

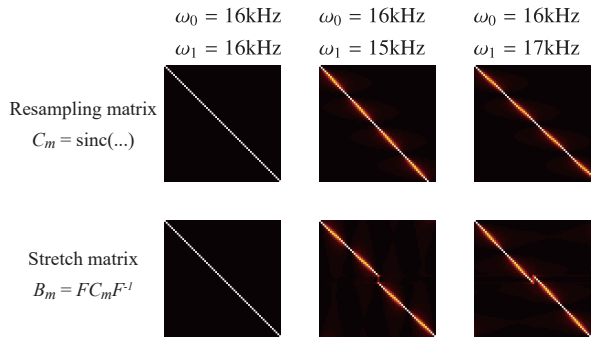



図 2: リサンプリング行列とスペクトル伸縮行列.

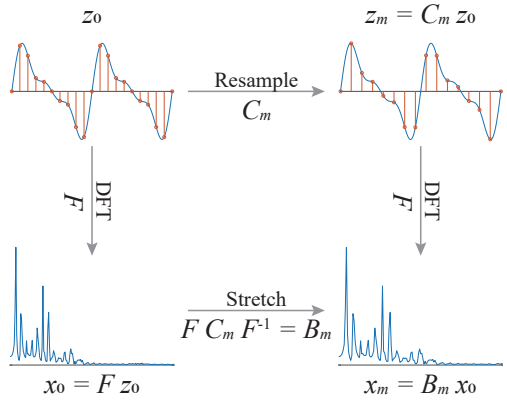



図 3: 時間領域での信号のリサンプリングと周波数領域でのスペクトルの伸縮の関係.

### 3.3 スペクトル伸縮モデル

サンプリングレート $\omega_0$ で録音された長さ $L$ の音響信号を $\boldsymbol{y}_0 = (y_{0,0}, \ldots, y_{0,L-1})^T$，同じ音響信号をサンプリングレート $\omega_1$ で録音した音響信号を $\boldsymbol{y}_1 = (y_{1,0}, \ldots, y_{1,L-1})^T$ とする．$y_{0,0}$ のサンプル時刻を 0，$y_{1,0}$ のサンプル時刻を $\tau_1$ とする．この $\tau_1$ はフレームのオフセットと解釈できる．同じ音響信号を録音しているため一方はもう一方のリサンプリングで表されると考えると，$\boldsymbol{y}_0$ と $\boldsymbol{y}_1$ の間には以下の関係が成り立つ．

$$z_{mt} = \sum_{t'=1}^{T} \mathrm{sinc}\left(\frac{\pi}{\omega_0}((t-1)\omega_m + \tau_m - (t'-1)\omega_0)\right) z_{0t'} \tag{5}$$

$$\mathrm{sinc}(t) = \frac{\sin(t)}{t} \tag{6}$$

これを行列形式で記述すると以下となる．

$$\boldsymbol{z}_m = C_m \boldsymbol{z}_0 \tag{7}$$

$B$ は以下で定義されるリサンプリング行列である．

$$C_m = \left(\mathrm{sinc}\left(\frac{\pi}{\omega_0}(t\omega_m + \tau_m - t'\omega_0)\right)\right)_{\substack{t=0,\ldots,T-1 \\ t'=0,\ldots,T-1}} \tag{8}$$

$\boldsymbol{y}_0$ と $\boldsymbol{y}_1$ をそれぞれ離散フーリエ変換して得られるスペクトルを $\boldsymbol{x}_0$ と $\boldsymbol{x}_1$ とする．

$$\boldsymbol{x}_0 = \mathcal{F}\boldsymbol{z}_0, \quad \boldsymbol{x}_m = \mathcal{F}\boldsymbol{z}_m \tag{9}$$

$$\mathcal{F} = \frac{1}{\sqrt{T}}\left(e^{-\frac{2\pi i t t'}{L}}\right)_{t=0,\ldots,T-1,\, t'=0,\ldots,T-1} \tag{10}$$

したがって，上記を合わせると以下が成り立つ．

$$\boldsymbol{x}_m = B_m \boldsymbol{x}_0, \quad B_m = \mathcal{F} C_m \mathcal{F}^{-1} \tag{11}$$

この $A$ をスペクトル伸縮行列と定義する．図 2 にリサンプリング行列とスペクトル伸縮行列の例を示す．

伸縮されたスペクトル $\boldsymbol{x}_{mt}$ は $\boldsymbol{y}_{mt}$ の線型変換であり，$\boldsymbol{y}_{mt}$ は式 (3) で示すように複素正規分布に従って生成されるため，$\boldsymbol{x}_{mt}$ もまた以下の複素正規分布に従って生成される．

$$\boldsymbol{x}_{mt} \sim \mathcal{N}_{\mathbb{C}}(\bar{\boldsymbol{x}}_{mt}, \sigma_y^2 B_m B_m^{\mathrm{H}}) \tag{12}$$

$\bar{\boldsymbol{x}}_{mt}$ は音源スペクトル，ステアリングベクトル，スペクトル伸縮行列によって定まる $\boldsymbol{x}_{mt}$ の予測値である．

$$\bar{\boldsymbol{x}}_{mt} = B_m \left(\sum_{n=1}^{N} a_{n1m} s_{nt1}, \ldots, \sum_{n=1}^{N} a_{nFm} s_{ntF}\right)^{\mathrm{T}} \tag{13}$$

### 3.4 パラメータ推定と音源定位・分離

サンプリング周波数 $\omega_m$ とオフセット $\tau_m$ の頑健な推定を実現するため，これらの事前分布を導入する．

$$\omega_m \sim \mathcal{N}(\hat{\omega}_m, \sigma_\omega^2),\ \tau_m \sim \mathcal{N}(0, \sigma_\tau^2) \tag{14}$$

これにより，モデル全体は以下で表現される．

$$\begin{aligned} & p(X, S, R, \boldsymbol{\omega}, \boldsymbol{\tau}) \\ &= \prod_{t=1}^{T}\prod_{m=1}^{M} p(\boldsymbol{x}_{mt} | \boldsymbol{s}_{1t}, \ldots, \boldsymbol{s}_{Nt}, R, \omega_m, \tau_m) \\ & \prod_{n=1}^{N}\prod_{t=1}^{T}\prod_{f=1}^{F} p(s_{ntf}) \prod_{n=1}^{N} p(\boldsymbol{r}_n) \prod_{m=1}^{M} p(\omega_m) \prod_{m=1}^{M} p(\tau_m) \end{aligned} \tag{15}$$

パラメータと観測スペクトルの対数同時確率は以下で表わされる．

$$\log p(X, S, R, \boldsymbol{\omega}, \boldsymbol{\tau}) \tag{16}$$

$$
\begin{aligned}
&= -\sum_{m=1}^{M} T \log |\sigma_y^2 B_m B_m^{\mathrm{H}}| \\
&\quad - \sum_{t=1}^{T}\sum_{m=1}^{M} (\boldsymbol{x}_{mt} - \bar{\boldsymbol{x}}_{mt})^{\mathrm{H}} \left(\sigma_y^2 B_m B_m^{\mathrm{H}}\right)^{-1} (\boldsymbol{x}_{mt} - \bar{\boldsymbol{x}}_{mt}) \\
&\quad - \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{f=1}^{F} \frac{\|s_{ntf}\|_2^2}{\sigma_s^2} + \sum_{n=1}^{N}\sum_{l=1}^{L} \alpha \log r_{nl} \\
&\quad - \sum_{m=1}^{M} \frac{(\omega_m - \hat{\omega}_m)^2}{2\sigma_\omega^2} - \sum_{m=1}^{M} \frac{\tau_m^2}{2\sigma_\tau^2} + \text{const.}
\end{aligned}
$$

サンプリング周波数 $\omega_m$ とオフセット $\tau_m$ の事後分布は正規分布を提案分布としたメトロポリス法によって推定する．音源方向 $\boldsymbol{r}_n$ の事後分布はディリクレ分布を提案分布としたメトロポリス・ヘイスティングス法によって推定する．

音源スペクトル $s_{ntf}$ の事後分布の推定，すなわち音源分離は線型フィルタリングによって実現される．スペクトル伸縮を補償する行列 $\bar{B}_m$ はサンプリング周波数とオフセットによって以下で定まる．

$$
\bar{C}_m = \left( \mathrm{sinc}\left( \frac{\pi}{\omega_m}(t\omega_0 - \tau_m - t'\omega_m) \right) \right)_{\substack{t=0,\ldots,T-1 \\ t'=0,\ldots,T-1}} \tag{17}
$$

$$
\bar{B}_m = \mathcal{F}\bar{C}_m\mathcal{F}^{-1} \tag{18}
$$

この行列を観測スペクトル $\boldsymbol{x}_{mt}$ に適用することで，伸縮を補償したスペクトルが推定される．

$$
\boldsymbol{y}_{mt} = \bar{B}_m \boldsymbol{x}_{mt} \tag{19}
$$

さらにウィーナーフィルタを適用することで，音源スペクトルの推定値 $\hat{s}_{ntf}$ が得られる

$$
\hat{s}_{ntf} = \frac{\boldsymbol{a}_{nf}^{\mathrm{H}}}{\boldsymbol{a}_{nf}^{\mathrm{H}}\boldsymbol{a}_{nf}} \bar{B}_m \boldsymbol{x}_{mt} \tag{20}
$$

# 4 評価実験

提案手法を評価するためシミュレーション実験を行った．図 4 に示すように，半径 10cm の円形 4 チャネルマイクロホンアレイを用いて，マイクロホンアレイの中心から 300cm の距離に，5°刻みの位置にランダムに音源を配置した．各音源の配置に対して，ステアリングベクトルを幾何的に計算した．信号は同期された状態で録音され，その後各マイクロホンのサンプリング周波数に応じてリサンプリングすることで非同期的に録音された観測音響信号を生成した．音源数 $N$ は 2，同期状態のサンプリング周波数は 16kHz，フレーム長は 512 とした．音源スペクトルは複素正規分布から生成されたホワイトノイズを用いた．

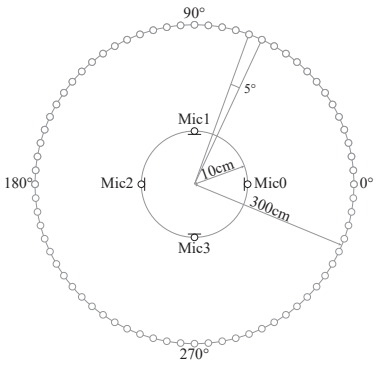



図 4: 実験環境．

表 1: サンプリング周波数の組み合わせ

| ID | サンプリング周波数 |
|---|---|
| S1 | 16010, 16000, 16000, 16000 |
| S2 | 16010, 15990, 16000, 16000 |
| S3 | 16010, 15990, 16005, 16000 |
| S4 | 16010, 15990, 16005, 15980 |

表 2: サンプリング周波数推定の実験結果．値はそれぞれ推定されたサンプリング周波数の平均値（上段）とその標準誤差（下段，括弧内）を表す．

| 組み合わせ | $\omega_0$ [Hz] | $\omega_1$ [Hz] | $\omega_2$ [Hz] | $\omega_3$ [Hz] |
|---|---|---|---|---|
| S1 | 16008.1 | 15997.4 | 15998.1 | 15997.2 |
| | (1.81) | (1.78) | (1.83) | (2.05) |
| S2 | 16009.6 | 15990.5 | 16001.1 | 15997.8 |
| | (2.35) | (3.19) | (2.05) | (2.48) |
| S3 | 16007.0 | 15991.0 | 16003.9 | 15996.9 |
| | (3.53) | (3.89) | (3.01) | (3.33) |
| S4 | 16007.2 | 15993.4 | 16004.0 | 15980.0 |
| | (2.74) | (2.81) | (2.00) | (4.90) |

表 3: 音源方向の組み合わせ

| ID | 音源方向 |
|---|---|
| D1 | 0°, 60° |
| D2 | 0°, 90° |
| D3 | 0°, 120° |
| D4 | 0°, 150° |

表 4: 音源方向推定の実験結果．値はそれぞれ推定された音源方向誤差の平均値とその標準誤差（括弧内）を表す．

| 組み合わせ | 誤差の平均値（標準誤差） |
|---|---|
| D1 | 8.42° (3.40°) |
| D2 | 11.0° (4.79°) |
| D3 | 8.48° (3.69°) |
| D4 | 8.50° (3.32°) |

サンプリング周波数推定精度は，表 1 に示すサンプリング周波数の組み合わせおよびランダムに与えられたオフセットと音源方向を用いて評価した．オフセットと音源方向を既知として，メトロポリス法を用いてサンプリング周波数の事後分布に従うサンプルを生成し，事後確率が最大となるサンプルを推定値とした．

音源定位精度は，表 3 に示す音源方向の組み合わせおよびランダムに与えられたサンプリング周波数とオフセットを用いて評価した．観測スペクトルのサンプリング周波数とオフセットを既知として，音源方向の重みパラメータをメトロポリス・ヘイスティングス法で生成し，事後確率が最大となるサンプルを音源方向の推定値とした．

実験結果を表 2 および表 4 に示す．サンプリング周波数推定値の平均誤差は最大で 3.4Hz であり，提案手法の有効性を一定の範囲で示している．16kHz からずれたサンプリング周波数のマイクロホンが増えると，推定値の標準誤差が増大している．音源定位誤差は D1, D3, D4 で類似した傾向を示し，平均誤差はおよそ 8.5 °であった．一方で， D2 では他の組み合わせよりも大きな誤差を示した．これは，音源からマイクロホンへの位相差がゼロとなる組み合わせが多くなったためだと考えられる．

## 5 おわりに

本稿では，非同期分散マイクロホンアレイでのスペクトル伸縮モデルに基づいてチャネル間の同期と音源定位を行う手法について述べた．複数の音源スペクトルと観測スペクトルとの関係性を表すモデルを確率的生成過程に基づいて表現し，その逆問題を解くことでサンプリング周波数，オフセット，音源方向，音源スペクトルを推定する．数値実験により，提案手法がサンプリング周波数と音源方向を推定狩野であることが示された．今後の課題は，低ランクもしくは深層学習に基づく音源モデル [24, 25] を導入し音源分離を実現することである．

## 謝辞

本研究の一部は JSPS 科研費 16H02884, 17K00365 および 19K12017 の助成を受けた．

## 参考文献

[1] 小野順貴, 宮部滋樹, 牧野昭二. 非同期分散マイクロホンアレイに基づく音響信号処理. 日本音響学会誌, Vol. 70, No. 7, pp. 391–396, 2014.

[2] Alexander Bertrand, Simon Doclo, Sharon Gannot, Nobutaka Ono, and Toonvan Waterschoot. Special issue on wireless acoustic sensor networks and ad hoc microphone arrays. *Signal Processing*, Vol. 107, pp. 1–3, 2015.

[3] Zicheng Liu, Zhengyou Zhang, Li-Wei He, and Phil Chou. Energy-based sound source localization and gain normalization for ad hoc microphone arrays. In *ICASSP*, Vol. 2, pp. 761–764, 2007.

[4] Minghua Chen, Zicheng Liu, Li-Wei He, Phil Chou, and Zhengyou Zhang. Energy-based position estimation of microphones and speakers for ad hoc microphone arrays. In *WASPAA*, pp. 22–25, 2007.

[5] Marius Hennecke, Thomas Plötz, Gernot A. Fink, Jörg Schmalenströer, and Reinhold Häb-Umbach. A hierarchical approach to unsupervised shape calibration of microphone array networks. In *SSP*, pp. 257–260, 2009.

[6] Ivan Himawan, Iain McCowan, and Sridha Sridharan. Clustered blind beamforming from ad-hoc microphone arrays. *IEEE Trans. Audio, Speech, and Language Process.*, Vol. 19, No. 4, pp. 661–676, 2011.

[7] Marco Crocco, Alessio Del Bue, and Vittorio Murino. A bilinear approach to the position self-calibration of multiple sensors. *IEEE Trans. Signal Process.*, Vol. 60, No. 2, pp. 660–673, 2012.

[8] Yubin Kuang and Kalle Åström. Stratified sensor network self-calibration from TDOA measurements. In *EUSIPCO*, pp. 1–5, 2013.

[9] Mohammad J. Taghizadeh, Reza Parhizkar, Philip N. Garner, HervéBourlard, Afsaneh Asaei. Ad hoc microphone array calibration: Euclidean distance matrix completion algorithm and theoretical guarantees. *Signal Processing*, Vol. 107, pp. 123–140, 2015.

[10] A. Canclini, F. Antonacci, A. Sarti, and S. Tubaro. Acoustic source localization with distributed asynchronous microphone networks. *IEEE Trans. Audio, Speech, and Language Process.*, Vol. 21, No. 2, pp. 439–443, 2013.

[11] Nobutaka Ono, Hitoshi Kohno, Nobutaka Ito, and Shigeki Sagayama. Blind alignment of asynchronously recorded signals for distributed microphone array. In *WASPAA*, pp. 161–164, 2009.

[12] Daobilige Su, Teresa Vidal-Calleja, and Jaime Valls Miro. Simultaneous asynchronous microphone array calibration and sound source localisation. In *IROS*, pp. 5561–5567, 2015.

[13] Kouhei Sekiguchi, Yoshiaki Bando, Katsutoshi Itoyama, , and Kazuyoshi Yoshii. Layout optimization of cooperative distributed microphone arrays based on estimation of source separation performance. *J. Robot. Mechatron.*, Vol. 29, No. 1, pp. 83–93, 2017.

[14] Fangyuan Jiang, Yubin Kuang, and Kalle Åström. Time delay estimation for TDOA self-calibration using truncated nuclear norm regularization. In *ICASSP*, pp. 3885–3889, 2013.

[15] Shigeki Miyabe, Nobutaka Ono, and Shoji Makino. Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation. *Signal Processing*, Vol. 107, pp. 185–196, 2015.

[16] Hiroaki Miura, Takami Yoshida, Keisuke Nakamura, and Kazuhiro Nakadai. SLAM-based online calibration of asynchronous microphone array for robot audition. In *IROS*, pp. 524–529, 2011.

[17] Keisuke Hasegawa, Nobutaka Ono, Shigeki Miyabe, and Shigeki Sagayama. Blind estimation of locations and time offsets for distributed recording devices. In *LVA/ICA*, pp. 57–64, 2010.

[18] Mohamad Hasan Bahari, Alexander Bertrand, and Marc Moonen. Blind sampling rate offset estimation for wireless acoustic sensor networks through weighted least-squares coherence drift estimation. *IEEE/ACM Trans. Audio, Speech, and Language Process.*, Vol. 25, No. 3, pp. 674–686, 2017.

[19] Alexander Bertrand. Applications and trends in wireless acoustic sensor networks: A signal processing perspective. In *SCVT*, pp. 1–6, 2011.

[20] Yoshifumi Chisaki, Dan Murakami, and Tsuyoshi Usagaway. Network-based multi-channel signal processing using the precision time protocol. In *APSIPA*, pp. 1–6, 2012.

[21] Rainer Lienhart, Igor Kozintsev, Stefan Wehr, and Minewa Yeung. On the importance of exact synchronization for distributed audio signal processing. In *ICASSP*, Vol. 4, pp. 840–843, 2003.

[22] Joerg Schmalenstroeer and Reinhold Haeb-Umbach. Sampling rate synchronisation in acoustic sensor networks with a pre-trained clock skew error model. In *EUSIPCO*, pp. 1–5, 2013.

[23] Takuma Otsuka, Katsuhiko Ishiguro, Hiroshi Sawada, and Hiroshi G. Okuno. Bayesian nonparametrics for microphone array processing. *IEEE/ACM Trans. Audio, Speech, and Language Process.*, Vol. 22, No. 2, pp. 493–504, 2014.

[24] Kousuke Itakura, Yoshiaki Bando, Eita Nakamura, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Tatsuya Kawahara. Bayesian multichannel audio source separation based on integrated source and spatial models. *IEEE/ACM Trans. Audio, Speech, and Language Process.*, Vol. 26, No. 4, pp. 831–846, 2018.

[25] Julio José Carabias-Orti, Joonas Nikunen, Tuomas Virtanen, and Pedro Vera-Candeas. Multichannel blind sound source separation using spatial covariance model with level and time differences and nonnegative matrix factorization. *IEEE/ACM Trans. Audio, Speech, and Language Process.*, Vol. 26, No. 9, pp. 1512–1527, 2018.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-055-6
(11/22)

# Between-class Learning for Sound and Image Classification

床爪 佑司 [1]　牛久 祥孝 [1]　原田 達也 [1,2]
Yuji Tokozume[1]　Yoshitaka Ushiku[1]　Tatsuya Harada[1,2]

[1] 東京大学
[1] The University of Tokyo
[2] 理化学研究所
[2] RIKEN

**Abstract:** We introduce our novel learning method for sound and image classification called between-class learning (*BC learning*). We generate between-class images by mixing two images belonging to different classes with a random ratio. We then input the mixed image to the model and train the model to output the mixing ratio. BC learning has the ability to impose constraints on the shape of the feature distributions, and thus the generalization ability is improved. As a result, classification performance of sounds and images was improved.

## 1 はじめに [1]

本発表では，実環境理解に関する研究として，ICLR 2018 および CVPR 2018 で提案した深層ニューラルネットワークの新しい教師付学習手法 between-class learning (BC learning) [Tokozume 18a, Tokozume 18b] について紹介する．

音や画像の認識において，深層学習を用いた手法が高い性能を発揮している．深層学習は，線形分離不可能なデータ空間から線形分離可能な特徴空間への関数を学習する．限られた学習データから出来る限り判別的な特徴空間を学習することが，深層学習における重要な課題である．

そこで本研究では，限られた学習データから判別的な特徴空間を学習できる，深層ニューラルネットワークの新しい教師付学習手法を提案する．新しい教師付学習手法には，ネットワーク構造や正則化等の従来の学習技術に影響を与えないこと，限られた学習データを効率的に使えること，判別的な特徴空間を学習できること，の 3 つが求められる．

ここで，判別的な特徴空間とはどのようなものだろうか．まず，クラス間の Fisher's criterion [Fisher 36] が大きい特徴空間は判別的である．Fisher's criterion とは，クラス内分散に対するクラス間距離の比のことであり，2 つのクラスがどの程度判別的であるかを表す指標である．また，各クラスが無相関な特徴空間は判別的である．識別タスクでは各クラスを等価に扱う必要があるため，特徴空間において各クラスが等間隔に並んでいることが望ましい．本研究ではこれら 2 つを判別的な特徴空間の要件とする．

従来の教師付学習では，学習データセットから単一の学習データを選択し，対応するクラスは 1，それ以外は 0 を出力するようにニューラルネットワークを学習していた．このような学習手法では，特徴空間において各クラスが線形分離可能であれば罰則が与えられないので，特徴空間が判別的になる保証は無い．本研究ではこの問題を解決する学習手法を提案する．

[1] 本稿の内容は JSAI2019 における講演予稿 (3E4-OS-12b) の転載 (一部改変) である．

## 2 Between-class Learning

### 2.1 概要と効果

本研究では，深層ニューラルネットワークの新しい教師付学習手法として，between-class learning (BC learning) を提案する．BC learning では，以下の手順でモデルを学習する．

- 異なるクラスに属する 2 つのデータを選択する．
- それらをランダムな比率で合成し，モデルに入力する．
- 合成比率を出力するようにモデルを学習する．

BC learning は，従来の学習技術に影響を与えない．また，データの合成によって学習データのパターン数が増えるため，限られた学習データを効率的に使うことができる．さらに，判別的な特徴空間を学習できる効果がある．その理由を以下に示す．

**効果 1. Fisher's criterion の増大**　図 1 (左) のように，特徴空間においてクラス A, B 間の Fisher's criterion が小さい場合を考える．クラス A, B に属するデータをある比率で合成してモデルに入力した際に，

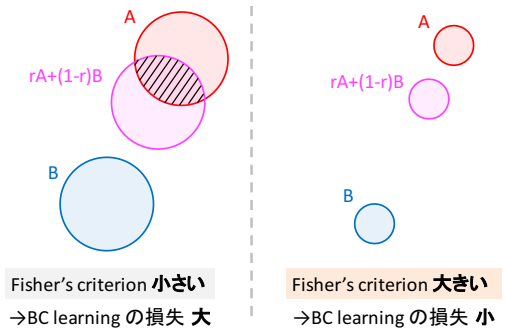



図 1: BC learning による Fisher's criterion の増大.

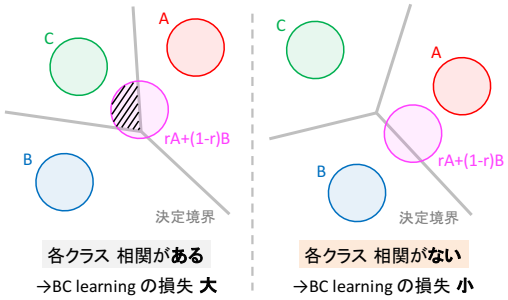



図 2: BC learning による各クラスの無相関化.

その特徴量分布 (桃色) はクラス A, B のいずれかの特徴量分布と重複することが予想される．このとき，合成するデータの組み合わせによっては，合成したデータがいずれかのクラスに分類されてしまい，モデルが合成比率を出力することができない．そのため，BC learning を行った場合の損失が大きい．一方，図 1 (右) のように Fisher's criterion が大きい場合，重複が発生しないため，BC learning による損失が小さい．学習は損失が小さくなる方向に進むため，BC learning によって図 1 (右) のような Fisher's criterion が大きい特徴空間が学習される．

**効果 2. 各クラスの無相関化**　図 2 (左) のように特徴空間において各クラスに相関がある場合，クラス A, B の合成物がクラス C に分類されるケースが発生するため，BC learning の損失が大きい．一方，図 2 (右) のように各クラスに相関がない場合，クラス A, B の合成物がクラス C に分類されないため，BC learning の損失が小さい．よって，BC learning によって図 2 (右) のような各クラスが無相関な特徴空間が学習される．

## 2.2　環境音識別への適用

音はデータ同士を合成しても音として成り立つため，BC learning が有効であると考えられる．選択された 2 つの学習データをそれぞれ $\mathbf{x}_1, \mathbf{x}_2$ とし，それらの one-hot ラベルをそれぞれ $\mathbf{t}_1, \mathbf{t}_2$ とする．また，合成比率 $r$ を一様分布 $U(0, 1)$ から生成する．ラベルの合成は単純に $r\,\mathbf{t}_1 + (1-r)\,\mathbf{t}_2$ とする．一方，データの合成は，

表 1: 環境音データセットにおける実験結果.

| モデル | 学習手法 | 誤識別率 (%) ESC-50 | ESC-10 | US8K |
|---|---|---|---|---|
| EnvNet-v2 | Standard | $21.2 \pm 0.3$ | $10.9 \pm 0.6$ | 24.9 |
| | BC (ours) | $\mathbf{15.1 \pm 0.2}$ | $\mathbf{8.6 \pm 0.1}$ | **21.7** |

表 2: 一般物体画像データセットにおける実験結果.

| モデル | 学習手法 | 誤識別率 (%) CIFAR-10 | CIFAR-100 |
|---|---|---|---|
| 11 層 CNN | Standard | $6.07 \pm 0.04$ | $26.68 \pm 0.09$ |
| | BC (ours) | $5.40 \pm 0.07$ | $24.28 \pm 0.11$ |
| | BC+ (ours) | $\mathbf{5.22 \pm 0.04}$ | $\mathbf{23.68 \pm 0.10}$ |
| ResNet-29 | Standard | $4.24 \pm 0.06$ | $20.18 \pm 0.07$ |
| | BC (ours) | $3.75 \pm 0.04$ | $19.56 \pm 0.10$ |
| | BC+ (ours) | $\mathbf{3.55 \pm 0.03}$ | $\mathbf{19.41 \pm 0.07}$ |
| Shake-Shake | Standard | 2.86 | **15.85** |
| | BC (ours) | $2.38 \pm 0.04$ | $15.90 \pm 0.06$ |
| | BC+ (ours) | $\mathbf{2.26 \pm 0.01}$ | $16.00 \pm 0.10$ |

同様に $r\,\mathbf{x}_1 + (1-r)\,\mathbf{x}_2$ とするのが単純であるが，$\mathbf{x}_1$, $\mathbf{x}_2$ それぞれの音圧レベル $G_1, G_2$ (dBA) の差を考慮した以下の合成式を提案する．

$$\frac{p\,\mathbf{x}_1 + (1-p)\,\mathbf{x}_2}{\sqrt{p^2 + (1-p)^2}} \quad \text{where} \quad p = \frac{1}{1 + 10^{\frac{G_1 - G_2}{20}} \cdot \frac{1-r}{r}} \tag{1}$$

## 2.3　画像識別への適用

画像を合成することは直感に反するが，画像データは $x$ 軸と $y$ 軸に沿った波であると考えられるので，環境音と同様に BC learning が有効であると考えられる．先程と同様に $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{t}_1$, $\mathbf{t}_2$, $r$ を定義する．ラベルの合成は単純に $r\,\mathbf{t}_1 + (1-r)\,\mathbf{t}_2$ とする．データの合成は，同様に $r\,\mathbf{x}_1 + (1-r)\,\mathbf{x}_2$ とするのが単純であるが，$\mathbf{x}_1$, $\mathbf{x}_2$ からそれぞれの平均値 $\mu_1$, $\mu_2$ を引いてゼロ平均にしたのちに，環境音と同様に合成することを提案する．音圧レベルの代わりに各画像の標準偏差 $\sigma_1$, $\sigma_2$ を用いた以下の合成式を提案する．前者の単純な合成方法を BC，後者を BC+と呼ぶことにする．

$$\frac{p\,(\mathbf{x}_1 - \mu_1) + (1-p)\,(\mathbf{x}_2 - \mu_2)}{\sqrt{p^2 + (1-p)^2}} \quad \text{where} \quad p = \frac{1}{1 + \frac{\sigma_1}{\sigma_2} \cdot \frac{1-r}{r}} \tag{2}$$

# 3　実験

環境音データセット ESC-50, ESC-10, UrbanSound8K, および一般物体画像データセット CIFAR-10, CIFAR-100 を用いて様々なモデルの学習・評価を行った．その結果の一部を表 1 および表 2 に示す．多くの条件において BC learning および BC+によって識別性能が向上した．特に CIFAR-10 において 2018 年 1 月現在の世界最高性能 2.26%を達成した．

次に大規模画像データセット ImageNet-1K を用いて実験を行った．その結果を図 3 に示す．BC learning に

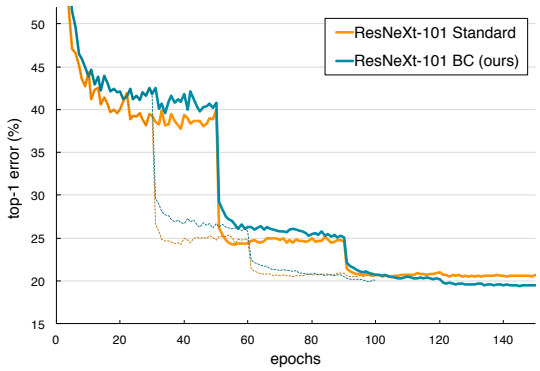



図 3: ImageNet-1K における実験結果．破線は 100 epoch，実線は 150 epoch での実験結果．

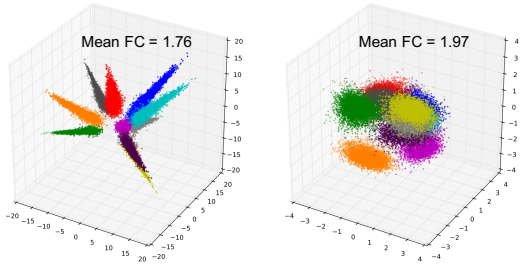



図 4: BC learning によって学習された特徴空間の可視化．

よって最終的な誤識別率が 20.4%から 19.4%へ約 1%向上した．BC learning は大規模データセットに対しても有効であることが示された．

CIFAR-10 で学習した 11 層 CNN の特徴空間（第 10 層）を PCA を用いて可視化した結果を図 4 に示す．BC learning によって学習された特徴空間は，各クラスが球状にまとまっていることが分かる．また，2 クラス間の Fisher's criterion の平均値も，BC learning の方が大きかった．BC learning によって判別的な特徴空間が学習されたといえる．

## 4 結論と今後の展望

本研究では，between-class (BC) learning という深層ニューラルネットワークの新しい教師付学習手法を提案した．実験の結果，BC learning によって音と画像の識別性能が大きく向上することが示された．BC learning は，音や画像以外のモダリティのデータの識別や，識別以外のタスクにも応用が期待される，非常に汎用性の高い技術である．また，考え方がシンプルで実装も容易であり，実用性も高い．さらに，理論的考察の余地もあり，今後さらなる研究がなされると考えられる．

## 参考文献

[Tokozume 18a] Y. Tokozume, Y. Ushiku, and T. Harada. Learning from between-class examples for deep sound recognition. In *ICLR*, 2018.

[Tokozume 18b] Y. Tokozume, Y. Ushiku, and T. Harada. Between-class Learning for Image Classification. In *CVPR*, 2018.

[Fisher 36] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, Vol. 7, No. 2, pp. 179–188, 1936.

一般社団法人 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-055-7
(11/22)

# 視聴覚統合による動的環境下における3次元再構成の提案

# Audio-visual integration for 3D reconstruction under dynamic environments

紺野 隆志[1]* 西田 健次[1] 糸山 克寿[1] 中臺 一博[1,2]
Takashi Konno[1] Kenji Nshida[1] Katsutoshi Itoyama[1] Kazuhiro Nakadai[1,2]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[2] Honda Research Institute Japan Co., Ltd.

**Abstract:** Structure from motion (SfM) usually assumes a stationary environment where no moving object exists. In practice, the environment is dynamic where many moving objects exist. It makes the performance of SfM poor. In this paper, to solve this problem, we focus on the fact that dynamic objects often emit sound, and propose an audiovisual 3D environment understanding method that integrates acoustic signal processing by microphone array processing into SfM. The performance of the proposed method is evaluated on a public dataset which simulates dynamic environments.

## 1 はじめに

3次元再構成は、ARやVR、ロボティクスなど様々な応用先があり、コンピュータビジョンにおける最も重要な分野の1つとして、多くのアルゴリズムが提案されている。例えば過去20年間で、Structure from Motion (SfM) [1] や Muti-View Stereo (MVS) [2] 、Simultaneous Localization and Mapping (SLAM) [3, 4] などが提案されている。

SfMは、物体やシーンに対して様々な視点で撮影した2次元の画像群から、カメラの位置と姿勢および、物体の3次元構造を復元する手法である [5]。物体が静的であるという仮定のもとで、カメラと物体の幾何学的な計算により3次元構造が復元される。画像内に動的物体が存在している場合は、多くの場合、画像間の特徴点マッチングにより動的物体は除外される。そのため、復元された3次元構造の点群が存在しない領域には、何も物体は存在しないのか、それとも動的物体が存在していたが除外されたのかを判別することはできない。また、特徴点マッチングにおける除外処理に失敗した場合、カメラの姿勢推定の誤りが大きくなり、いびつな形状の物体が復元されてしまう場合がある。この場合、動的物体だけではなく静的な物体の復元性能にも影響を与えてしまう。

*連絡先：東京工業大学 工学院 システム制御系
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: itoyama@ra.sc.e.titech.ac.jp

現実世界では、動的物体は、その動作や振動などにより音を発している場合が多い。例えば、走行している車はエンジンの振動音やロードノイズ、風切り音などを発し、歩行している人は足音を発している。これは、従来のSfMでは再構成から除外される動的物体は、音情報を利用することにより再構成ができる可能性があることを示唆している。そこで本稿では、動的物体は常に音を発していると仮定をし、音源定位とSfMを統合した動的環境下における3次元再構成を提案する。音と画像の空間的な対応関係を利用することにより、画像内の静的物体と各動的物体を分け、それぞれの物体ごとにSfMを行い3次元再構成をする。最後に、静的物体と動的物体を統合し、全体構造を復元する。動的物体とその物体が発する音の対応関係がとれていることにより、定位した音の視覚的な3次元構造も確認可能となる。そのため、画像だけを用いて3次元再構成を行うよりも環境理解が深まることが期待される。

## 2 関連研究

**動的環境下における3次元再構成** 動的物体の3次元復元は、コンピュータビジョンの研究でも困難な問題とされている。そのため、動的環境下における3次元再構成の多くの従来研究は、動的領域を外れ値として除外することを目指している。外れ値除去アルゴリズムとして、RANSAC [6] が最も用いられている。近年

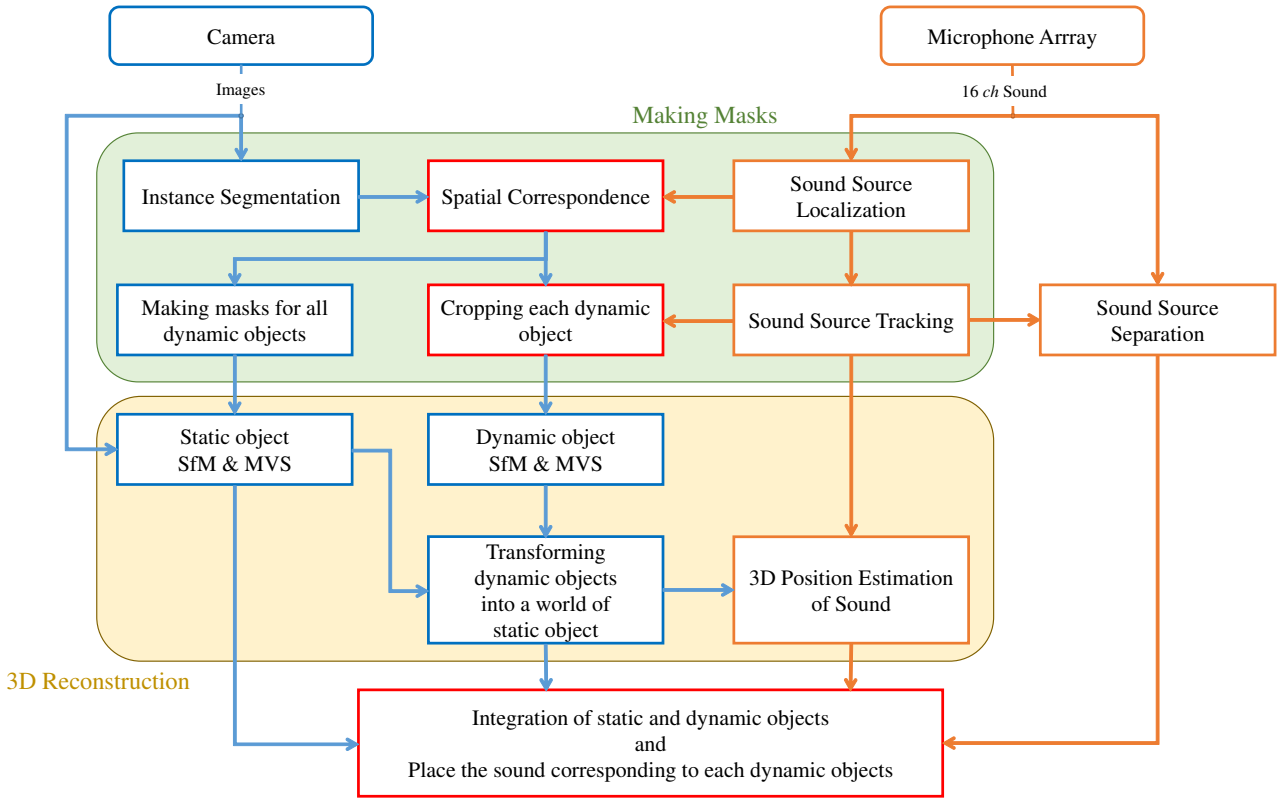



図 1: Flowchart of the proposed system. Blue frame modules indicate processing of images. Orange frame modules indicate the processing of sound. Red frame modules indicate the processing of both image and sound. Blue arrows indicate the flow of information about images. Orange arrows indicate the flow of information about sound. The modules on the green background perform making masks. Those on the yellow background perform 3D reconstruction.

は、深層学習による性能向上を背景に物体検出などを利用した新しい動的物体除去のアルゴリズムが提案されている [7, 8]。RGB 画像だけでなく、Depth 画像も用いた動的物体の 3 次元構造復元アルゴリズムとして、Co-Fusion [9] や MaskFusion [10]、MID-Fusion [11]、DetectFusion [12] などが提案されている。いずれも、本稿と同じく静的物体と動的物体を分離し、それぞれの物体ごとに 3 次元復元を行っている。Depth 画像を利用しているため、RGB 画像のみを利用する SfM よりも復元が容易となり復元性能もよいが、depth の測定可能距離には限界があり、屋外などの直射日光環境では使用が困難である。本稿は、RGB 画像と音を用いることにより、屋外での使用も目指し、さらに音の情報を利用しより高次元な環境理解を目指す。

**音情報を利用した 3 次元再構成** マイクロホンアレイ処理を用いた音源定位と、カメラや LiDAR による SLAM を用いて、音源の 3 次元位置を推定し、SLAM により復元したポイントクラウド上に音源を表示する研究が行われている [13, 14, 15, 16]。しかし、いずれの研究も定位結果の三次元空間へのマッピングが目的であり、音響情報と画像情報の相補的な統合をし性能を向上させるという取り組みはされていない。このため、本稿で目指すような、視聴覚情報を統合して、SfM の問題点を解決するという課題に対して、これらの手法を適用することは難しい。

## 3 提案手法

図 1 に、提案手法のフレームワークを示す。まず、音と画像の空間的な関係を利用し、各画像ごとに各動的物体のバイナリマスクを作成する。音源追跡により、画像間の各動的物体をトラッキングし、全画像の動的物体それぞれに対応するバイナリマスクを得る。次に、このバイナリマスクを用いて、静的物体と各動的物体ごとに SfM と MVS を適用し、それぞれの物体ごとに 3 次元構造を復元する。最後に、静的物体と動的物体を統合し、全体シーンを復元する。別のフローとして、音源定位により得られた音源の空間情報を用いて音源分離を行うことにより、各動的物体に対応する音および

その視覚的な 3 次元構造を得る。各モジュールの詳細については、以降の節で述べる。

次に、カメラとマイクロホンアレイの配置について述べる。カメラとマイクロホンアレイの相対的な位置と姿勢の関係を常に一定に保つため、カメラの上部にマイクロホンアレイを取り付ける。その際、カメラの光軸方向とマイクロホンアレイの 0 度方向が同じ方向を向くようにする。そのため、カメラの動きに合わせてマイクロホンアレイの位置と姿勢も変動する。

## 3.1 音と画像の空間的な対応関係によるバイナリマスク生成

### 3.1.1 インスタンスセグメンテーション

全画像 $N$ に対して、インスタンスセグメンテーションを適用し、画像 $\{I_i\}_{i=1}^{N} \in \mathbb{R}^{w\times h\times 3}$ 内に映る物体 $o \in \{1, ..., K\}$ の Boundin Box(BBox) $b_{i,o} \in \mathbb{R}^4$ およびそのバイナリバイナリマスク $M_{i,o} \in \mathbb{R}^{w\times h}$ を得る。$w, h$ は画像の幅と高さであり、$K$ は画像 $i$ において検出される物体数である。インスタンスセグメンテーションのアルゴリズムとして、オフラインで最も性能のよい Mask-RCNN [17] を利用する。検出される物体には、静的な物体も含まれる。

### 3.1.2 音源定位

全音源数を $L$ とする。マイクロホンアレイ処理を用いた音源定位により、画像 $i$ におけるマイクロホンアレイに対する音源 $s \in \{1, ..., L\}$ の方位角 $\theta_{i,s}$ と仰角 $\phi_{i,s}$ を得る。音源定位のアルゴリズムとして、MUSIC (MUltiple Signal Classification) 法 [18] を利用し、実装にはロボット聴覚 OSS である HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [19] を用いる。得られた音源の方向とカメラの内部パラメータ $A \in \mathbb{R}^{3\times 3}$ を利用して、音源の 3 次元位置 $P_s \sim [\tan\theta_{\mathrm{i,s}}\cos\phi_{\mathrm{i,s}}, \tan\theta_{\mathrm{i,s}}\sin\phi_{\mathrm{i,s}}, 1]^{\mathrm{T}}$ を画像に投影することにより、音源 $s$ の画像 $i$ 内の位置 $P_{i,s}(\sim AP_s) \in \mathbb{R}^2$ を得る。あらかじめ任意に定めたオフセット $off$ を用いて、以下の式により音源の BBox $b_{i,s} \in \mathbb{R}^4$ を得る。

$$b\_\mathrm{xmin_{i,s}} = \mathrm{P_{i,s}\_x} - \mathrm{off}, \quad b\_\mathrm{ymin_{i,s}} = \mathrm{P_{i,s}\_y} - \mathrm{off} \,(1)$$

$$b\_\mathrm{xmax_{i,s}} = \mathrm{P_{i,s}\_x} + \mathrm{off}, \quad b\_\mathrm{ymax_{i,s}} = \mathrm{P_{i,s}\_y} + \mathrm{off} \,(2)$$

### 3.1.3 音と画像の空間的な対応関係

画像 $i$ において、インスタンスセグメンテーションにより推定された全 BBox $b_{i,o}$ と、音源定位により推定された全 BBox $b_{i,s}$ から全ペアを抽出し、各ペアの Intersection-over-Union ($\mathrm{IoU_{i,o,s}}$) を計算する。IoU が任意のしきい値 $th_{iou}$ を超えた場合は、そのペアの $b_{i,o}$ は音源つまり動的物体の BBox であるとする。この動的物体のバイナリマスクとして、物体 $o$ に対するバイナリマスク $M_{i,o}$ を用いる。いずれの音源の BBox $b_{i,s}$ とも IoU がしきい値 $th_{iou}$ を超えなかった BBox $b_{i,o}$ は、静的な物体である可能性が高いため、この物体のバイナリマスク $M_{i,o}$ は後の処理では使用しない。しかし、いずれの BBox $b_{i,o}$ とも IoU がしきい値 $th_{iou}$ を超えなかった音源の BBox $b_{i,s}$ は、動的物体の可能性が高いが、インスタンスセグメンテーションによるバイナリマスクは得られない。そこで、この音源の BBox $b_{i,s}$ に含まれる領域を動的物体のマスクとするバイナリマスク $M_{i,s} \in \mathbb{R}^{w\times h}$ を生成し、静的な物体の復元のみに使用する。上記より、画像 $i$ における音源 $s$ に対応する動的物体のバイナリマスク $M_i^s \in \mathbb{R}^{w\times h}$ は以下のように再定義される。

$$M_i^s \leftarrow \begin{cases} M_{i,o} & \text{if} \quad \mathrm{IoU_{i,o,s}} \geq \mathrm{th_{iou}} \\ M_{i,s} & \text{if} \quad \text{otherwise} \end{cases} \tag{3}$$

### 3.1.4 全動的物体に対するバイナリマスクの生成

静的物体の復元の際に使用する、全動的物体に対するバイナリマスクの生成について述べる。画像 $i$ における全動的物体のマスクをすべて含むように、以下のように画像 $i$ におけるバイナリマスク $M_i \in \mathbb{R}^{w\times h}$ を生成する。$m$ は、$M_i^s$ と同次元で各値が 1 の行列である。

$$M_i = \left(Lm^{-1}\right) \sum_{s}^{L} M_i^s \tag{4}$$

### 3.1.5 音源追跡

音源 $s$ を音源追跡することにより、対応する動的物体を画像間でトラッキングし、式 (5) に示す全画像の各動的物体に対応するバイナリマスク群 $M^s \subset \mathbb{R}^{w\times h}$ を得る。音源追跡のアルゴリズムとして、HARK の SourceTracker[1] を利用する。

$$M^s = \{M_i^s \mid i = 1 \ldots N\} \tag{5}$$

### 3.1.6 各動的物体のみが映った画像の生成

各動的物体の復元の際に使用する、各動的物体のみが映った画像の生成について述べる。全画像に対して各動的物体に対応するバイナリマスクを掛けあわせる

[1] https://www.hark.jp/document/2.0.0/hark-document-ja/subsec-SourceTracker.html

ことにより、以下のように音源 $s$ に対応する動的物体のみが映った画像群 $\boldsymbol{D}^s \subset \mathbb{R}^{w \times h \times 3}$ を生成する。

$$\boldsymbol{D}^s = \{\boldsymbol{D}_i^s \mid \boldsymbol{D}_i^s = M_i \times I_i, \quad i = 1 \ldots N\} \tag{6}$$

## 3.2 静的物体と動的物体の３次元構造復元

### 3.2.1 静的物体の復元

画像 $i$ と対応する全動的物体に対するバイナリマスク $M_i$ をペア $(I_i, M_i)$ として、全ペアを SfM と MVS へと入力し、各カメラ姿勢と静的物体の３次元構造を復元する。SfM の処理の際に、バイナリマスクによりマスクされる領域からは特徴点を抽出しないようにし、動的物体を除外する。動的物体を除外することにより、性能向上が期待される。上記の処理の実装のベースとして、OSS の COLMAP [20] を用いる。

### 3.2.2 動的物体の復元

静的物体と動的物体が両方画像内に映っている場合、多くの場合 SfM の特徴点マッチングにおける幾何学的な外れ値処理により、動的物体は除外されてしまう。そこで、本稿では画像から動的物体のみ抽出して、動的物体のみが映った画像群を新しく生成する。この画像群においては、動的物体が剛体の場合は、擬似的に静的物体とみなすことができるため、SfM によって復元が可能となる。そのため、セクション 3.1.6 によって生成した音源 $s$ に対応する動的物体のみが映った画像群 $\boldsymbol{D}^s$ を SfM に入力することにより、各動的物体のみの３次元構造が復元可能となる。

### 3.2.3 各動的物体を静的物体のワールドへ変換

SfM では、物体は任意のスケールで復元されるため、動的物体の復元物のワールド (DW) と静的物体の復元物のワールド (SW) は、それぞれワールド座標系が異なる。そのため、各動的物体を静的物体のワールドへ変換する必要がある。

動的物体に対する相対的なカメラ位置と姿勢は、DW と SW でスケールを除き共通である。そのため、カメラ座標系を介することにより動的物体を、DW のワールド座標系に対する３次元位置 ${}^{\text{world}}P_{i,DW}^s$ から SW のワールド座標系に対する３次元位置 ${}^{\text{world}}P_{i,SW}^s$ へと変換する。

まず、式 (7) により、動的物体を DW におけるワールド座標系からカメラ座標系へ変換する。DW におけるワールド座標系からカメラ座標系への回転行列を $R_{DW} \in \mathbb{R}^{3\times3}$、並進行列 $T_{DW} \in \mathbb{R}^3$ と表す。

$${}^{\text{cam}}P_{i,DW}^s = R_{DW} \times {}^{\text{world}}P_{i,DW}^s + T_{DW} \tag{7}$$

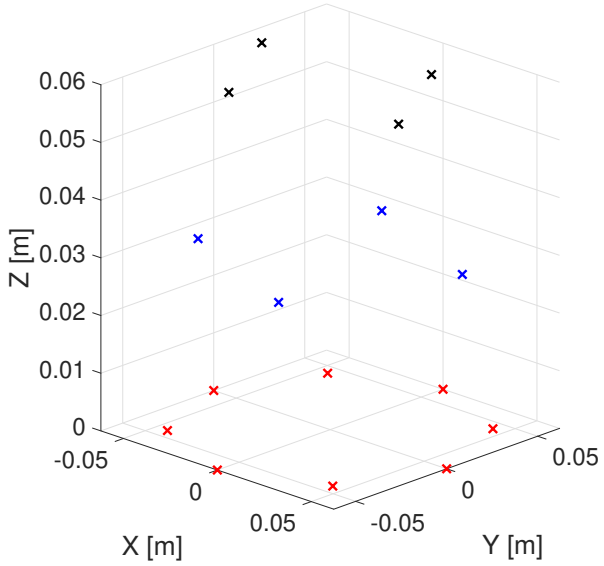



図 2: Microphone configuration of Microphone Array.

次に、式 (8) により、動的物体を DW におけるカメラ座標系 ${}^{\text{cam}}P_{i,DW}^s$ から、SW におけるカメラ座標系 ${}^{\text{cam}}P_{i,SW}^s$ へ変換する。DW から SW へのスケール変換を $S_{DW2SW} \in \mathbb{R}$ と表す。

$${}^{\text{cam}}P_{i,SW}^s = S_{DW2SW} \times {}^{\text{cam}}P_{i,DW}^s \tag{8}$$

最後に、式 (9) により、動的物体を SW におけるカメラ座標系 ${}^{\text{cam}}P_{i,SW}^s$ からワールド座標系 ${}^{\text{world}}P_{i,SW}^s$ へ変換する。SW におけるワールド座標系からカメラ座標系への回転行列を $R_{SW} \in \mathbb{R}^{3\times3}$、並進行列 $T_{SW} \in \mathbb{R}^3$ と表す。

$${}^{\text{world}}P_{i,SW}^s = R_{SW}^{-1} \times ({}^{\text{cam}}P_{i,SW}^s - T_{SW}) \tag{9}$$

## 3.3 全体シーンの復元および音源分離

式 (9) により、SW における画像 $i$ に対する音源 $s$ に対応する動的物体の３次元位置 ${}^{\text{world}}P_{i,SW}^s$ が得られる。画像 $i$ に対応する時刻 $t$ において、SW の ${}^{\text{world}}P_{i,SW}^s$ に各動的物体を配置することにより、時間的に変動する３次元構造が復元される。${}^{\text{world}}P_{i,SW}^s$ に、音源分離により分離した音源 $s$ の音を配置することにより、各動的物体に対応する音およびその視覚的な３次元構造を得る。音源分離のアルゴリズムとして、HARK に実装されている GHDSS[2] (Geometric High-order Dicorrelation-based Source Separation) を用いる。

[2] https://www.hark.jp/document/2.0.0/hark-document-ja/subsec-GHDSS.html

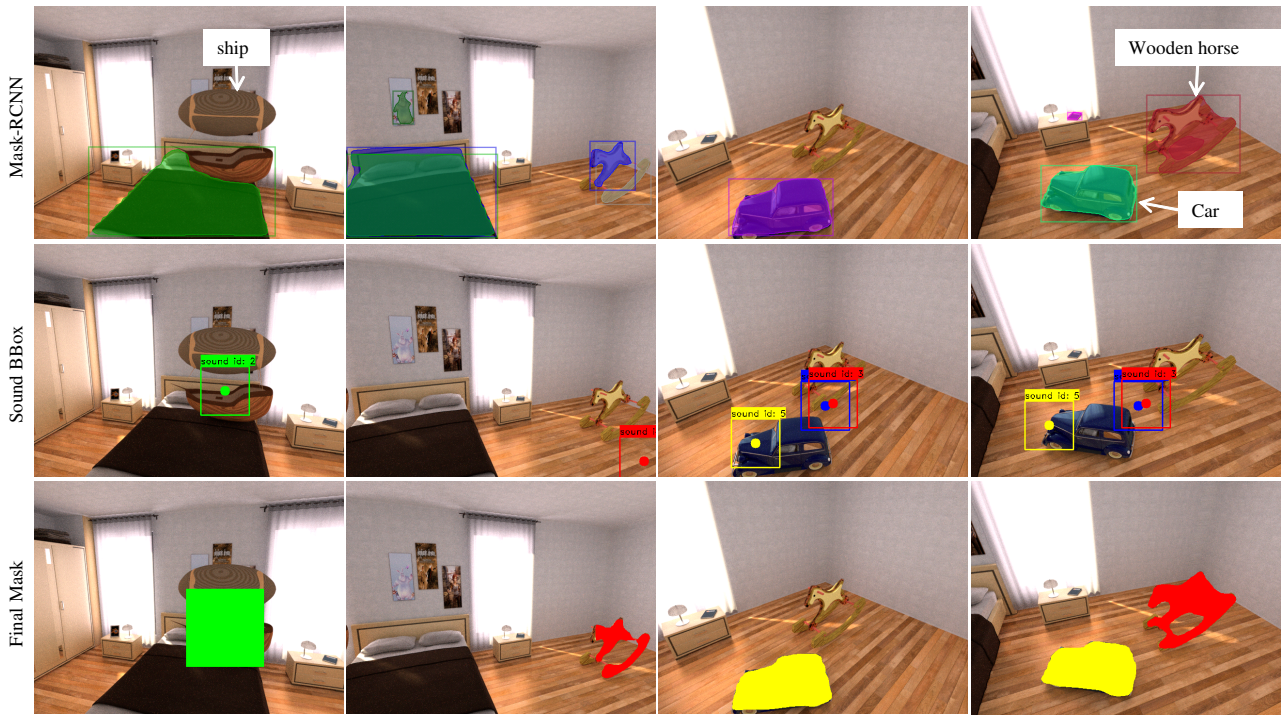
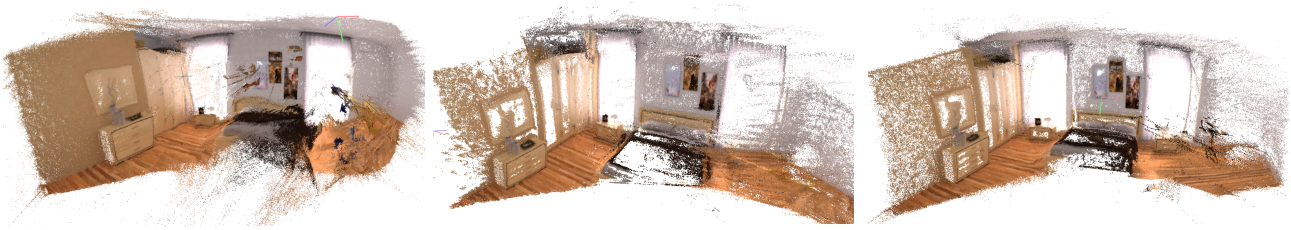



図 3: Qualitative results for making binary masks of dynamic objects. Color-coded for each detected object and sound source.



(a) Without Masks (b) With Predicted Masks (c) With Ground Truth Masks

図 4: Qualitative results for 3D reconstruction of static objects

## 4 評価実験

動的物体の 3 次元構造復元の評価を目的として、Martin らによって作成された Co-Fusion データセット [9] を用いて、提案手法の定性評価を実施し、手法の有効性および現状の限界を示す。評価実験として、動的物体のバイナリマスクの性能評価 (4.2)、静的物体の復元性能評価 (4.3)、動的物体の復元性能評価 (4.4)、全体シーンの復元性能評価 (4.5) を実施した。

### 4.1 実験設定

**Co-Fusion データセット：**Co-Fusion データセットには、複数の物体 (静的物体と動的物体いずれも) が存在する環境でカメラを動かして撮影した画像 (RGB 画像と Depth 画像) や、各時刻におけるカメラや動的物体の 3 次元位置の真値などが含まれている。シミュレーション環境と実環境で取得した、合計 4 つの環境でのデータが含まれる。本稿では、シミュレーション環境における 850 枚の RGB 画像を使用した。シミュレーションで再現した部屋の中に、3 つの動的物体 (Ship, Wooden Horse, Car) がそれぞれ独立して動いており、常に画像内に動的物体が写っているとは限らない。

**音のシミュレーション：**Co-Fusion データセットには、音が含まれていないため、シミュレーションで音を再現した。動的物体は常に音を発していると仮定し、各時刻における各動的物体の 3 次元位置の真値に音源を置いた。音は各動的物体の見た目に合わせて、16.1[kHz] で録音されたモノラル音を用いた。音の録音には、16 ch のマイクロホンアレイを用い、0 度方向がカメラの光軸方向と合うようにカメラに固定した。16 個のマイ

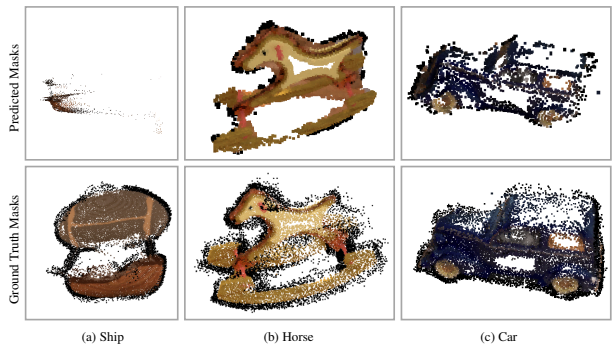



図 5: Qualitative results for 3D reconstruction of dynamic objects.

クロホンは図 2 のように、最下段に 8 個、高さ 3 cm の中段に 4 個、高さ 6 cm に 4 個配置した。音源定位には、このマイクロホンアレイに対して幾何的に計算した伝達関数を用いた。実際は音源とマイクロホンアレイどちらも動いているが、マイクロホンアレイは固定し音源を相対的に動かした。各フレームにおいて各マイクロホンと各音源の伝達関数を作成し、そのフレームの音に畳み込み、すべての音源の音を足し合わせることにより 16 ch の混合音を作成した。この混合音を用いて、システムの評価を行った。

**インスタンスセグメンテーション：**Mask-RCNN は、Detectron2 [21] で実装されているコードを利用し、ResNet-101 [22] と FPN をバックボーンとし MS COCO データセットの train2017 で学習済みのモデルを使用した。

## 4.2 動的物体のバイナリマスクの性能評価

図 3 に、Mask-RCNN(図の 1 段目) と Sound BBox(図の 2 段目) により動的物体のバイナリマスク (図の 3 段目) を生成した結果を示す。Ship は、学習済みモデルに含まれていないため Mask-RCNN では検出されない。そのため、4.2 で示した方法で音を用いてバイナリマスクを生成しているが、Ship 全体を覆うマスクは生成できていない。Horse と Car については、ある程度精度よくバイナリマスクを生成できている。しかし、音源間の距離が近づく場合に、音源定位の分解能の限界によりうまく二つの音源を定位することができず、音源追跡に失敗している場合があった。

## 4.3 静的物体の復元性能評価

図 4 に、静的物体の復元結果を示す。(a) は動的物体のバイナリマスクなし、(b) は提案手法により推定したバイナリマスクあり、(c) は Ground Truth のバイナリマスクありで、それぞれ SfM と MVS により復元した結果である。(a) は、動的物体が存在している領域に歪みが生じて復元されている。動的物体のマスクを使用しないため、画像間のマッチングで動的物体の特徴点除去に失敗し、カメラ姿勢推定誤差が大きくなってしまうためと考えられる。提案手法では 4.2 で示した通り完全なマスクを推定することができていないが、(b) の結果から (a) で見られる歪みをある程度抑えられていることが確認できる。さらに、動的物体を完全にマスクした (c) の復元結果に近い結果が得られている。完全ではないものの動的物体の特徴点をある程度除去することができているため、画像間マッチングの除去処理がうまく働いたと考えられる。

## 4.4 動的物体の復元性能評価

図 5 に、各動的物体の復元結果を示す。1 段目は提案手法、2 段目は Ground Truth のバイナリマスクを用いて復元した結果である。Ground Truth のマスクを用いた場合でも、画像から動的物体のみを抽出することにより画素数が小さく、動的物体の特徴点数が少ないため若干歪みが生じている。提案手法では、Ship は 4.2 の通りマスクの性能がよくなく、Ship 全体を覆うマスクではないため、全体を復元することはできていない。そのため Ship のマスクは、静的物体の復元に影響を与えないように生成することが主な目的となる。Horse と Car については、ある程度よく復元ができているが、マスクが生成できていないフレームもあり、Ground Truth よりも復元に使用する画像数が少なくなり、復元される点群数が少ない。

## 4.5 全体シーンの復元性能評価

図 6 に、動的物体に対する Ground Truth のバイナリマスクを用いて、全体シーンを復元した結果を示す。1 段目は元の画像、2 段目は 1 段目に対応する時間に復元されたポイントクラウドを元画像と同じサイズの空の画像に投影した結果、3 段目は 1 段目に対応する時間に復元されたポイントクラドを上から見た図である。3.2.2 と 3.2.3 の性能評価を行うため、動的物体に対するバイナリマスクは Ground Truth を用いた。ある程度よく、動的物体を DW から SW へ変換することができているが、4.4 で述べたように、動的物体の復元性能 (カメラ姿勢推定) があまりよくないため、各時刻で動的物体が振動していたり位置がずれてしまっている。また、$S_{DW2SW}$ は任意に決めているため、自動で推定する方法の開発は Future work である。

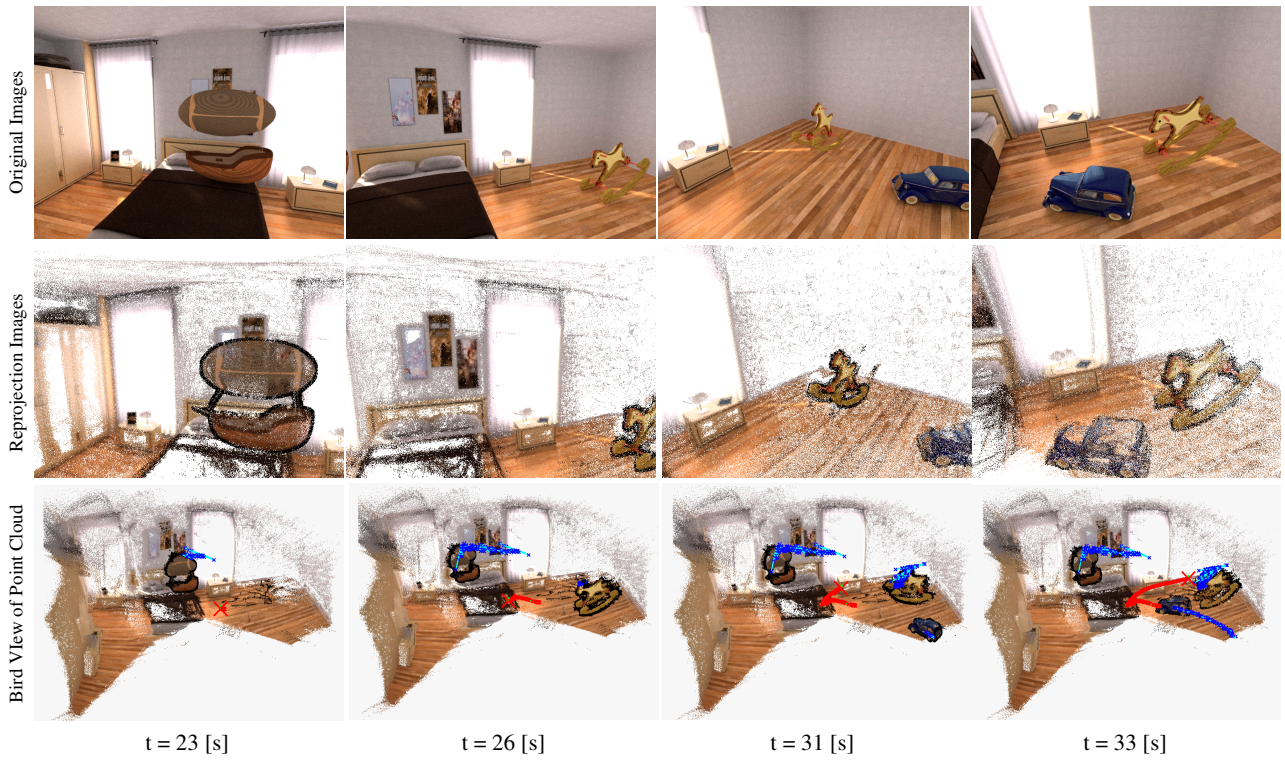



図 6: Qualitative results for 3D reconstruction of all scenes. The red line in the third row figure represents the camera trajectory, and the blue line represents the dynamic object trajectory.

## 5 おわりに

本稿では、SfM ではうまく再構成ができない動的環境下において、音響信号を手がかりに 3 次元再構成を行う手法について述べた。Co-Fusion データセットを用いて提案手法の定性評価を実施し、音響信号を用いることによって、SfM の性能を向上できる可能性があることを示した。複数の音源が近くに存在する場合に音源定位がうまくいかなくなる場合や、動的物体のスケールに関する問題など、本手法の限界も示した。今後は、設定が異なるシミュレーション環境や実環境、屋外などでの評価実験を行う。また、画像内に動的物体が写っていない場合の手法や、画像と音を統合した音源分離の手法などに取り組み、画像と音を両方利用することによる 3 次元環境理解の有効性を示していく。

## 謝辞

音のシミュレーションに関して助言を頂いた、山田泰基氏と鍾知氏には深く感謝したします。

## 参考文献

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[2] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR)*, 2007.

[3] J. Engel, T. Schöps, and D. Cremer. Lsd-slam: Large-scale direct monocular slam. In *IEEE European Conference on Computer Vision (ECCV)*, 2014.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5), 2015.

[5] R. Hartley and A. Zisserman. Multiple view geometry in computer vision., 2004. Cambridge University Press.

[6] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–3951, 1981.

[7] Bescos Berta, Fácil JM., Civera Javier, and Neira José. DynaSLAM: Tracking, mapping and inpainting in dynamic environments. *IEEE Robotics and Automation Letters*, 2018.

[8] F. Zhong, S. Wang, Z. Zhang, C. Chen, and Y. Wang. Detect-slam: Making object detection and slam mutually beneficial. In *2018 IEEE Winter Conference*

*on Applications of Computer Vision (WACV)*, pages 1001–1010, 2018.

[9] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478, 2017.

[10] M. Runz, M. Buffier, and L. Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20, 2018.

[11] Binbin Xu, Wenbin Li, Dimos Tzoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *IEEE International Conference on Robotics and Automation*, International Conference On Robotics and Automation. IEEE, 5 2019.

[12] Ryo Hachiuma, Christian Pirchheim, Dieter Schmalstieg, and Hideo Saito. Detectfusion: Detecting and segmenting both known and unknown dynamic objects in real-time slam. In *British Machine Vision Conference (BMVC)*, 2019.

[13] Y. Sasaki, R. Tanabe, and H. Takemura. Probabilistic 3d sound source mapping using moving microphone array. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1293–1298, 2016.

[14] D. Su, T. Vidal-Calleja, and J. V. Miro. Towards real-time 3d sound sources mapping with linear microphone arrays. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1662–1668, 2017.

[15] D. Su, K. Nakamura, K. Nakadai, and J. V. Miro. Robust sound source mapping using three-layered selective audio rays for mobile robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2771–2777, 2016.

[16] D. Su, T. Vidal-Calleja, and J. V. Miro. Split conditional independent mapping for sound source localisation with inverse-depth parametrisation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2000–2006, 2016.

[17] He Kaiming, Gkioxari Georgia, Dollar Piotr, and Girshick Ross. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[18] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[19] 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*, 24:739–761, 2010.

[20] Schönberger, Johannes Lutz, Frahm, and Jan-Michael. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

一般社団法人 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-055-8
(11/22)

# リハビリテーション効果推定のための感情識別器の構成と評価

# Evaluation of Emotion Detector for Estimating the Effect of Rehabilitation

西田健次[1*]　山田亨[2]　藤村友美[2]　糸山克寿[1]　中臺一博[1,3]
Kenji Nishida[1], Toru Yamada[2], Tomomi Fujimura[2], Katsutoshi Itoyama[1], Kazuhiro Nakadai[1,3]

[1] 東京工業大学
[1] Tokyo Institute of Technology
[2] 国立研究開発法人産業技術総合研究所
[2] National Institute of Advanced Industrial Science and Technology
[3] ホンダ・リサーチ・インスティチュート・ジャパン
[3] Honda Research Institute Japan

**Abstract:** 笑顔度検出を，脳卒中後遺症における音楽療法や認知症の進行抑制のための認知活性化療法での介入効果の推定手法に取り入れる提案がなされてきており，その有効性が確認されつつある脳機能障害患者の表情変化は健常者の表情変化よりも乏しいことが多く汎用の笑顔度識別器では「笑顔」と判定されないことが多いため，個人内での表情変化を捉える識別器を構成しなくてはならない．更に介入の効果判定のために個人内での表情変化を正規化し，個人間の比較を行う手法なども提案されてきている．本稿では，笑顔度検出と同様の手法で笑顔（喜び）以外の感情（怒、嫌悪、恐れ、悲しみ、驚き）の識別器を構成し，その有効性の統計的検証を行った．その結果，嫌悪と悲しみの判別など数例に関しては有意な結果が得られたが，笑顔の検出性能が特異的に高いことが示された．

## 1　はじめに

2018 年（平成 30 年）10 月 1 日現在、我が国の総人口における 65 歳以上の割合は 28.1%に達し，更に 75 歳以上人口が 12.4%を越えるなど本格的な高齢化社会を迎えている [1]．このような高齢化への社会構造の変化にともないアルツハイマー患者数は年々増加し，また患者数は漸減しているものの脳卒中の患者数も 100 万人を越えており [2]，脳機能障害に対するリハビリテーションの重要性は年々増加している．脳卒中後遺症のリハビリテーションの一つに音楽療法が挙げられる．音楽療法は、「音楽の持つ生理的，心理的，社会的働きを用いて，心身の障害の回復，機能の維持改善，生活の質の向上，行動の変容などに向けて，音楽を意図的，計画的に使用すること」と定義されている [3]．本療法においては従来，患者に対する療法の効果を，病院や音楽療法士が独自に設けた評価基準と介入内容の記録などを通じて質的・量的に評価することが試みられてきたため，客観的で統一的な評価方法を確立することは困難であった．そこで，表情の変化（笑顔度）を検出することによって音楽療法効果の客観的評価手法が提案され，その有効性が示されている [4]．また，認知症患者に対する心理療法の一つである回想法においても，従来より心理療法士の観察によって効果の評価が行われてきたが [5]，客観的な評価手法の確立が求められており，笑顔度による介入効果の評価への期待が持たれている．

笑顔という表情は「快」あるいは「幸福」の感情を示すものとすると，笑顔度は患者に対する療法の正の効果を測るものと考えることができる．逆に，「怒」，「嫌悪」，「悲しみ」，「恐怖」の感情を示す表情は，療法の負の効果を測るものであり，これらの表情を検出することは，療法の不適切な適用を避けるために有効は手段と考えることができる．そこで，本稿では，6 種の感情類型，「怒り (anger)」，「嫌悪 (disgust)」，「恐怖 (fear)」，「幸福 (happiness)」，「悲しみ (sadness)」，「驚き (surprise)」に対応する識別器を構成し，感情識別器としての有効性を評価した．表情の本稿の構成は，2 章では脳機能障害患者の表情識別に関する課題を述べ，本稿で用いた表情識別器による感情推定手法を説明する．3 章では本稿で用いたデータセットについて述べ，表情識別器の訓練手法を説明する．4 章において表情識別器による感情推定結果に対し統計的検証によってそれぞれの感情推定結果の有効性を考察する．5

*連絡先：東京工業大学
152-8552 東京都目黒区大岡山 2–12–1 W8-18
E-mail: nishida@sc.e.titech.ac.jp

章でまとめを述べ，今後の課題についても言及する．

## 2　表情識別による感情推定手法

表情識別はコンピュータビジョン，および，人工知能技術においても重要な課題であり，これまでにも多くの研究がなされてきた．そして，表情は人間の感情を推定する重要な手掛かりでもある．本稿では，脳機能障害患者の感情を推定するために，其々の感情を代表する表情を識別し，その推定値を感情の推定値として用いることを提案し，その妥当性を検証した．

### 2.1　表情識別に関する関連研究

表情識別は，コンピュータビジョンの分野において重要な課題の一つであり，多くの研究成果があげられてきた．画像による表情識別は特徴点ベースの手法 [6] とアピアランス・ベースの手法 [7, 8, 9] の二つに大別される．特徴点ベースの手法は，顔検出後に顔器官（目頭，目尻，口角，鼻など）から特徴点を抽出し，それらの位置関係から表情を検出する．顔の検出位置のずれや顔の向きに対する補正が可能なため，表情の検出性能は高いと言える．その一方で，特徴点の検出性能，特徴点のアノテーションの精度により，表情の検出性能が影響を受けるなどの問題点がある．アピアランスベースの手法は，顔検出後に顔器官を検出する必要がないため処置が単純で高速であるが，顔の位置ずれ，顔の向き，個人ごとの顔器官配置，そして，画像の撮影条件（照明，フォーカス）の影響を受けやすい．しかし，訓練サンプルに十分なバリエーションを持たせ，かつ，位置ずれや顔の向きの変化にロバストな特徴を用いることで，これらの問題点を解決する手法が提案されてきた [10]．さらに，大規模データセットと深層学習を組み合わせることで，数多くのアピアランス・ベースの表情識別手法が提案されている [11, 12, 13]．また，真顔から笑顔に変化するシーケンスを学習サンプルとすることで，笑顔という表情を検出するだけでなく，その強度（笑顔度）を推定する手法も提案されている [14]．

### 2.2　脳機能障害患者の表情識別

脳機能障害患者は表情が乏しくなることが多く [16, 17]，前節で述べたような汎用的な表情識別器では充分な検出性能を得ることが難しいことは容易に想像できる．しかし，一方で，乏しいながらも表情に変化があるならば，一般的には検出できないものであっても，その変化を捉えることによって，ある個人の表情を検出できる．さらに，無表情からある表情への変化が単調であると仮定するなら，変化の度合いをその表情の強度と考えることができる．そして，単調な変化を線形近似すると考えると顔画像の表情の強度は，式（1）で表すことができる．この表情の強度を，感情推定値として用いることとする．

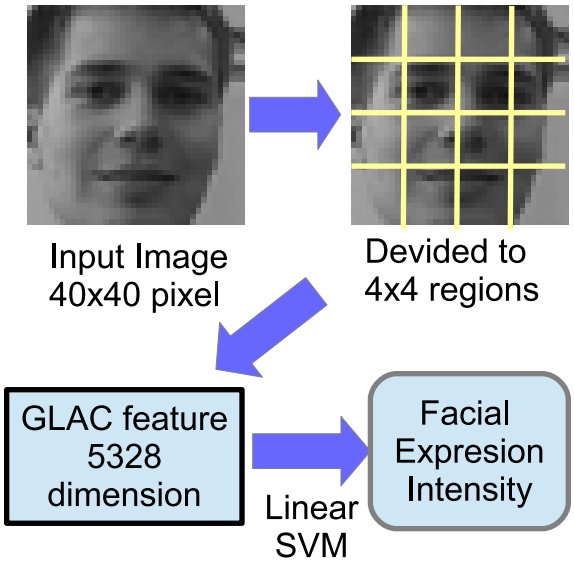



図 1: 表情識別器の構成

$$y = \boldsymbol{w}^T \boldsymbol{x} - h \tag{1}$$

ここで，$\boldsymbol{y}$ は表情の強度（スコア），$\boldsymbol{x}$ は顔画像から抽出された特徴量，$\boldsymbol{w}$ は係数ベクトル，$\boldsymbol{h}$ はバイアス値と示す．ある個人の表情の変化は $\boldsymbol{y}$ の変化によって示すことができるが，個人間での表情強度は直接比較することができないため，何らかの方法で正規化する必要がある．この正規化手法については，後述する．

本稿では，アピアランス·ベースの表情識別手法を採用し，特徴量として位置ずれや照明条件に頑健な GLAC (Gradient Local Auto-Correlation)[15] 特徴を採用した．検出された顔画像は $40 \times 40$ のグレースケール画像に変換され，$4 \times 4$，計 16 個の領域に分割される．各領域について 333 次元の GLAC 特徴が抽出され，顔画像 1 枚につき 5328 次元の特徴量が $\boldsymbol{x}$ として抽出される（図 1 ）．係数ベクトル $w$，バイアス $h$ は，特定の表情に対する 2 クラス線形サポート・ベクトル・マシン (SVM) を学習することによって得られる．

## 3　感情識別器の学習・評価用データセット

感情識別器の学習用のデータセットとして The Face-Grabber Database and Software [18]（図 2） を使用

図 2: FaceGrabber DB 顔画像の例

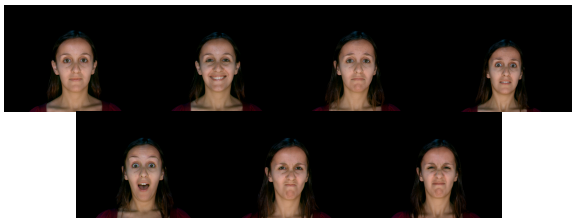


図 3: VISGRAF Faces DB 顔画像の例

した．FaceGrabber データベースは，40 人の怒り，嫌悪，恐怖，喜び（笑顔），悲しみ，驚きの 6 つの表情とニュートラルとされる表情に分類されており，一つの表情あたり 30 枚（ニュートラルに関しては 90 枚）の計 10800 枚の顔画像が含まれている．左右反転画像まで含めた計 21600 枚の画像を，1 表情分 2400 枚とそれ以外の表情全て（ニュートラル含む）19200 枚の 2 クラスに分け，2 クラス識別器による表情識別器 6 種（怒り，嫌悪，恐怖，喜び，悲しみ，驚き）の訓練を行った．

感情識別器の評価には，AIST 顔表情データベース [19] と VISGRAF Faces DB[20] （図 3）を用いた．AIST 顔表情データベースは，日本人 8 人（女性 4 人，男性 4 人）の 12 種類（怒り（閉口），怒り（開口），嫌悪（閉口），嫌悪（開口），興奮，恐怖，喜び，ニュートラル，リラックス，悲しみ，眠気，驚き）の表情が含まれているものであり，VISGRAF Faces DB は，主に欧米人 36 人の 7 種（怒り，嫌悪，喜び，ニュートラル，悲しみ，驚き）の表情が含まれている．

# 4 実験結果

## 4.1 識別器スコアの正規化

図 4 に，AIST 表情 DB の f01（女性）の感情ごとの識別器スコアのプロットを示す．6 種の識別器は独立に訓練されているため，識別器スコアの値そのものには意味がないため，識別器間での正規化を行う必要がある．本稿では，最大値-最小値での正規化（式 (2)) を

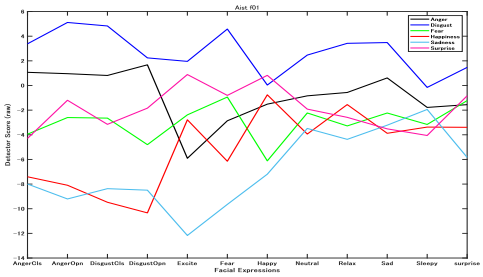



図 4: AIST 表情 DB: f01 の識別器スコア

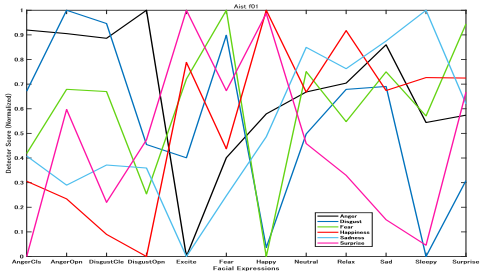



図 5: AIST 表情 DB: f01 正規化スコア

行った．これは，一人のサンプルでの感情識別器の全表情に対するスコアの最大値と最小値により，感情識別器のスコアを正規化するものである．図 5 は最大最小値での正規化の結果である．

$$\tilde{y} = \frac{y - \min(y_e)}{\max(y_e) - \min(y_e)} \tag{2}$$

$\tilde{y}$ は識別器出力 $y$ の正規化値，$y_e$ は表情 $\{e|e = anger, disgust, fear, happy, neutral, sad, surprise\}$ での識別器出力を示す．

また，個人間での感情強度の比較を行うためには，一人一人の感情のベースラインと振れ幅が異なるため，識別器のスコアそのままでは比較することができない．しかし，最大値最小値による正規化を行うことによって，各人のベースラインからの変位を求めることとなるため，個人間での比較も可能になると考えられる．

## 4.2 AIST 顔データベースに対する感情識別器の正規化スコア

AIST 顔表情データベースに対する識別器のスコアを図 6 から図 11 に示す．

図 9 に示される喜び識別器の結果は，一例を除き，怒り，嫌悪，恐怖，悲しみの負の表情に対するスコアが低く，喜び（笑顔），リラックスなどに対するスコアが

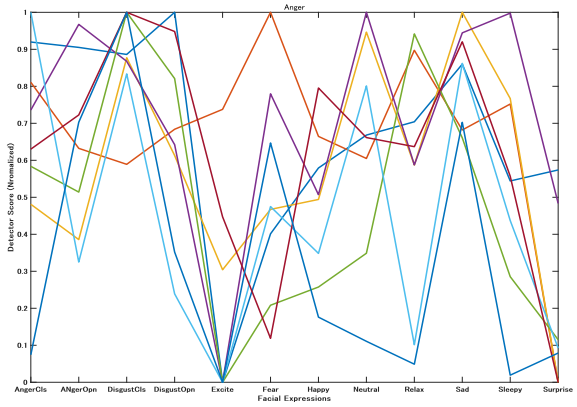



図 6: Anger 識別器の AIST-DB8 人に対する正規化スコア

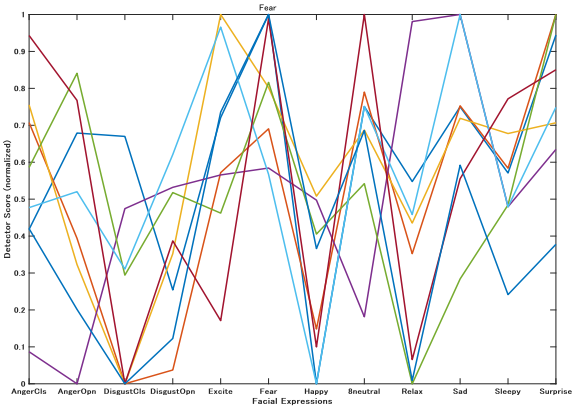



図 8: Fear 識別器の AIST-DB8 人に対する正規化スコア

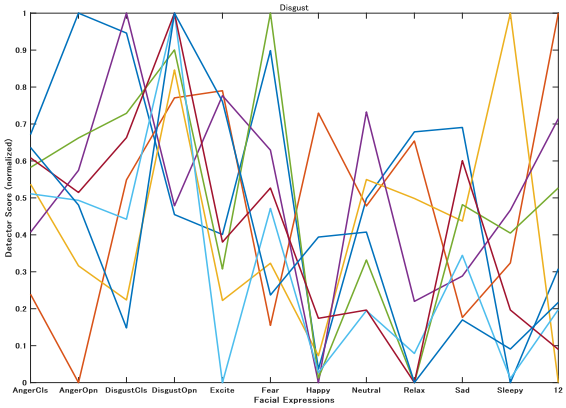



図 7: Disgust 識別器の AIST-DB8 人に対する正規化スコア

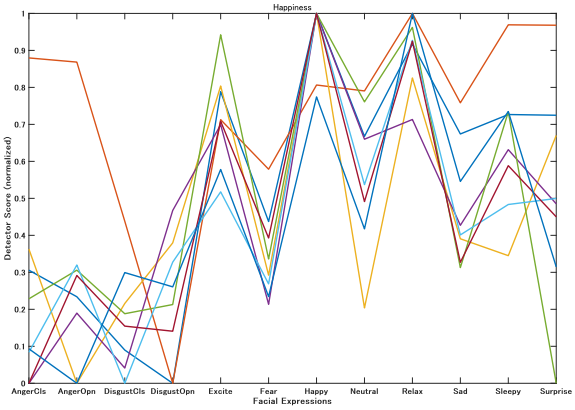



図 9: Happiness 識別器の AIST-DB8 人に対する正規化スコア

高い傾向がみられる．一方，図 6, 図 7, 図 8, 図 10 に示される怒り，嫌悪，恐怖，悲しみに対する識別器の出力は，喜びの表情に対するスコアが低いことは共通するが，対応する表情に対するスコアが必ずしも高いわけではない．

図 11 に示される驚き識別器の結果は，驚きという表情に喜び由来のものと恐怖由来のものの二種類があることを示唆している．

## 4.3 VSIGRAF DB に対する感情識別器の正規化スコア

VISGRAF-DB に対する識別器のスコアを図 12 から図 17 に示す．

喜び識別器（図 15）の特異性が際立つ結果となっている．一部に，怒り，嫌悪などにも反応している例もあるが，喜びの表情に対しては喜び識別器は高値を出力している．一方，AIST-DB と共通の傾向であるが，怒り，嫌悪，恐怖，悲しみに対する識別器は，喜び表情に対して低値を示すが，負の表情に対しては相互に高値を示す傾向にあり，その識別は難しいと考えられる．

驚き識別器（図 17）の結果も，AIST-DB に対する結果と共通し，喜び，恐怖の表情に対しても高値を示している例がある．

## 4.4 識別器スコアの統計的解析

前節で述べた識別器の傾向を確認するため，分散分析 (ANOVA)，および，Tukey-Kramer 法による多重比

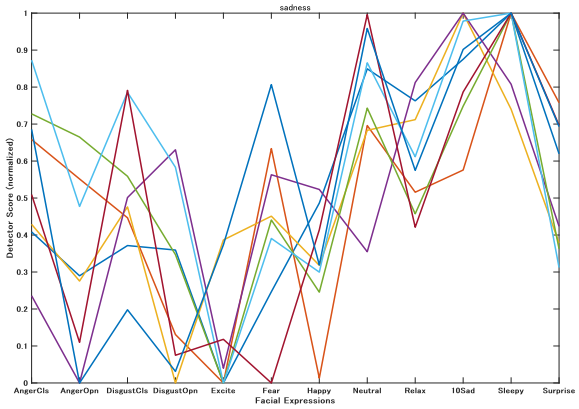



図 10: Sadness 識別器の AIST-DB8 人に対する正規化スコア

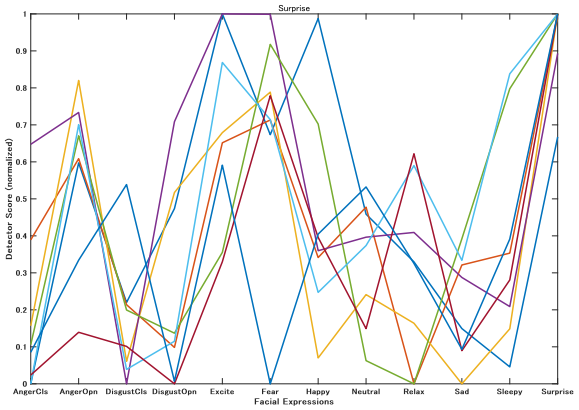



図 11: Surprise 識別器の AIST-DB8 人に対する正規化スコア

較検定 (Post-Hoc) を行った．

表 1 に AIST-DB の結果に対する分析，表 2 に VISGRAF-DB の結果に対する分析を示す．AIST-DB では，ANOVA において感情識別器の出力と顔表情 (12 種) 間には $p < 0.01$ で交互作用あり，一連の顔表情に対して感情識別器が異なる出力を示すことが示唆された．これを受けて行った post-hoc 解析では，恐怖識別器と悲しみ識別器間には $p < 0.01$ で，怒り識別器と嫌悪識別器 間，怒り識別器と驚き識別器間には $p < 0.05$ で有意差があると判定された．前節で述べた喜び識別器と他の感情識別器との差異が大きくなっていない．VISGRAF-DB では，ANOVA において感情識別器の結果と顔表情 (8 種) 間 には $p < 0.01$ で交互作用があり，やはり一連の顔表情に対して感情識別器が異なる出力を示すことが示唆された．これを受けて行った post-hoc 解析では，

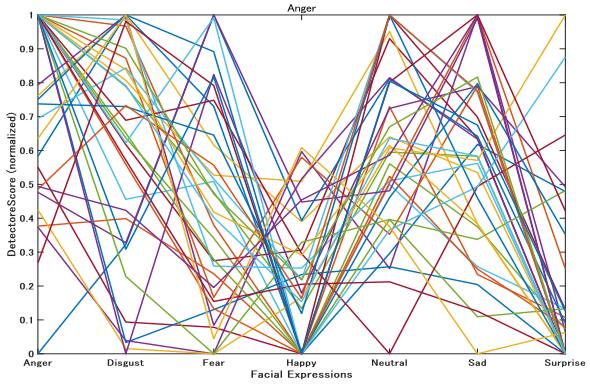



図 12: Anger 識別器の VISGRAF-DB36 人に対する正規化スコア

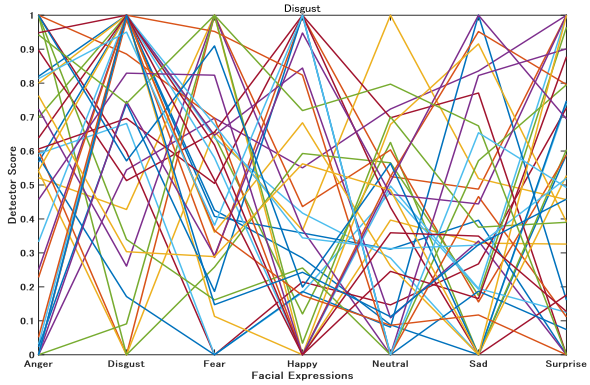



図 13: Disgust 識別器の VISGRAF-DB36 人に対する正規化スコア

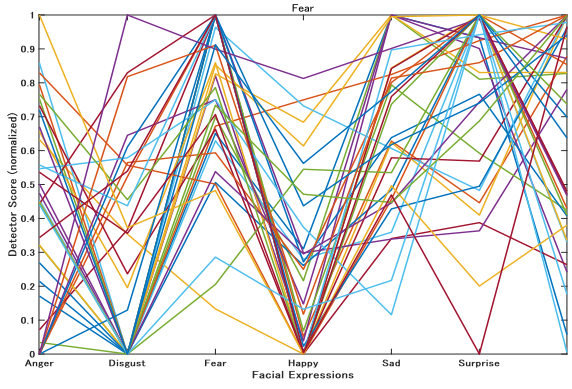



図 14: Fear 識別器の VISGRAF-DB36 人に対する正規化スコア

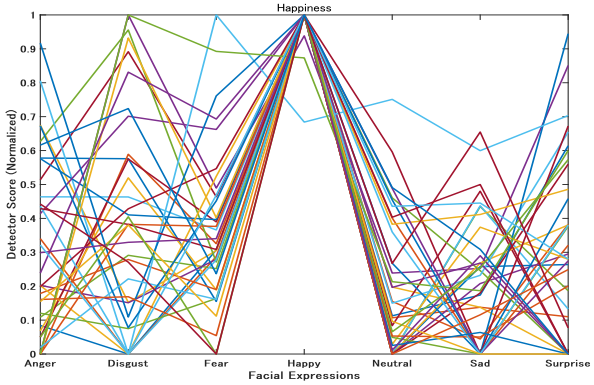



図 15: Happiness 識別器の VISGRAF-DB36 人に対する正規化スコア

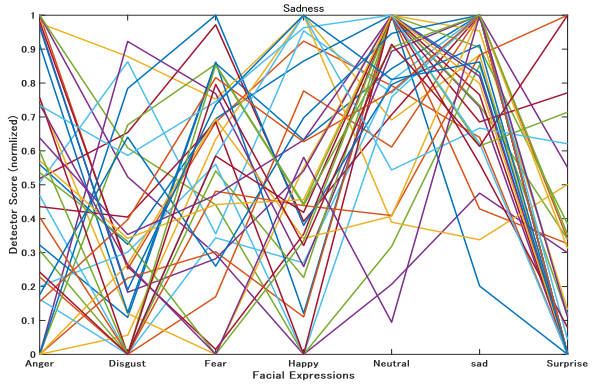



図 16: Sadness 識別器の VISGRAF-DB36 人に対する正規化スコア

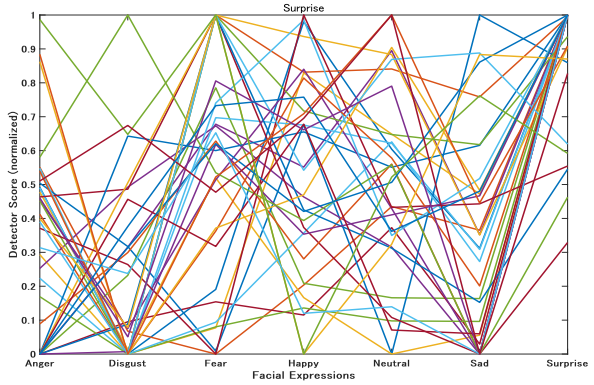



図 17: Surprise 識別器の VISGRAF-DB36 人に対する正規化スコア

表 1: AIST-DB での分散分析結果

| ANOVA | Tukey-Kramer | $p <$ |
|---|---|---|
| $p < 0.01$ | anger-disgust | 0.01 |
| | anger-surprise | 0.01 |
| | fear-surprise | 0.05 |

表 2: VISGRAF-DB での分散分析結果

| ANOVA | Tukey-Kramer | $p <$ |
|---|---|---|
| $p < 0.01$ | anger-happiness | 0.01 |
| | disgust-happiness | 0.01 |
| | fear-happiness | 0.01 |
| | happiness-sadness | 0.01 |
| | happiness-surprise | 0.01 |

喜び識別器に対して他 の 5 種の感情識別器が $p < 0.01$ で有意差があると判定され，識別器の他の組み合わせに対しては差異が認められなかった．

## 5 結論

AIST-DB，VISGRAF-DB ともに，喜び表情（笑顔）による喜びの感情推定の信頼性は高いものと考えられる．一方で，喜び以外の感情推定は，表情によっては判別が難しい組み合わせがあることが示されていると考えられる．驚き識別器が，驚き表情と同時に，喜び表情と恐怖，悲しみなどの表情に対しても高値を示すことから，複数の感情が組み合わされていると考えられる例もあり，一つの表情が必ずしも一つの感情にのみ結び付けられるものではないことが示唆されている．訓練に用いた画像は，主に欧米人の多く含まれるデータセットであったため，AIST-DB の日本人の表情に対しては，推定精度が下がっていた可能性もある．表情に対する文化的背景，性差なども考慮していく必要がある可能性がある．逆に考えると，文化的背景，人種，性差などを越えて，喜び表情（笑顔）の識別は普遍的なものと考えることができ，興味深い結果となった．

統計的な分析結果は，喜び（笑顔）とそれ以外の表情の判別性が高いため，他の表情間の判別性がマスクされてしまった可能性を示唆している．今後は，喜びの影響を排除した他の表情の識別手法を検討していく必要があると考えられる．

## 謝辞

表情識別器構成手法に関して有益なご助言をいただいた産業技術総合研究所人間情報研究部門松田圭司氏，

ならびに，笑顔度識別器のプロトタイプを回想法に適用することによってその有用性を示していただいた筑波大学人間系山中克夫准教授に感謝いたします.

## 参考文献

[1] 総務省統計局: 人口推計（2018 年（平成 30 年）10 月 1 日現在）, https://www.stat.go.jp/data/jinsui/new.html

[2] 厚生労働省: 平成 29 年（2017）患者調査の概況, https://www.mhlw.go.jp/toukei/saikin/hw/kanja/17/dl/05.pdf

[3] 関 啓子:音楽と高次脳機能障害, 音楽医療研究, Vol. 10, No. 1, pp. 14–25 (2017).

[4] 嶋田敬士, 山田亨, 高橋友香, 野口祥宏, 山崎郁子, 福井和広: SVM による笑顔度推定技術を用いた音楽療法効果の評価, 情報処理学会論文誌, Vol. 55, No. 12, pp. 2569–2581, (2014).

[5] 中谷淳, 山中克夫: 認知症ケアにおける回想法, 保険の科学, Vol. 48, No. 4, pp. 254–258, (2006).

[6] Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M. and Movellan, J.: Toward Practical Smile Detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.31, No.11, pp.2106–2111 (2009).

[7] Deniz, O., Castrillon, M., Lorenzo, J., Anton, L. and Bueno, G.: Smile Detection for User Interfaces, *ISVC' 08: Proc. 4th International Symposium on Advances in Visual Computing*, Part II, pp.602–611, Springer-Verlag, Berlin, Heidelberg (2008).

[8] Shan, C., Gong, S. and McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study, *Image and Vision Computing*, Vol.27, No.6, pp.803–816 (2009).

[9] Moore, S. and Bowden, R.: Local binary patterns for multi-view facial expression recognition, *Computer Vision and Image Understanding*, Vol.115, No.4, pp.541–558 (2011).

[10] Shimada, K., Noguchi, Y., Kurita, T.: Fast and Robust Smile Intensity Estimation by Cascaded Support Vector Machines, *Int. natnal. J. of Computer Theory and Engineering*, Vol. 5, No. 1, pp. 24–30, (2013).

[11] Yu, Z., Cha, Z.: Image Based Static Facial Expression Recognition with Multiple Deep Network Learning, *Proc. 2015 ACM International Conf. on Multimodal Interaction*, pp. 435–442, (2015).

[12] Zhang, K., Huang, Y., Du,Y., Wang, L.:Facial Expression Recognition Based on Deep Evolutional Spatio-Temporal Networks, *IEEE trans. on Image Processing*, Vol. 29, No. 9, pp. 4193–4203, (2017).

[13] Zhang, K., Huang, Y., WU, H., Wang, L.: Facial smile detection based on deep learning features, *3rd IAPR Asian Conference on PatternRecognition (ACPR)*, pp. 534–538, (2015).

[14] Sabri, M., Kurita, T.: Facial expression intensity estimation using Siamese and triplet networks, *Neurocomputing*, Vol. 313, No. 3, pp. 143–154, (2018).

[15] Kobayashi, T., Otsu, N.: Image Feathre Extraction Using Gradient Local Auto-Correlations, *European Conference on Computer Vision (ECCV)*, pp. 346–356. (2008).

[16] Borod,J. C., Koff, E., Perlman Loach, M., Nicholas, M., Welkowittz, J.: Emotional and non-emotional facial behaviour in patients with unilateral brain damages, *J. of neurology, Nuerosurgery, and Psyhiatry*, Vo. 51, pp. 826–832, (1988).

[17] Patel, S., Oishi, K., Wright, A., Sutherland-Foggio, H., Saxena, S., Shppard, S. M., Hillis, A. E.: *Frontiers in Neurology*, Vol. 9, Article 224, pp. 1–7, (2018).

[18] D. Merget, T. Eckl, M. Schw?rer, P. Tiefenbacher, and G. Rigoll, Capturing Facial Videos with Kinect 2.0: A Multithreaded Open Source Tool and Database, in *Proc. WACV*, IEEE, 2016.

[19] Fujimura, T., Umemura, H.: Development and validation of a facial expression database based on the dimensional and categorical model of emotions. *Cognition & Emotion*, Vol. 32, pp. 1663-1670, (2018).

[20] Mena-Chalco J., Marcondes R., Velho, L.: Banco de Dados de Faces 3D: IMPA-FACE3D, *TR 01, IMPA - VISGRAF Laboratory*, (2008).

# パラメトリックスピーカを用いたオーディオスポット形成における広帯域信号の安定化の検討

# Study of stabilization of wide-band signal in forming audible spot with parametric speakers

袴田拓実[1*]　干場功太郎[1]　土屋健伸[1]　遠藤信行[1]
Takumi Hakamata[1]　Kotaro Hoshiba[1]　Takenobu Tsuchiya[1]　Nobuyuki Endoh[1]

[1] 神奈川大学
[1] Kanagawa University

**Abstract:** われわれは, 特定領域にのみ音を再生させることを目的に, 二つのパラメトリックスピーカを用いた局所的可聴領域形成について研究を行っている. これまで, 再生する可聴音の周波数が変化した際，可聴領域の面積も変化してしまうという問題点を解決するため，可聴領域を安定化させるための音圧制御手法を提案してきた. しかし, 単一の周波数のみを可聴音として再生させた場合のみの評価しか行っていなかった．本稿では, 提案手法の有効性について，より複雑な条件での評価を行った．2 つの周波数成分が含まれる可聴音を再生する場合，提案手法により音圧は安定化されたが，側帯波に含まれる二成分間の差音がエイリアスとして発生することがわかった．また，広帯域の周波数成分が含まれる可聴音を再生した場合，本実験状況では側帯波の周波数帯域の狭かったため提案手法の有効性を確認することができなかった．以上の結果から，提案手法が有効な条件と解決すべき課題が明らかになった．

## 1　はじめに

### 1.1　背景

超音波の指向性を用いて, 超音波を特定方向に伝搬させることができるパラメトリックスピーカという音響デバイスがある [1]. パラメトリックスピーカは, 高い指向性故にビーム状に可聴音を形成することができ, 直線状に音を届けることが可能である. しかし, 指向性が鋭いため, 床や壁, 天井などからの反射の影響が大きく, 非対称者にまで音が聞こえてしまう. そこで, 音を対象者にのみ届けることができるような局所的な可聴領域を形成する技術が求められている.

### 1.2　関連研究

以下に可聴領域形成に関するいくつかの研究を紹介する. 深澤らは, 5m 四方の空間を 128 個からなるスピーカアレイで囲うように設置し, 可聴領域を形成するためのシステムを構築している [2]. その大きなシステムでは, 5m 四方の空間内に 4 つの可聴領域を形成した. また松井らは, 二つのスピーカから, 振幅変調した変調信号を送信することで可聴領域を形成した [3]. しかし, これらの手法では，スピーカの周波数特性を考慮しておらず，再生する可聴音の周波数が変化した際，可聴領域の面積も変化してしまうという問題点があった [4].

### 1.3　先行研究

そこでわれわれは, 安定した可聴領域の形成を目的に，パラメトリックスピーカの周波数特性を考慮した音圧制御手法を提案した [5]．提案手法では, スピーカから照射する音の音圧を周波数毎で同レベルとなるよう制御することで，スピーカの周波数特性に影響されない，安定した可聴領域形成を実現する．評価実験の結果, 可聴音の周波数が変化した場合でも一定の面積を保つことができた. しかし, 可聴音が単一の周波数のみの場合でしか検討していなかった.

### 1.4　目的

そこで本稿では，可聴音が単一周波数でない場合の可聴領域形成における提案手法の有用性について評価を行う．可聴音として，スパイク状の複数の周波数成

*連絡先：神奈川大学大学院工学研究科電気電子情報工学専攻
〒 221-8686 神奈川県横浜市神奈川区六角橋 3-27-1
E-mail: r201870089qz@jindai.jp

分が含まれる信号，そして広帯域の周波数成分を持つ信号の二通りの信号を可聴領域にて再生する．音圧制御の有無による比較を行うことで，提案手法の有効性を確認する．

# 2　手法

## 2.1　二つのパラメトリックスピーカを用いた局所的可聴領域形成手法

本研究では, 二つのパラメトリックスピーカを用いて局所的可聴領域を形成する．Fig. 1 に可聴領域形成の概略図を示す. 両スピーカから変調信号を送信し, それらの交差領域にてのみ信号が復調され可聴領域を形成する.

以下に変調手法について紹介する．片方のスピーカから送信するキャリア波 $v_c(t)$, 可聴信号 $v_s(t)$ を以下の様に定義する.

$$v_c(t) = A_c \sin 2\pi f_c t \tag{1}$$

$$v_s(t) = A_s \sin 2\pi f_s t \tag{2}$$

$A_c$, $A_s$ は, キャリア波と再生される可聴信号の振幅を表しており，$f_c$, $f_s$ は, キャリア波と可聴信号の周波数, $t$ は時間を表している. 振幅変調の一種である DSB (Double Side Band) 変調方式 [6] を用いた場合の変調波は次式の様に表すことができる.

$$\begin{aligned} v_{dsb}(t) =& A_c \sin 2\pi f_c t + \frac{A_s}{2} \sin 2\pi (f_c + f_s) \\ &+ \frac{A_s}{2} \sin 2\pi (f_c - f_s) t \end{aligned} \tag{3}$$

第一項はキャリア波, 第二項は上側側帯波, 第三項は下側側帯波である. 第一項と第二項, 第一項と第三項の差音として可聴音が復調される．この DSB 変調方式は, キャリア波と両側波帯との差音を利用するため, 再生音圧レベルが大きいという利点があるが, 高調波歪の発生が原因で音質が劣化してしまう問題点がある．そこで, 高調波歪を低減する方法として, SSB (Single Side Band) 変調方式 [7] が用いられる．SSB 変調波は次式の様に表される.

$$v_{ssb}(t) = A_c \sin 2\pi f_c t + \frac{A_s}{2} \sin 2\pi (f_c - f_s) t \tag{4}$$

第一項はキャリア波, 第二項は側帯波を表しており, これらの差音として可聴音が復調される．また, SSB 変調波は DSB 変調波と比べ, 可聴信号の振幅が半分になるが, 高調波歪が抑制される利点がある．本研究では, 変調の簡潔さから SSB 変調方式を用いている．キャリア波と側帯波を二つのパラメトリックスピーカから別々に送信することで，送信された二信号の交差領域でのみ可聴領域を形成する.

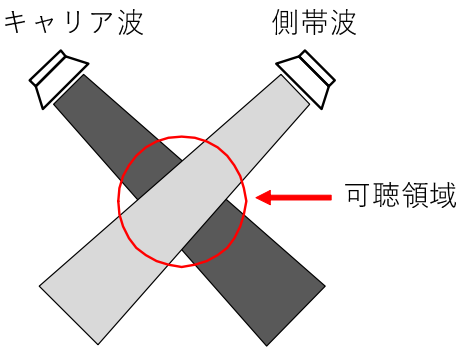



Fig. 1: 2 つのパラメトリックスピーカを用いた可聴領域形成の概略図.

## 2.2　音圧制御手法

スピーカの周波数特性により, 可聴信号の周波数 $f_s$ が変化すると, 側帯波の振幅 $A_s/2$ が変化する．そのため, 可聴領域が変化してしまい, 安定した可聴領域を形成することが困難である．そこで, キャリア波の振幅 $A_c$ と側帯波の振幅 $A_s/2$ を, 全ての周波数で等しくなるように制御する．送波系統が増幅器とパラメトリックスピーカから構成されていると仮定する．送波系統への入力電圧を $V_{in}$，周波数を $f_b = f_c - f_s$ とすると，送波系統の入出力特性は以下のように表すことができる.

$$A_s/2 = \alpha\beta V_{in}(af_b + b) \tag{5}$$

ここで，$\alpha$ は増幅器の増幅率，$\beta$ はスピーカの電圧から音圧への変換係数，$a$，$b$ は定数である．出力音圧は，$V_{in}$ に対して線形であり，$V_{in} = 0$ の場合，$Y = 0$ が成り立つため，$V_{in}$ に対しては定数項のない一次関数であるとみなせる．また，$f_b$ に関しては，大きく周波数が変化しないという仮定のもと，定数項がある一次関数で表している．このように，送波系統の入出力特性は，二変数一次関数で近似することができる．Eq. 5 を $V_{in}$ に対して解くと以下のようになる.

$$V_{in} = \frac{A_s}{2\alpha\beta(af_b + b)} \tag{6}$$

$A_s$ に $A_c$ を代入することで，周波数毎で所望の出力音圧を得るための $V_{in}$ を求めることができる．これにより, キャリア波と側帯波の振幅が全ての $f_s$ で同じレベルにすることができ, 安定した可聴領域を形成することが可能になる.

# 3　評価実験

## 3.1　測定方法

提案手法の有用性を確認するため，可聴領域を形成し評価を行った．実験状況図を Fig. 2 に示す．先行研

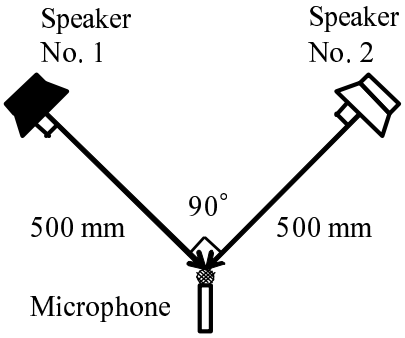



Fig. 2: 実験状況.

究にて，可聴領域の面積変化と，領域の中心音圧に相関性があることがわかっている [5]．そこで，本実験では，領域の中心音圧のみを計測する．スピーカ No. 1 とスピーカ No. 2 を，音軸が垂直に交わるよう，500mm の距離に設置した．また, マイクロホンをそれぞれのスピーカの音軸の交点に設置し, 信号を取得した．スピーカ No. 1 からキャリア波を，スピーカ No. 2 から側帯波をそれぞれ送波し，可聴領域を形成する．再生する可聴音として，二通りの実験を行った．はじめに，2 つの周波数成分が含まれる可聴音を再生した (Exp1)．キャリア波として $f_c = 40$ kHz の信号，側帯波として $f_{b-1}$, $f_{b-2}$ の 2 つの周波数成分を含む信号を用いた．$f_{b-1}$ は 39 kHz とし，$f_{b-2}$ は 38〜38.8 kHz の間で変化させた．つまり，$f_s$ として 1 kHz と 1.2〜2 kHz の 2 つの周波数成分が含まれる可聴音が再生される．続いて，広帯域の周波数成分が含まれる可聴音を再生した (Exp2)．前実験同様，キャリア波として $f_c = 40$ kHz の信号，側帯波として 39〜39.15 kHz の帯域の周波数を持つ信号を用いた．これらの二通りの実験を行い，音圧制御を導入し，その効果を評価する．Fig. 3 に側帯波を送波するスピーカ No. 2 の入出力特性の実測値を示す．横軸が周波数，縦軸が入力電圧であり，カラーマップにて出力音圧を表している．こちらのデータを用い，Eq. 5 で示した近似式を求めると以下のようになる．

$$A_s/2 = V_{in}(1.02 f_b - 36.81) \tag{7}$$

近似式から求めた近似値を Fig. 4 に示す．このように，二変数近似により入出力特性が精度良く近似できていることがわかる．得られた近似式を用いて制御を行う．

### 3.2 測定結果

はじめに Exp1 の測定結果を示す．$f_c = 40$ kHz, $f_{b-1} = 39$ kHz，$f_{b-2} = 38.2$ kHz の場合の受信信号の周波数スペクトルを Fig. 5 に示す．つまり，この場合

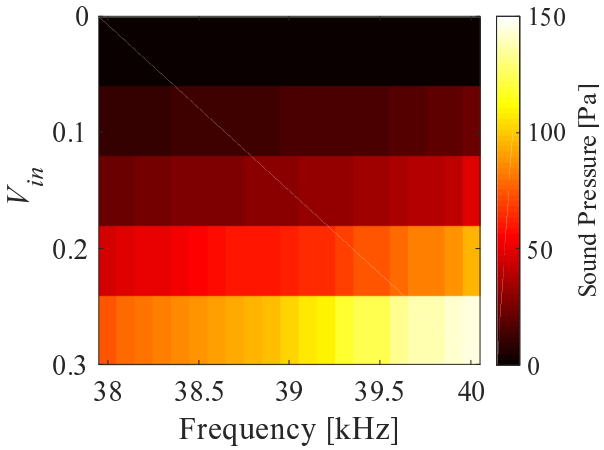



Fig. 3: スピーカ No. 2 の入出力特性の実測値.

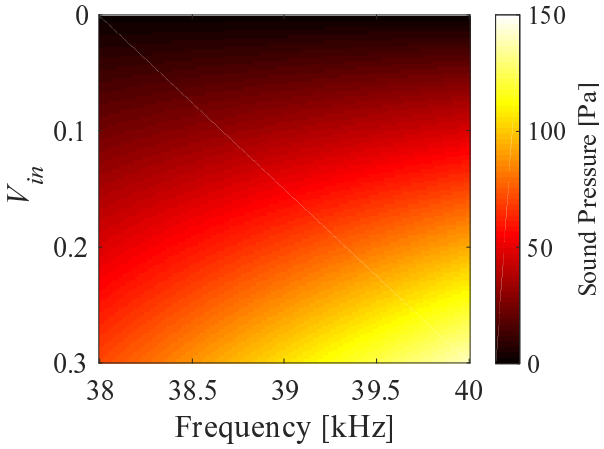



Fig. 4: スピーカ No. 2 の入出力特性の近似値.

1，1.8 kHz の成分を持つ可聴音が再生される．(a) にキャリア波と側帯波の周波数成分，(b) に再生された可聴音の周波数成分を示す．また，(I) は音圧制御をしていない場合，(II) は音圧制御を導入した場合の結果である．各周波数における音圧は SPL (Sound Pressure Level) で表している．キャリア波と側帯波の周波数成分を比較すると，音圧制御導入前はキャリア波に比べ側帯波の音圧が低いが，音圧制御を導入すると側帯波の音圧はキャリア波と同等となっていることがわかる．再生された可聴音の周波数成分を比較すると，音圧制御導入前は 1 kHz と 1.8 kHz における音圧の差が大きいが，音圧制御を導入するとその差が小さくなっている．しかし，0.8 kHz にも大きいピークを確認することができる．これは，$f_{b-1} - f_{b-2}$ 間の差音がエイリアスとして再生されたものである．同様に $f_c = 40$ kHz, $f_{b-1} = 39$ kHz，$f_{b-2} = 38$ kHz の場合の受信信号の周波数スペクトルを Fig. 6 に示す．Fig. 5 同様，キャリア波と側帯波の周波数成分では音圧制御の効果が確認

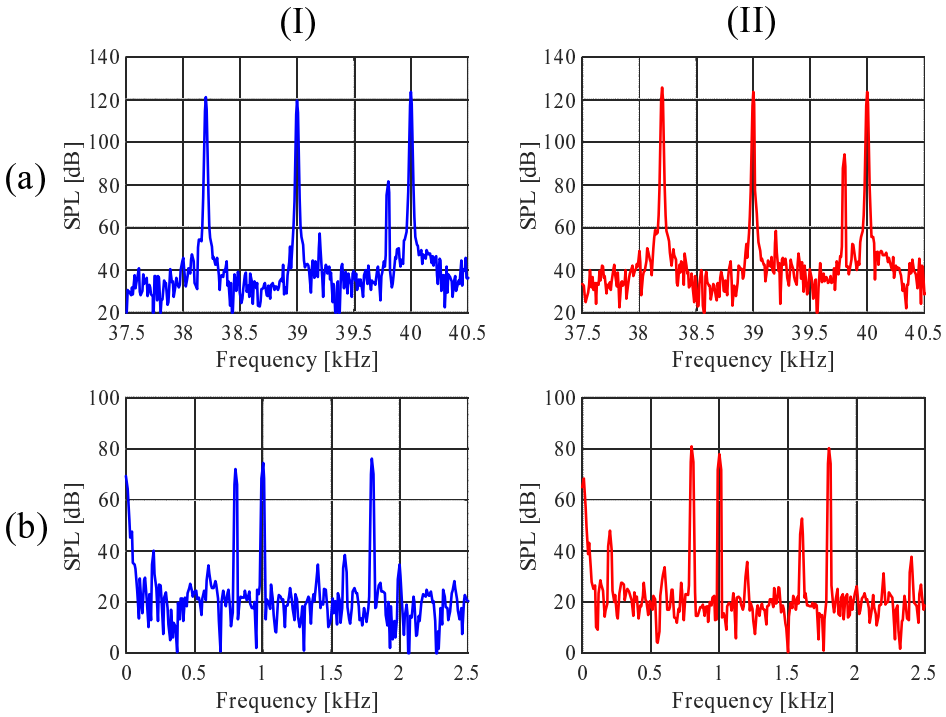



Fig. 5: Exp1: 受信信号の周波数スペクトル ($f_{b-2} = 38.2$ kHz)．(a) キャリア波と側帯波の周波数成分，(b) 再生された可聴音の周波数成分．(I) 音圧制御なし，(II) 音圧制御あり．

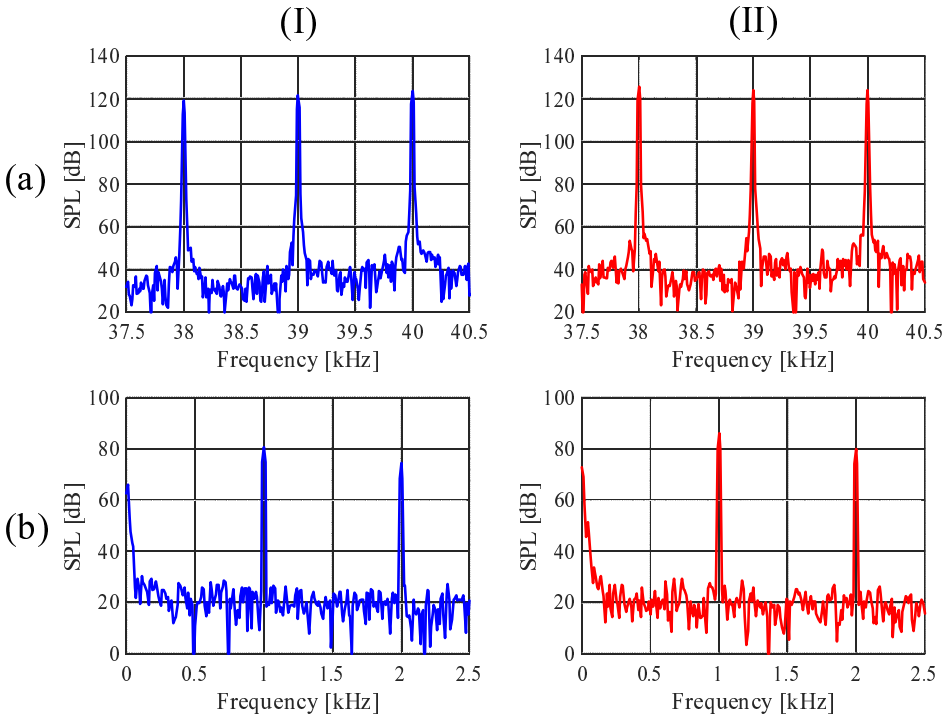



Fig. 6: Exp1: 受信信号の周波数スペクトル ($f_{b-2} = 38$ kHz)．(a) キャリア波と側帯波の周波数成分，(b) 再生された可聴音の周波数成分．(I) 音圧制御なし，(II) 音圧制御あり．

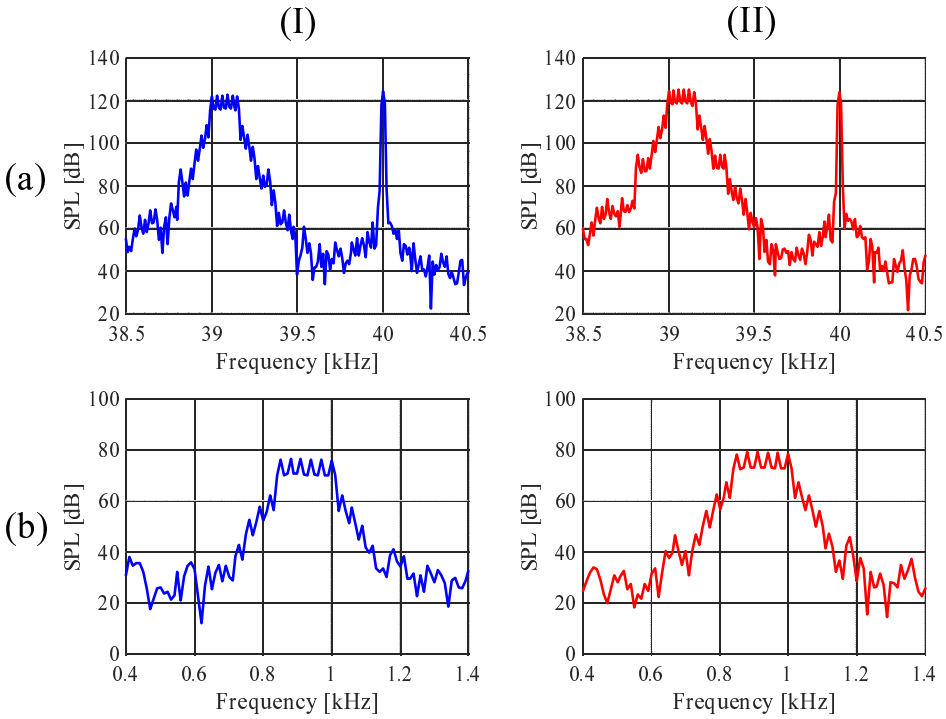



Fig. 7: Exp2: 受信信号の周波数スペクトル．(a) キャリア波と側帯波の周波数成分，(b) 再生された可聴音の周波数成分．(I) 音圧制御なし，(II) 音圧制御あり．

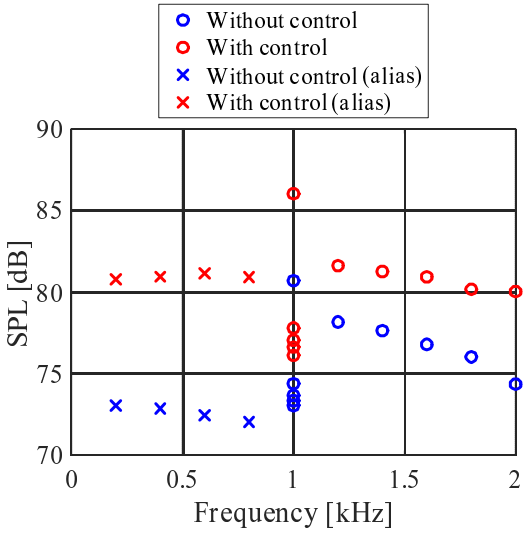



Fig. 8: Exp1: 各条件における周波数スペクトルのピーク値．

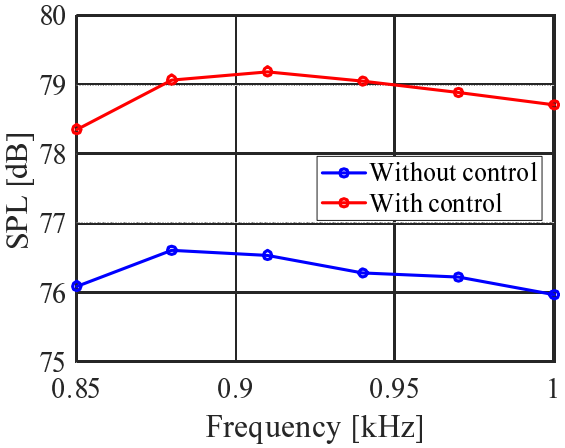



Fig. 9: Exp2: 各条件における周波数スペクトルのピーク値．

できるが，再生された可聴音の成分である 1，2 kHz における音圧の差は音圧制御を導入しても大きな変化はない．これは，再生したい成分である $f_c - f_{b-1}$ 間の差音 (1 kHz) と，エイリアスである $f_{b-1} - f_{b-2}$ 間の差音 (1 kHz) が干渉してしまっている影響だと考えられる．このように，複数の周波数を再生させる場合，エイリアスが発生してしまうという問題点が明らかになった．

続いて，Exp2 の測定結果を Fig. 7 に示す．キャリア波と側帯波の周波数成分を比較すると，39〜39.15 kHz にある側帯波成分が，音圧制御導入によりキャリア波

と同等の音圧となったことが確認できる．音圧制御導入前はキャリア波に比べ側帯波の音圧が低いが，音圧制御を導入すると側帯波の音圧はキャリア波と同等となっていることがわかる．また，再生された可聴音の周波数成分を比較すると，安定化の効果は見られないが，音圧制御導入により可聴音の音圧が全体的に大きく出力されていることがわかる．

### 3.3 考察

Exp1 にて得られた，各条件での受信信号の周波数スペクトルにおける 0〜2.5 kHz の範囲でのピーク値をプロットしたものを Fig. 8 に示す．再生させた可聴音の成分である 1.2〜2 kHz の音圧は，音圧制御をしない場合，周波数が上がるにつれて下がっていくが，音圧制御を導入するとその傾きが小さくなり，有効性が確認できる．しかし，前述したように，1.2〜2 kHz の音圧と同程度のエイリアスが 0.2〜0.8 kHz に発生してしまっている．また，1 kHz の成分もエイリアスにより不安定になっていることがわかる．Exp2 における同様の結果を Fig. 9 に示す．音圧制御導入により可聴音の音圧が全体的に大きく出力されているが，周波数変化による音圧変化の傾きは大きく変わっていない．これは，側帯波の周波数帯域が狭いため，音圧の変化も小さく，音圧制御の効果が表れにくいことが原因と考えられる．以上の結果から，提案手法の有効性と課題が明らかになった．

## 4 おわりに

本稿では，先行研究にて提案した，二つのパラメトリックスピーカを用いた局所的可聴領域形成における音圧制御手法について，より複雑な条件での評価を行った．2 つの周波数成分が含まれる可聴音を再生する場合，提案手法により音圧は安定化されたが，側帯波に含まれる二成分間の差音がエイリアスとして発生することがわかった．また，広帯域の周波数成分が含まれる可聴音を再生した場合，本実験状況では側帯波の周波数帯域の狭かったため提案手法の有効性を確認することができなかった．以上から，提案手法の有用性と課題が明らかになった．今後は，エイリアスを低減させる手法についての検討と，より広帯域な可聴音を再生させた場合の評価を行っていく予定である．

## 参考文献


[1] T. Kamakura, M. Yoneyama, K. Ikegaya: Developments of Parametric Loudspeaker for Practical Use, *Proceedings of 10th International Symposium on Nonlinear Acoustics*, pp. 147–150 (1984)

[2] Y. Fukasawa, K. Shinagawa, K. Horio, K. Mituhashi, A. Deguchi, F. Kusunoki, S. Inagaki, H. Mizoguchi: Loudspeaker array system for exhibition support into museum Effectiveness verification of sound spot for science education, *Proceedings of IEEE International Conference on Systems*, pp. 1175–1180 (2008)

[3] T. Matsui, D. Ikefuji, M. Nakayama, T. Nishiura: A design of audio spot based on separating emission of the carrier and sideband waves, *Proceedings of Meetings on Acoustics*, pp. 1–9 (2013)

[4] 袴田拓実, 干場功太郎, 土屋健伸, 遠藤信行; パラメトリックスピーカを用いた局所的可聴領域形成の検討, 電子情報通信学会ソサイエティ大会講演論文集, p. 17 (2018)

[5] T. Hakamata, H. Yamashita, K. Watanabe, K. Hoshiba, T. Tsuchiya, N. Endoh: Control of sound pressure in audible spot using parametric speakers, *Proceedings of the 23rd International Congress on Acoustics*, pp. 2690–2695 (2019)

[6] Y. Wang, M. Chen, H. Li, Z. Zhou: Defining the Parameters of Truncated Square-rooting DSB for Parametric Loudspeaker, *Proceedings of IEEE International Conference on Mechatronics and Automation*, pp. 1689–1693 (2007)

[7] M. Chen, X. Qin, L. Xu, Y. Du, L. Xu: The Distortion Analysis of the Single Side Band Method for Parametric Loudspeaker Based on Orthogonal Envelope Detection, *Proceedings of the 2nd International Symposium on Systems and Control in Aerospace and Astronautics*, pp. 1–5 (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 web page; `http://www.osaka-kyoiku.ac.jp/~challeng/`