人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-057 (11/20, 11/21)

ＡＩチャレンジ研究会(第 57 回)

Proceedings of the 54th Meeting of Special Interest Group on AI Challenges

CONTENTS

日時:2020年11月20日～21日, 場所: オンライン

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

生態系保全と AI

Ecosystems Conservation and Artificial Intelligence

田中　章

Akira Tanaka

東京都市大学環境学部環境創生学科
Department of Restoration Ecology and Built Environment,
Faculty of Environmental Studies, Tokyo City University
連絡先：東京都市大学環境学部　〒224-8551　横浜市都筑区牛久保西３－３－１
tanaka@tcu.ac.jp

Abstract

The destruction and loss of natural ecosystems on this planet have long been recognized as the greatest challenge to humankind. The main cause is human economic activities, especially development projects. In other words, the "economy" or "development projects" to enhance the well-being of human society and the preservation of nature and ecosystems, which are the basis for the survival of humans as organisms, are unfortunately still incompatible. The concept of "sustainable development," discussed at the 1992 Earth Summit, was devised to reconcile the two. The SDGs as the measures to realize them, but the global environment continues to deteriorate, and humanity has not been able to change such trend.

AI, on the other hand, is developing at a tremendous pace, and the possibilities for AI are limitless. Therefore, if it is possible to integrate ecosystem conservation and AI, which have rarely been seen on the same stage, new possibilities may be born. With this in mind, I would like to introduce my research encompassing Environmental Impact Assessment (EIA), Habitat Evaluation Procedure (HEP), Biodiversity Offset/Bank, ecological restoration, Landscape, Soundscape, Aroma-scape, green infrastructure, “Satoyama” Banking and Earth Banking.

概要

地球上の生態系の破壊や消失は人類最大の課題と認識されてから久しい。その主原因は人間による経済活動、特に開発事業である。つまり、人間社会を幸福にするための「経済」や「開発事業」と、生物としてのヒトの生存基盤である「自然」や「生態系」の保全は、残念ながら未だに両立できない関係にある。1992 年の地球サミットで議論された「持続可能な開発」は、その 2 つが両立するための概念であり、それを実現するための方策である SDGs の掛け声は高いが、依然として地球環境は悪化し続けており、人類はその傾向を変えることはできていない。一方、AI はすさまじい勢いで発展しており、AI の可能性は無限である。そこで、これまでほとんど、同じ舞台で捉えられてこなかった生態系保全と AI が融合できたなら新しい可能性が生まれるかもしれない。そんな思いから、環境アセスメント、HEP、自然復元、グリーンインフラ、里山バンク、アースバンクなどする次第である。

謝辞

自然環境保全分野とは畑違いにも拘わらず、人工知能学会の AI チャレンジ研究会の招待講演という貴重な機会をいただいた研究会幹事の方々に心より感謝申し上げます。

参考文献

[1] 『アジェンダ 21 持続可能な開発のための人類の行動計画（92'地球サミット採択文書）』（1993, 海外環境協力センター 461pp)

[2] 『ＨＥＰ入門 ―〈ハビタット評価手続き〉マニュアル― Theory and practices for Habitat Evaluation Procedure(HEP) in Japan』（2006, 朝倉書店 266pp)

[3] 相野田幸司、田中章（2017）サウンドスケープ概念の生態系評価への応用．東京都市大学横浜キャンパス情報メディアジャーナル，Vol.18，No1，p61-70

マイクロホンアレイ搭載ドローンによる音源方向尤度統合に基づく音源追跡

3D Sound Source Tracking for Drones Using Direction Likelihood Integration

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，ドローンに搭載された複数のマイクロホンアレイを用いた 3 次元音源追跡手法を提案する．一般的にマイクロホンアレイは，信号処理を通じて音源方向を推定するツールとして用いられている．マイクロホンアレイを搭載することで，ドローンは危険な災害現場で助けを呼ぶ人の捜索がすることが可能になる．多くの音源が存在する屋外環境では，音源方向の情報だけでは音源を特定するには不十分であり，代わりに音源位置を取得する必要がある．そこで，単一のマイクロホンアレイでは三次元位置を取得することは困難であるため，三角測量を用いた三次元音源追跡手法が多数報告されている．しかし，実際の屋外環境ではドローン自体からのノイズや未知の外部ノイズが存在するため，方向推定が悪化しやすく，音源方向推定に含まれる離散性も相まって，三角測量による音源位置推定もより悪化しやすい．本研究では，三角測量に代えて，音源位置の尤度分布を推定することで音源位置を追跡することを提案する．従来，各マイクロホンアレイが音源の方向を定位する際には，音源の方向の尤度分布を計算し，最も尤度の高い方向を取るのが一般的であるが，本研究では，最大尤度に着目するのではなく，全てのマイクロホンアレイの方向尤度を統合し，音源位置尤度の分布を推定しながら音源追跡を行う手法を提案する．これにより，三角測量では困難な音源位置分布を非ガウス形で表現することが可能となる．提案手法は，実在するドローンのノイズを用いて数値的に推定し，三角測量に基づく他のトラッキング手法と比較して評価した．シミュレーションの結果，提案手法の有効性が実証され，40m 先の音源を 4m 以下の RMSE で追跡できることがわかった．

1　はじめに

音響シーン解析の分野では，マイクロホンアレイを用いた音源の定位が盛んに研究されており，有用な技術として期待されている．例えば，ドローンと組み合わせることで，災害現場の瓦礫の下に埋もれている人をドローンが捜索することが可能となる [1,2]．一般に，音源の方向を推定するには，単独のマイクロホンアレイを用いるが，単独マイクロホンアレイを用いるのではなく，複数マイクロホンアレイを用いることで，推定された方向を統合して音源の位置を推定することができる．複数のマイクロホンアレイを用いて音源位置を推定する方法の一つとして，音源方向に基づく三角測量が報告されている [3–5]．しかし，ドローン聴覚の分野では，定位すべき音源がドローンから遠く離れている場合が多く，加えてドローンのノイズによって方向推定が悪化するため，三角測量が悪化する可能性がある．そこで，本稿では，方向推定ではなく，位置尤度の分布を推定することで音源位置を追跡する手法を提案する（図 1）．これにより，音源位置の分布を一般的な形で表現することができ，マイクロホンアレイのペアごとの情報ではなく，すべてのマイクロホンアレイの情報を用いた音源位置の推定を行うことができる．本稿の残りは以下のように構成されている．第 2 節では提案手法を説明し，第 3 節では数値シミュレーションによる評価を行う．最後に結論と今後の課題を述べる．

*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

本稿はシンポジウム Quiet Drones 2020 で発表した内容を和訳したものである．
https://www.quietdrones.org/conferences/1-quiet-drones-2020/

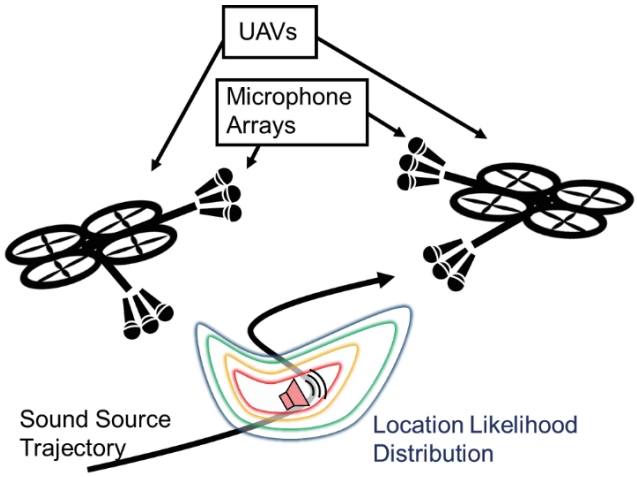


図 1: 複数マイクロホンアレイ搭載ドローンを用いた音源追跡

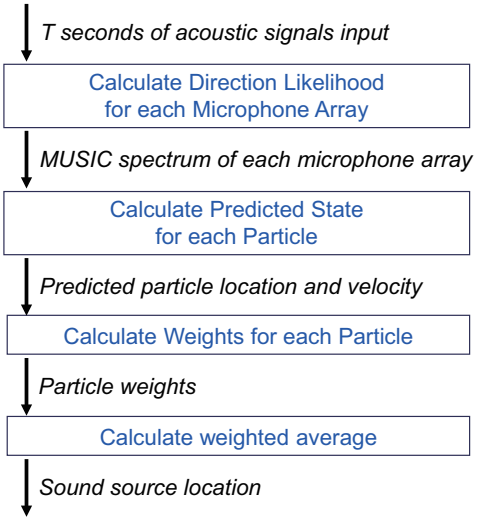


図 2: 追跡手法のフロー (1 サイクル分)

2 三次元音源追跡

本節では，位置尤度推定に基づく音源追跡手法を提案する．音源位置尤度分布を，各マイクロホンアレイから得られた音源方向尤度分布を統合することで求め，得られた音源位置尤度を元に Particle Filter を通じて音源位置を追跡する．

2.1 問題設定

複数マイクロホンアレイを用いた音源位置追跡を考える．マイクロホンアレイは全部で N 個存在すると仮定し，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 MA_n の時刻 t における状態を

$$\boldsymbol{m}_n(t) = [\boldsymbol{m}_{n,xyz}^T, \boldsymbol{m}_{n,\phi\theta\psi}^T]^T \quad (1)$$

$$\boldsymbol{m}_{n,xyz} = [x_n(t), y_n(t), z_n(t)]^T \quad (2)$$

$$\boldsymbol{m}_{n,\phi\theta\psi} = [\phi_n(t), \theta_n(t), \psi_n(t)]^T \quad (3)$$

とおき，既知であるとする．$x_n(t), y_n(t), z_n(t)$ は 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 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，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．追跡する音源は点音源であると仮定し，音源の三次元座標は，

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

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

本手法は方向推定手法によって得られる方向尤度分布の統合に基づいており，図 2 に提案追跡手法のフローを示している．一般的に，音源方向推定を行う際は，音源方向尤度 $P(\phi, \theta)$ を計算し，$P(\phi, \theta)$ の最大値をとるような方位角 ϕ，仰角 θ を推定方向としている． 本稿では各マイクロホンアレイで得られた音源方向尤度 $P(\phi, \theta)$ を統合し，三次元音源位置尤度として変換することで，三次元音源位置の追跡を図る．推定された三次元音源位置尤度は Particle filter に適用し，各粒子に位置尤度を重みとして与えることで，音源軌跡を推定する．

2.2.1 方向尤度分布の推定

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 [6] 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 [7] 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．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 [10] 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 ω, 方位角 ϕ, 仰角 θ 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, (ϕ,θ) における空間スペクトル $P(\phi,\theta)$ は

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

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, ω_L,ω_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 MA_n で求めた MUSIC スペクトルを $P_n(\phi,\theta)$ と記述し, 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, 音源位置尤度分布を表現することで, 音源位置を追跡する. 任意の三次元位置 $\boldsymbol{x}$ において，$\boldsymbol{m}_{n,xyz}$ から x の方向を (ϕ_n,θ_n) とおく．このとき，三次元位置 $\boldsymbol{x}$ における音源位置尤度は以下のように定義する．

$$L(\boldsymbol{x}) = \sum_n P(\tilde{\phi}_{ni}^{\mathrm{round}},\tilde{\theta}_{ni}^{\mathrm{round}}) \quad (6)$$

$$\tilde{\phi}_{ni}^{\mathrm{round}} = \mathrm{round}(\tilde{\phi}_{ni}),\quad \tilde{\theta}_{ni}^{\mathrm{round}} = \mathrm{round}(\tilde{\theta}_{ni}) \quad (7)$$

$$\begin{bmatrix}\cos\tilde{\phi}_{ni}\cos\tilde{\theta}_{ni}\\ \sin\tilde{\phi}_{ni}\cos\tilde{\theta}_{ni}\\ \sin\tilde{\theta}_{ni}\end{bmatrix} = \boldsymbol{R}_n^{-1}\begin{bmatrix}\cos\phi_{ni}\cos\theta_{ni}\\ \sin\phi_{ni}\cos\theta_{ni}\\ \sin\theta_{ni}\end{bmatrix} \quad (8)$$

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 (ϕ_{ni},θ_{ni}) を丸める関数であり, $\boldsymbol{R}_n$ は 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点 $\boldsymbol{x}$ 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, 点 x の音源位置尤度としている.

2.2.3　音源位置尤度分布に基づく追跡

本手法では，前小節で求めた音源位置尤度 $L(\boldsymbol{x})$ を Particle filter に適用し，音源位置を追跡する．パーティクルフィルタに用いるパーティクルの個数を I とし，時刻 k におけるパーティクル i の状態と重みをそれぞれ $\boldsymbol{x}_k^i, w_k^i$ とする．状態 $\boldsymbol{x}_k^i$ はパーティクルの三次元位置と速度を含み，以下のように記述する．

$$\boldsymbol{x}_k^i = \left[x_k^i, y_k^i, z_k^i, \dot{x}_k^i, \dot{y}_k^i, \dot{z}_k^i\right]^T \quad (9)$$

また，パーティクルは excitation-damping モデル [9] に従うと仮定し，以下のような挙動を示すと仮定する．

$$\boldsymbol{x}_k^i = F\boldsymbol{x}_{k-1}^i + H\boldsymbol{v} \quad (10)$$

$$F = \begin{bmatrix}\boldsymbol{I} & T\boldsymbol{I}\\ \boldsymbol{O} & a\boldsymbol{I}\end{bmatrix},\qquad H = \begin{bmatrix}\boldsymbol{O}\\ b\boldsymbol{I}\end{bmatrix}, \quad (11)$$

$$\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}) \quad (12)$$

ここで，$\boldsymbol{I}\in\mathbb{R}^3$ は単位行列，$\boldsymbol{O}$ はゼロ行列を示す．各パーティクルは $L(\boldsymbol{x}_{k,pos}^i)$ に比例した重みが与えられるため，重みは下式で表される．

$$w_k^i = w_{k-1}^i\frac{L(\boldsymbol{x}_{k,pos}^i)}{\sum_i L(\boldsymbol{x}_{k,pos}^i)} \quad (13)$$

ここで，$\boldsymbol{x}_{k,pos}^i) = [x_k^i, y_k^i, z_k^i]^T$ である．リサンプリングは有効パーティクル数が閾値 N_{thr} を下回ったときに行う．つまり，下式の条件が満たされるとき，各パーティクルの重みは $1/I$ にリセットされる．

$$\frac{1}{\sum_i (w_k^i)^2} \le N_{thr} \quad (14)$$

パーティクルの初期化は，以下の分布からパーティクルをサンプルすることで得られる．

$$\boldsymbol{x}_0^i \sim \mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \quad (15)$$

$$\boldsymbol{\mu}_0 = [\boldsymbol{\mu}_{0,pos}^T, 0, 0, 0]^T \quad (16)$$

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix}\sigma_{pos}^2\boldsymbol{I} & \boldsymbol{O}\\ \boldsymbol{O} & \sigma_{vel}^2\boldsymbol{I}\end{bmatrix} \quad (17)$$

ここで，$\boldsymbol{\mu}_0$ は三角測量を用いることで得る三次元位置である [4,5]．具体的には，複数の三角測量点を計算し，全ての三角測量点の平均点を $\boldsymbol{\mu}_0$ とおく．

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　シミュレーション設定

図 3 のように，2 台のドローン（黒点）に 2 個ずつマイクロホンアレイ（黒丸）を取り付け，高さ 30m でホバ

表 1: 音源追跡に用いた変数値

Variables	Value
a	0.5
b	3
I	500
N_{thr}	350
σ^2_{pos}	25
σ^2_{vel}	25

リングして停止しているシナリオを考える．追跡する音源（赤線）は半径 5m の円を描くように等速で移動する．音源の速さは π m/s であり，10 秒で 1 周する．円軌道の中心と両ドローンとの水平距離は $l = 10, 20, 30, 40, 50$ m の 5 種類のパターンに対してシミュレーションを行った．各マイクロホンアレイは図 4 のような球型マイクロホンであり，1 つのマイクロホンアレイは $M = 16$ 個のマイクロホンから成っている．音響信号は 16 kHz, 24 bit で収録される．音源位置は 0.2 秒ごとに 10 秒間推定されることで，音源軌跡を推定する．MUSIC スペクトルは水平角，仰角共に 5 度刻みで推定される．提案手法に用いられる各種パラメータは表 1 に示す．音源は日本音響学会 新聞記事読み上げ音声コーパス (JNAS) 内の男声コーパス・女声コーパス各 10 種，ホワイトノイズ 10 種の計 30 種を出力させ，各種音源に対してシミュレーションを行った．また，他手法との比較のため，三角測量点の平均点に対してカルマンフィルタを適用した手法 [8] と推定音源方向を元にパーティクルフィルタで推定する手法 [9] を実装し，同じ条件の元でシミュレーションを行った．

3.2 結果

シミュレーション結果を図 5 と図 6 に示す．いずれの図からも，提案手法が音源軌跡の概形を追跡できていることがわかる．推定序盤では z 軸方向に推定誤差が大きく発生してるが，これはパーティクル初期化のために行った三角測量がドローンノイズの影響で不安定であり，初期位置が音源に近くないためである．しかし，z 軸の初期位置が外れているのにもかかわらず，提案手法によって真値に収束している様子が図 6 で見られ，提案手法が移動音源に対して有効であることが分かる．追従誤差は表 2-4 に示しており，表は以下のように定義される Root Mean Squared Error (RMSE) を比較したものである．

$$\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum \mathrm{error}^2} \tag{18}$$

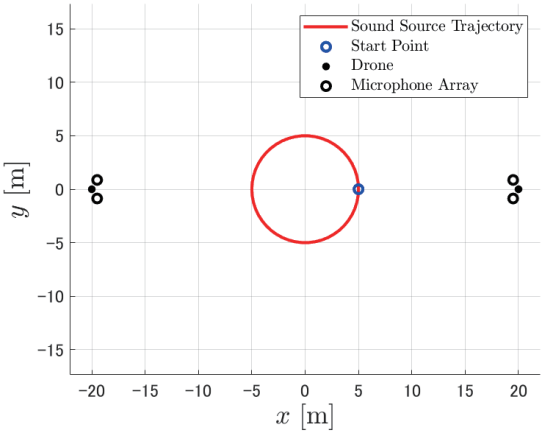


図 3: 上から見たシミュレーションシナリオ ($l = 20$ m)

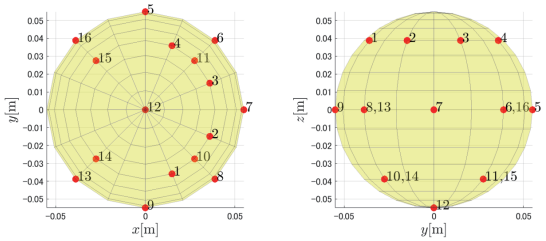


図 4: マイクロホンアレイ中のマイクロホン配置

ここで，error は推定位置と真値のユークリッド誤差を指しており，K は一シミュレーション中のタイムステップ数を示している．また，表 2-4 中の太字で示されている数値は，同じ水平距離において最も小さい RMSE を指している．表 2-4 より，提案手法は他手法と比べて最も小さい誤差で音源追跡を行ったことが分かる．これは提案手法は三角測量点を求めるのではなく，位置尤度分布を推定しようとするため，三角測量の離散性より生じる外れ値の影響を受けにくくなっているからだと考えられる．また，他手法と同じく，音源距離が長くなるにつれて RMSE が増加しているが，これは方向尤度に分解能があるため，位置尤度分布にも離散性が生じるからである．

4 終わりに

本稿では，音源の方向尤度分布を位置尤度として，統合することで音源追跡を行う手法を提案した．音源位置尤度分布は各マイクロホンアレイから得られる方向尤度分布を足し合わせることで得られる．音源位置尤度分布を元に音源追跡を行うことで，音源方向推定誤差や三角測量誤差が発生しやすいケースに対してロバストな追跡が行えることが期待される．提案手法を数

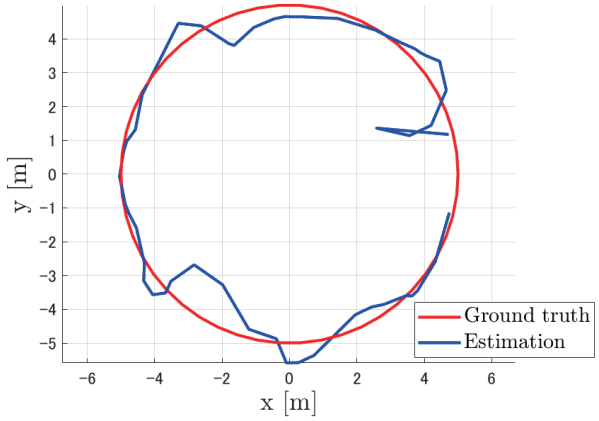


図 5: 上から見た推定軌跡

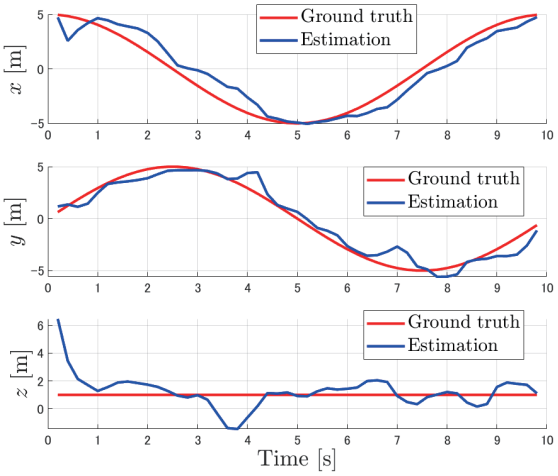


図 6: 各軸における推定結果

表 2: 各音源水平距離の追跡誤差 (RMSE) (女声)

	10 m	20 m	30 m	40 m	50 m
Proposed method	**1.66**	**2.43**	**1.75**	**2.89**	**5.61**
Method [8]	12.47	13.76	14.15	14.68	16.43
Method [9]	4.97	4.01	3.85	4.30	5.69

表 3: 各音源水平距離の追跡誤差 (RMSE) (男声)

	10 m	20 m	30 m	40 m	50 m
Proposed method	4.80	**4.10**	4.18	**3.92**	**4.32**
Method [8]	12.47	13.76	14.34	15.17	16.27
Method [9]	**4.77**	4.20	**3.98**	4.26	5.32

表 4: 各音源水平距離の追跡誤差 (RMSE) (ホワイトノイズ)

	10 m	20 m	30 m	40 m	50 m
Proposed method	**1.81**	2.39	**1.93**	**2.26**	**2.95**
Method [8]	8.73	10.57	10.87	11.16	11.32
Method [9]	2.14	**1.81**	2.14	2.55	3.66

値シミュレーションにより評価した結果，提案手法は 40m 離れた音源に対して 4m 以下の誤差で追跡できることがわかった．実環境での評価や，音源分離・音源認識を加えた音環境理解システムへの拡張は今後の課題である．

謝辞

科研費 JP19K12017, JP19KK0260 および JP20H00475 の助成を受けた．

参考文献

[1] Nakadai, K. et al. (2017). Development of microphone-array-embedded UAV for search and rescue task, 2017 IEEE International Conference on Intelligent Robots and Systems (IROS), pp. 5985–5990

[2] Washizaki, K., Wakabayashi, M., and Kumon, M. (2016). Position estimation of sound source on ground by multirotor helicopter with microphone array, 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1980-1985

[3] Brandstein, M. S., and Silverman, H. F. (1997). A practical methodology for speech source localization with microphone arrays, Computer speech & language, 11(2), 91-126.

[4] Gabriel, D., Kojima, R., Hoshiba, K., Itoyama, K., Nishida, K., and Nakadai, K. (2019). 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system, Advanced Robotics, 33(7-8), pp. 403-414.

[5] Yamada, T., Itoyama, K., Nishida, K., and Nakadai, K. (2020). Sound Source Tracking by Drones with Microphone Arrays, IEEE/SICE International Symposium on System Integration (SII), pp. 796-801

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

[7] Valin, J. M., Michaud, F., and Rouat, J. (2006). Robust 3D localization and tracking of sound sources using beamforming and particle filtering, 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (Vol. 4, pp. IV-IV).

[8] Potamitis, I., Chen, H., and Tremoulis, G. (2004). Tracking of multiple moving speakers with multiple microphone arrays, IEEE Transactions on Speech and Audio Processing, 12(5), pp. 520-529.

[9] Lauzon, J. S., Grondin, F., Létourneau, D., Desbiens, A. L., and Michaud, F. (2017). Localization of RW-UAVs using particle filtering over distributed microphone arrays, 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 2479-2484)

[10] Schmidt, R. (1986). Multiple emitter location and signal parameter estimation, IEEE transactions on antennas and propagation, 34(3), pp. 276-280.

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-057-3 (11/20)

オンライン音環境認識のための低次元埋め込み手法の高速化

Acceleration of Low-dimensional Embedding Method for Online Auditory Scene Analysis

新里 顕大[1*]　小島 諒介[2]
Kenta Shinzato[1]　Ryosuke Kojima[2]

[1] 京都大学工学部情報学科数理工学コース
[1] Undergraduate School of Informatics and Mathematical Science, Kyoto University
[2] 京都大学大学院医学研究科ビッグデータ医科学分野
[2] Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto University

Abstract: Summarizing and visualizing auditory scenes are essential tasks for environmental monitoring applications. In this paper, we address the mapping of sound events into low-dimensional space by unsupervised manner considering long-term recordings. Because recorded data is accumulated continuously or periodically in long-term recording data analysis, computational costs and the incremental nature of the analysis methods are significant. We propose an unsupervised auditory scene analysis system, which uses an incremental low-dimensional embedding technique SONG for visualization. To realize real-time analysis by using this method, we improve the implementation of SONG algorithm by accelerating the algorithm utilizing just-in-time compile and matrix computation techniques. In this study, we investigated the bottleneck of this algorithm by the profiling and quantitatively evaluated the effectiveness of our accelerated implementation.

1　はじめに

長期的な環境モニタリングは，継続的な環境調査や動物の生態調査において重要である．特に，環境中の音をマイクロフォンで収録・分析することは，視界の悪い環境など視覚ベースのセンサが効果的でないような場合にも有効であることから注目されている．一方で，実環境には様々なノイズが含まれるため，収録した音から有用な音の情報を分離・抽出することは重要な研究課題である．特に音環境認識では，複数のマイクロフォンからなるマイクロホンアレイを用いることでより多くの情報を抽出することが可能である [1]．マイクロホンアレイを用いた研究は，ロボット聴覚の分野で多くなされており，例として Voice Activity Detection (VAD), Sound Source Localization (SSL), Sound Source Separation (SSS), Sound Source Identification (SSI) などの技術が挙げられる [2, 3]．VAD, SSL, SSS は古くから信号処理の分野で研究されており，教師あり・教師なし機械学習によるアプローチも研究されている．一方で，SSI に関しては教師あり機械学習の手法を用いることが一般的である．しかし，教師ありアプローチでは, 対象音源が明確にカテゴリ分類されている必要があり，それらの教師データは予めアノテーションをする必要があり，これを人間が行う場合には，その労力や正確性が問題となる．

近年，事前にカテゴリ分類がされていない，あるいは曖昧な音源を認識する手法として低次元埋め込みの手法が注目されている [4, 5, 6]．低次元埋め込みでは，高次元のベクトルを低次元，特に可視化の用途では 2 次元や 3 次元のベクトルとして表現することを考える．低次元埋め込みを使った可視化は既に，バイオインフォマティクスやデータサイエンスの分野などで広く応用されている [7, 8]．

我々は，野生生物のモニタリングや調査を想定した低次元埋め込みを用いた音環境モニタリングシステムを提案する．野生生物の音環境モニタリングでは，環境変化の調査や希少な事象の調査のために長時間の記録が欠かせないが，このような用途では，インクリメンタル性が重要である．インクリメンタルな手法を考える理由の一つは長時間の録音では季節や時間帯の変化などにより環境が動的に変化することで， 過去のデータには存在しない，新たな音源クラスが生じることを

*連絡先： 京都大学工学部情報学科数理工学コース
京都市左京区聖護院川原町 54
E-mail: sinzato.kenta.82r@st.kyoto-u.ac.jp

考慮する必要があるためである．加えて，長時間の連続録音では録音された音声データのサイズが膨大であるため，全データを一度に処理し，解析することは現実的ではない．また，インクリメンタルな処理を実現することで，重要なイベントのみをデータとして蓄積するといった応用も可能である．

低次元埋め込みアルゴリズムの進歩は目覚ましく，特に t-SNE [5] や UMAP [6] は，データの可視化，すなわち二次元への埋め込みを行うためのスタンダードな手法となっている一方で，従来手法の多くは，インクリメンタルに設計されていないため，新たなデータが与えられた際にすべてのデータについて再計算を行わなければならない．近年，低次元への埋め込みをインクリメンタルに構築できる Self-Organizing Nebulous Growths (SONG) [9] が提案され，これにより，t-SNE や UMAP に匹敵する可視化がインクリメンタルに可能であると報告されている．

長時間録音に対しインクリメンタルな音環境認識のシステムを構築するにあたり，求められる性能としてインクリメンタル性に加え, リアルタイム性がある．インクリメンタルな音響解析では，収録した音の長さと同程度の処理時間で低次元埋め込み処理が実行できることが理想的なので， システムの各処理は高速に動作することが求められる．本研究では，音環境認識を想定した埋め込みアルゴリズムの各処理に要する時間の分析を行い，実行速度面で改善が可能である部分に関し最適化を施し，評価を行う．

2 収録音の低次元埋め込みおよび可視化

低次元埋め込みを用いた環境音モニタリングシステム [10] は前処理・特徴抽出と可視化のための低次元埋め込みの 2 つのステップから構成される．まず，2.1 章で収録データから特徴量を抽出する方法について述べ，2.2 章でインクリメンタルな低次元埋め込みおよび可視化のための SONG 法について述べる．

2.1 前処理

我々は，これまでにインクリメンタルなマイクロホンアレイで収録を行い，定位，分離，可視化を行うデータフロー的な設計を行ったシステムを開発した [10]．本研究では，このシステムに組み込むことを想定し，特に，低次元埋め込みとその前処理部分に関する検討を行う．以降では分離後の音データから特徴量を計算する方法について述べる．

まず，音のフレーム毎の音響特徴量を抽出する．本稿では，メルスペクトログラムによる音響特徴量抽出を利用する．メルスペクトログラムは，メルスケーリングされたスペクトログラムであり，本研究で用いる特徴量はメルスペクトログラムの 2 つの隣接フレーム間の差のデルタ特徴とそのさらに差分の二重デルタ特徴を合わせて用いる．その後, スライディングウィンドウを利用して複数のフレームに関連付けられた特徴を構造化する．本稿における実験では, 10 フレームの窓を使用しているため, 最終的な特徴ベクトルは各フレームの特徴ベクトルの 10 倍の長さになる．本稿では, このスライディングウィンドウによって計算されたベクトルを「サンプル」と呼ぶことにする．実際の次元数や詳細は 4.2 章で述べる．

各サンプルの低次元埋め込みを行う前に，上述の特徴ベクトルに対し正規化と次元削減を行う．これらの前処理は低次元埋め込みを効果的に計算するために経験的に重要である [5]．我々のプロトタイプの設計では，PCA (Principal Component Analysis) を用いて特徴ベクトルを 20 次元のベクトルに削減している．この PCA によってパラメータを得る過程はインクリメンタルではないが，インクリメンタルな設計の PCA [11] を利用することで，真にインクリメンタルなシステムを構築できる．

2.2 SONG アルゴリズムの概要

Self-Organizing Nebulous Growths (SONG) [9] は，t-SNE や UMAP 等と同等に高次元のデータを低次元空間に埋め込むアルゴリズムであり，主に高次元でのデータを可視化するために用いられる． これらのアルゴリズムではデータ間の距離に基づいて，類似するデータ同士が低次元空間上に射影された際にも近距離になるよう埋め込みを行う．SONG は t-SNE や UMAP とは異なり，パラメトリックな手法と呼ばれている．SONG では埋め込みを行う際にパラメータを利用しており，一度埋め込みを計算した後に新たにデータを追加した場合にも，そのパラメータを用いて新たな埋め込みが容易に追加可能というインクリメンタルな性質を備えている．

SONG アルゴリズムでは，入力次元のユークリッド空間中にコーディングベクトル（以下 CV と呼ぶ）と呼ばれる点を複数用意し， データのクラスターを代表する点として扱う．2 つの CV の間には非負の値として関連度を設け，ある入力データの k-近傍以内の CV 同士の関連度を高くし，入力データの近くにそのデータに関連付けられた CV を移動させることで，関連する CV 同士がより密になるようにする．逆に，着目データ点の k-近傍でない CV と最近傍の CV とは関連度を下げる．このように，関連度の高い CV 同士が近くなり，関連しない CV 同士は離れるように CV を配置し，

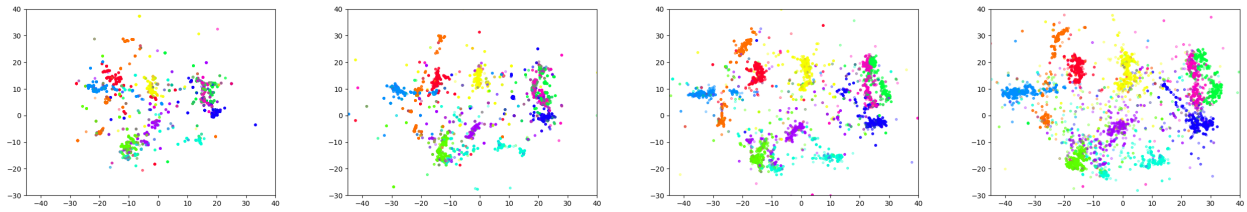


図 1: SONG 法により MNIST データが可視化される様子. 左から順に 10, 20, 50, 100 エポック後の埋め込み.

各 CV から低次元空間の点への写像を構成することで, 低次元のへの埋め込みを実現する.

より具体的には, SONG アルゴリズムは以下の 4 つのステップから構成され, それらを繰り返すことで低次元のへの埋め込みを行う. 初期時では, CV および埋め込み点はランダムに初期化されている.

ステップ A 入力データからランダムに一点選び $\boldsymbol{x}$ とする. $\boldsymbol{x}$ の 最近傍 の CV である $\boldsymbol{c}'$ と, その他の CV との間の関連度を更新する. 具体的には, $\boldsymbol{x}$ の k-近傍以内の点と $\boldsymbol{c}'$ との関連度を 1 にし, それ以外の点とは関連度を $\varepsilon(<1)$ 倍し, さらにその操作によって関連度が閾値以下になった場合は関連度を 0 にする.

ステップ B ある コーディングベクトル $\boldsymbol{c}$ との関連度が 0 ではない CV の集合を $R(\boldsymbol{c})$ で表し, これらを $\boldsymbol{c}$ と関連する点と呼ぶ. $R(\boldsymbol{c}')$ の各点を $\boldsymbol{x}$ に近くなるように移動させる. この際, $\boldsymbol{x}$ の近傍の点ほど大きく移動させることで関連度の情報を崩さないようにする. これには適切な損失関数を設定し, 最急降下法で最適化する.

ステップ C $\boldsymbol{c}'$ に対応する低次元空間の埋め込み点を $\boldsymbol{y}'$ とする. $\boldsymbol{y}'$ とその他の埋め込み点の位置を関連度の情報をもとに更新する. 具体的には, $\boldsymbol{y}'$ と, $R(\boldsymbol{c}')$ に対応する埋め込み点との距離が関連度に応じて近くなるようにする. 一方で $\boldsymbol{c}'$ と関連しない点 (CV の $R(\boldsymbol{c}')$ の補集合 $\overline{R(\boldsymbol{c}')}$ で表される点) からランダムにサンプリングしたコーディングベクトルに対応する埋め込み点と $\boldsymbol{y}'$ は離れるように, それぞれ適切な損失関数を設定し, 最急降下法で各埋め込み点の位置を最適化する.

ステップ D CV の数が不足している場合には, $\boldsymbol{x}$ と $\boldsymbol{x}$ の k-近傍の CV の重心座標に新たな CV を追加し, 同様に埋め込み点も追加する. この操作は, 各 $\boldsymbol{x}$ に対し $\|\boldsymbol{x}-\boldsymbol{c}'\|$ の累積が閾値を超えた場合に行う.

このステップ A-D を, 全入力データに対し行う操作を 1 エポックと数える.

我々は SONG の提案論文 [9] を元に, 独自に NumPy を用いた Python 実装を行い, これを公開している[1].

3 SONG 法の分析と高速化

ここでは, SONG 法のナイーブな実装 (以降, 初期実装と呼ぶ) とその改善した実装に関して, SONG 法の各処理における計算時間の分析を行う.

初期実装とボトルネックについて改善を施したコードに対しプロファイルをとり, 各処理に要する累積時間が全体に占める割合を表した結果が図 2 である. 図中の A-C は 2.2 章の各ステップと対応する. A-1 は, ステップ A において $\boldsymbol{x}$ の k-近傍のコーディングベクトルを探索する関数の累積時間の割合を示す. A-2 は, ステップ A において $\boldsymbol{c}'$ とその他のコーディングベクトルとの関連度を更新するのに要した時間の割合を表す. 速度の計測実験は, MNIST データセットから 10,000 個サンプリングしたデータの次元を PCA で 20 次元に削減し, ノルムの正規化を施したものに対して, SONG 法を 100 エポック適用するという条件の下で行った.

実行速度においてボトルネックだった部分を解消するために, 我々は 1) 最急降下法の実装の改善 2) 実行時コンパイラ (Jsut-In-Time Compiler) による高速化を行い, 速度の向上を検証した.

以下では, 本研究で実施した実装の改善について述べる. 2.2 章のステップ B, C において, 初期実装では各 $R(\boldsymbol{c}')$ および $\overline{R(\boldsymbol{c}')}$ の要素に対して一つずつ最急降下法の処理を実行していたが, このプロセスを行列演算として実行することにより, ボトルネックであったステップ B, C の速度を改善するに至った. その結果が図 2 の SGD optimization に相当する. また, 行列計算化した最急降下法の処理を Numba [12] を用いた実行時コンパイラを適用することで, わずかな速度向上が見られた. この時, 各実行時間の割合は図 2 の JIT に示した通りであった. このとき, ステップ A については特に改善を施していないため変化は見られないが, ステップ B については 9 倍の速度向上がみられた.

[1] https://github.com/hoppiece/song

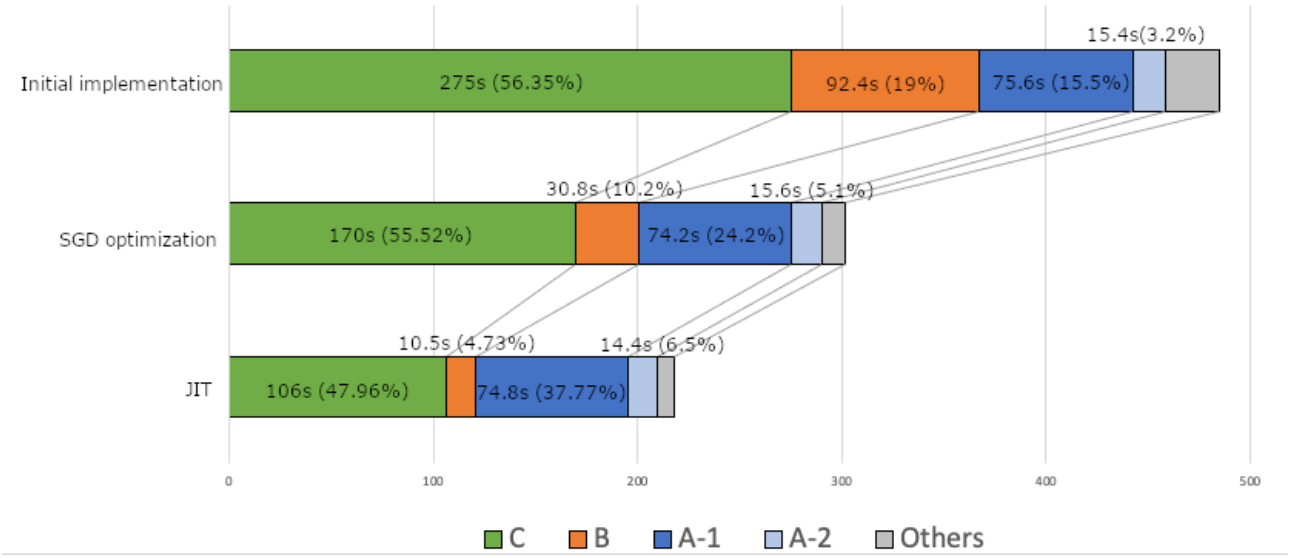


図 2: 各実装における SONG の各ステップに要した累積時間

本手法によって，全体でおおよそ 2 倍の高速化が実現できたが，SONG 法は UMAP と類似の操作で構成されており，実行時コンパイラによってさらに高度な高速化を行うと UMAP と同様にさらに高速化可能であると考えられる．この SONG 法のさらなる高速化法については 5 章で考察する．

4 実験

4.1 MNIST による評価

MNIST データセット [13] において， エポックごとに埋め込みが自己組織化される様子が図 1 である．我々の以前の研究 [10] では, SONG 法による MNIST データセットの埋め込みと，その他の埋め込み手法に対し比較を行った．

4.2 SONG 法による音響データの可視化

本システムを評価するために鳥の歌の音声データセットを用意し，音響特徴量を用いた低次元埋め込みを行う．まず，既報論文で利用されているものと同様のデータセットを構築する [4]．単一のマイクで録音された鳥の鳴き声のデータセットが `http://taylor0.biology.ucla.edu/birdDBQuery/` [14] から利用可能である．用いたデータセット内には 4 種の鳥の鳴き声を含む 645 の録音ファイルが収録されている．このデータセットには，鳥の鳴き声に関する音節のセグメンテーションとアノテーションが含まれているが， 本研究では音響特徴の埋め込みに焦点を当てているため，セグメンテーションは正しく行われているものとしてデータセット中のセグメントを用いた．また，音響特徴を抽出するには短すぎるセグメントは削除した．このデータセットから抽出された音響データの大部分は 50 フレーム未満であったので，50 フレームのデータを用いた．最終的に，こうして得られた音データから，2.1 章に記したメルスペクトラムを用いた特徴ベクトルを 349386 個のベクトルを用意した．

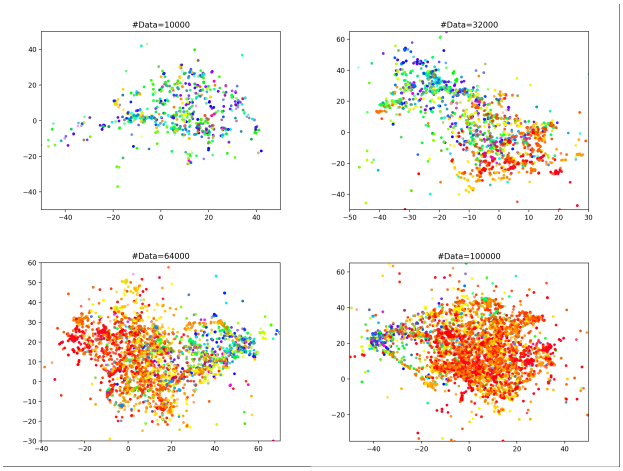


図 3: 鳥の鳴き声データの SONG 法による可視化．色は音節の種類を表す．左上，右上，左下，右下の順にデータ数が 10000, 32000, 64000, 100000 に対応する．

このデータから実行速度を評価するために，N 個のベクトル $(N = 10000, 32000, 64000, 100000)$ を取り出して，SONG 法を用いて可視化したものが図 3 となる．

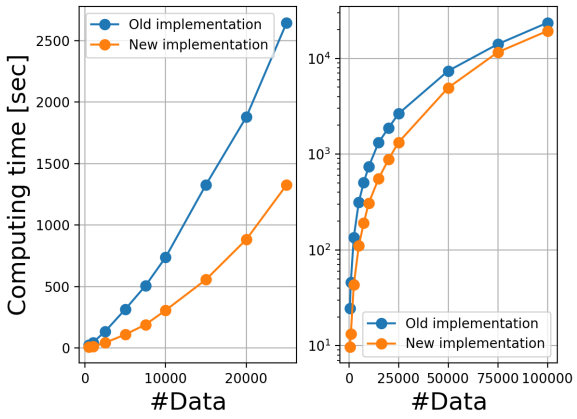


図 4: 入力データ数に応じた計算時間

上述のデータセットに関して，データ量に対しどのように増加するかを計測した結果が図 4 となる．実装は図 2 の JIT のものを用いた．データ数が 10^4 程度まではデータ数に対し線形よりやや悪いペースで実行時間が増加しているが，その後，計算時間が悪化している．その原因を分析するため様々なデータ数に対してプロファイルをとったところ， データ数が 10^4 未満の場合図 2 の JIT と同様の累積計算時間の割合を示したが，データ数が $30,000$ 程度からは，ステップ A の処理に大きく時間がかかっていることがわかった．データ数 50,000 の際の際にプロファイルをとったところ, 全体に要した時間 9770s に対して，ステップ A-1(近傍探索) が 6980 sec (70.2%) を占め，ステップ C が 2520 sec (25.31%)，ステップ B は 104 sec (1.04%)，ステップ A-2 は 181 sec (1.82%) となった．このことからデータ数が大きい際の近傍探索の速度を改善が今後の速度改善に関して新たな課題であることが分かったが，これについては 5 章で考察する．

5 考察

今回，特に基本的かつ効果的な 2 つの高速化を行った．SONG 法についてさらなる高速化の方法として次の 2 つを適用することが考えられる．

1. バッチ処理あるいは並列化による高速化
2. 近傍探索手法の高速化

の 2 つである．

バッチ化 (並列化) による高速化　有効な埋め込みアルゴリズム高速化の手法として，バッチ処理による高速化が考えられる．この方法では, 主にステップ B およびステップ C の計算に要する時間を改善することができる．2.2 のアルゴリズムでは，入力データ $\boldsymbol{x}$ を一つずつ選んで処理しているが，この処理をバッチ化し，複数のデータをまとめて扱うことで行列計算ライブラリの恩恵を受けることができる．しかし，$R(\boldsymbol{c}')$ の要素数が $\boldsymbol{x}$ ごとに異なるので，単純なテンソル計算に落とし込むには工夫が必要である．

類似のアプローチとして，バッチ内の $\boldsymbol{x}$ ごとに並列的に処理を行うという手法が考えられる．入力データが多い際に各 $\boldsymbol{x}$ を独立に処理できるので，並列化を用いた手法は計算機の並列演算性能に応じて高速化可能である．近年のニューラルネットワークの飛躍的な性能向上には GPU コンピューティングの利用が大きく寄与しているが，t-SNE や UMAP などの埋め込みアルゴリズにおいても GPU を利用した高速化の研究がされており [15, 16]，特にサンプル数が大きい ($> 10^5$) 場合に，既存の実装の 100 倍以上の高速化を実現している．SONG 法においても並列処理に GPU コンピューティングを用いることで，インクリメンタルな可視化の飛躍的な速度向上が期待できる．

近傍探索の高速化　SONG の高速化の方法として，図 2 での実験では全体の 37.7%，さらにデータ数 50,000 の際には全体の 70.2% を占める近傍探索の改善が重要である．2.2 節のアルゴリズム中のステップ A では, 入力データ近傍のコーディングベクトルを探索している．この近傍探索の速度を改善することは, 大規模なデータを可視化する際に重要であるほか，同様の改善は UMAP をはじめとした近傍探索を行う多くの埋め込み手法の改善にも応用可能である．

計算量に関して考えると，入力データ数を N，入力データの次元を D として，コーディングベクトルは入力データを代表する点のため，データ数に応じて増加することが経験的にわかっているので， コーディングベクトルの数を $O(N)$ と考えてもよい．

初期実装は線形探索により，クエリ点とすべてのコーディングベクトルとの距離を計算し，最小となるものを探索しているため, $O(ND)$ の計算量を要している．およそ $D < 30$ 以下の場合には，ユークリッド空間中の近傍探索を行う方法である KD 木 [17] や Ball 木 [18] を利用した近傍探索が有効であることが知られている．KD 木は探索空間を次元毎に二分割することで探索の効率を図っており，Ball 木は探索空間を次元毎ではなく超球面に沿って分割することで KD 木の高次元での非効率性を改善している． これらの手法は，探索のクエリに対して $O(D \log N)$ の時間計算量を要する (いずれも D がある程度小さい場合) が，木構造の構築にオーバーヘッドがあるので 前述のバッチ化により木の構築回数を減らすことで高速化の効果が相乗効果により大きく得られると考えられる．また，高次元で線形探索

以上の効率で厳密な最近傍探索を行うことは困難であることが知られている [19] が，GPU の特性を利用することで線形探索を高速に行う手法も考案されている [20].

近似的な最近傍探索の手法を可視化アルゴリズムに適用すると，D や N が大きい際にも高速に近傍探索が可能である．近似最近傍探索 (Approximate Nearest Neighbor, ANN) には情報検索や画像認識を始めとした広範な応用があるため，近年も盛んに研究されいる．Navigable Small World Graphs (NSW) [21] などのグラフ系の近傍探索手法では，探索データをノードとするグラフ構造を考え，探索データ間に k-近傍グラフを作成しておき，探索時には辺の張られたノード間の距離のみを計算することにより計算量を改善できる．SONG 法では，アルゴリズム中でコーディングベクトル間に関連度を辺の重みとした k-近傍グラフを構築するため，この情報を利用することで D, N が大きい際でも探索コストを大幅に抑えることが期待できる．

6　おわりに

本稿では，インクリメンタルな低次元埋め込み手法である SONG 法を用いた教師なし音環境認識システムを提案した．実装上の改善をすることで可視化フェーズにおいて初期実装と比較し 2 倍以上の高速化を実現した．さらに，改善した実装を用いて，実際の鳥の鳴き声データに関しても適用し，同様の高速化の効果があることを確認した．また，データ数を増やした場合の調査により，さらなる高速化の可能性について考察し，さらなる高速化の可能性とリアルタイムの音環境認識への応用可能性が示唆された．

謝辞

本研究は JSPS 科研費 No.20H00475, 19KK0260 の助成を受けた．実験には京都大学情報学研究科 先端数理科学専攻 応用解析学講座の計算機を利用した．

参考文献

[1] Jacob Benesty, Jingdong Chen, and Yiteng Huang. *Microphone array signal processing*, Vol. 1. Springer Science & Business Media, 2008.

[2] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, pp. 832–839. AAAI, 2000.

[3] Kazuhiro Nakadai and Hiroshi G Okuno. Robot audition and computational auditory scene analysis. *Advanced Intelligent Systems*, Vol. 2, No. 9, p. 2000050, 2020.

[4] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Latent space visualization, characterization, and generation of diverse vocal communication signals. *bioRxiv*, 2020.

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, Vol. 9, No. Nov, pp. 2579–2605, 2008.

[6] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

[7] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature communications*, Vol. 10, No. 1, pp. 1–14, 2019.

[8] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016.

[9] Damith Senanayake, Wei Wang, Shalin H. Naik, and Saman Halgamuge. Self Organizing Nebulous Growths for Robust and Incremental Data Visualization. *arXiv:1912.04896 [cs]*, June 2020. arXiv: 1912.04896.

[10] Kenta Shinzato and Ryosuke Kojima. An unsupervised auditory scene analysis system using incremental low-dimensional embedding. In *Proceedings of the 2021 IEEE/SICE International Symposium on System Integration(SII2021), (to appear)*, 2021.

[11] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International journal of computer vision*, Vol. 77, No. 1-3, pp. 125–141, 2008.

[12] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15, New York, NY, USA, 2015. Association for Computing Machinery.

[13] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[14] Julio G Arriaga, Martin L Cody, Edgar E Vallejo, and Charles E Taylor. Bird-db: A database for annotated bird song sequences. *Ecological Informatics*, Vol. 27, pp. 21–25, 2015.

[15] David M. Chan, Roshan Rao, Forrest Huang, and John F. Canny. t-SNE-CUDA: GPU-Accelerated t-SNE and its Applications to Modern Data. *arXiv:1807.11824 [cs, stat]*, July 2018. arXiv: 1807.11824.

[16] Corey J. Nolet, Victor Lafargue, Edward Raff, Thejaswi Nanditale, Tim Oates, John Zedlewski, and Joshua Patterson. Bringing UMAP Closer to the Speed of Light with GPU Acceleration. *arXiv:2008.00325 [cs, stat]*, August 2020. arXiv: 2008.00325.

[17] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, Vol. 18, No. 9, p. 509–517, September 1975.

[18] Stephen M Omohundro. *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.

[19] 和田俊和. 高次元空間における近似最近傍探索技術の進歩とその展望 (特集; 大規模画像データ処理). 人工知能, Vol. 25, No. 6, pp. 761–768, 2010.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus, 2017.

[21] Yu. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, 2018.

Visualizing Soundscape of Animal Vocalizations in Forests Using Robot Audition Techniques

Hao Zhao[1*] Reiji Suzuki[1] Shinji Sumitani[1] Shiho Matsubayashi[2] Takaya Arita[3]
Kazuhiro Nakadai[3, 4] Hiroshi G. Okuno[5, 6]

[1] Nagoya University [2] Osaka University
[3] Tokyo Institute of Technology [4] Honda Research Institute Japan
[5] Kyoto University [6] Waseda University

Abstract: A visualisation of the soundscape dynamics is one of the important topics in ecoacousics. In this paper, We try to use robot audition techniques and ecological methods to visualize the soundscape dynamics of forest animals for a long time. We create two false-color spectrograms based on acoustic indices and direction of arrival of sounds to show the overall dynamics of the soundscape of birds and cicadas in an about four-hour recording in a forest. The preliminary quantitative analysis of their vocal activities also implied that there might exist temporal avoidance behaviors among them.

1 Introduction

Visualization is one of the key techniques when considering roles of sounds in ecoacoustics: the subject to understand their own properties and functions in environments, and the tool for the indirect measurement of biodiversity or habitat quality of environments [1]. Extracting a spatio-temporal structure of a soundscape, which is a combination of sounds that arise from both natural and artificial environments, is essential for both roles in order to track active interactions among individuals and to grasp the overall properties of acoustic events.

We have been proposing and discussing novel applications of robot audition techniques to visualize soundscape dynamics in the directional or spatial domain by using the direction of arrival (DOA) of sound sources obtained from HARKBird, which is a bird song localization software based on a robot audition software HARK (explained later) [2, 3]. Inspired by Towsey et al. [4, 5], we created a false-color spectrogram that visualizes directional (DOA-based) soundscapes in which the color of the spectrogram reflects the direction of arrival of sounds, expecting that we can intuitively recognize directional variations of acoustic events (e.g., different vocalizing individuals or an individual vocalizing in different positions) [2]. We applied this to a 5 min recording with individuals of Zebra Finch, each put in a cage around the microphone array unit, and showed that the extracted visual information can reflect acoustic structures among this simulated group of individuals in the directional domain.

This paper further discusses an application of our framework to a soundscape analysis of a complex situation of vocalizing animals in forests. In particular, we focus on the acoustic dynamics of birds and arthropods, which are major species that dominate the soundscape in forests in early summer. It has been reported that birds are able to adjust both the timing and frequency of their signals to reduce overlap with the signals of other bird species [6, 7], other animals[8] and abiotic noise [9]. Hart et al. showed that birds significantly avoid temporal overlap with cicadas by reducing and often shutting down vocalizations at the onset of cicada signals that utilize the same frequency range [8].

We first illustrate the overall dynamics of the soundscape in about four-hour recording, by showing two false-color spectrograms based on acoustic indices and direction of arrival of sounds. Then, we further illustrate inter- and intra-specific interactions by classifying localized sound sources into bird and cicada vocalizations by making use of a typical acoustic index used in ecoacoustics. The preliminary analysis of their vocalization activities indicated that there might exist temporal overlap avoidance behaviors between birds and cicadas, and intra-specitic turn-taking between cicada individuals.

2 Materials and methods

2.1 HARKBird

HARK is an open-sourced robot audition software consisting of multiple modules for sound source localization, sound source separation, and automatic speech recognition of separated sounds that work on any robot with any microphone configuration [10]. See the website of HARK for detail[1].

HARKBird is a collection of Python scripts that enable us to conduct a field recording using microphone arrays connected to a laptop PC and analyze

*Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601
E-mail:zhao.hao@c.mbox.nagoya-u.ac.jp

[1] https://hark.jp

Figure 1: An experimental field (left) and a recording node (right).

the recording using a network of HARK which are designed to localize and separate bird songs in fields. The HARKBird can estimate the existence and the direction of arrival (DOA) of each sound source by using the MUltiple SIgnal Classification (MUSIC) method [11] based on multiple spectrograms with the short time Fourier transformation. We can further extract separated sounds as wave files for each localized sound using GHDSS (Geometric High order Decorrelation based Source Separation) method. The detailed description of HARKBird and the scripts are available from [12] and our website[2].

2.2 Recording and vocalization localization

We conducted an about 4-hour recording trial in the Inabu field, the experimental forest of Field Science Center, Graduate School of Bioagricultural Sciences, Nagoya University, in central Japan (Fig 1). The forest is mainly composed of conifer plantation (Japanese cedar, Japanese cypress, and red pine), with small patches of broadleaf trees (Quercus, Acer, Carpinus, etc.). In this forest, common bird species are known to vocalize actively during a breeding season.

The recording system is composed of the following components: a server node composed of a single PC; a microphone node (1 (right)) which has a microphone array (TAMAGO-03; System in frontier Inc.) connected with a Rasberry Pi 4; and a Wi-Fi router. The server and the node are connected together by the Wi-Fi, which enables us to control the node remotely. We placed the node in the field where there were some songbirds and cicadas (1 (left)). A recording started at 11:00am, June 27th, 2020 and ended at 3:20pm. In the end, we got thirteen 20-minute recordings with a total duration of four hours and 20 minutes.

We used the HARKBird to export the information about localized sound sources (i.e., the beginning and end time, DOA, and its separated sound file (wave file)). In this paper, we limited the frequency range for sound source localization to 2.5 - 3.5kHz, in order to localize vocalizations of birds and cicadas around this range. This is because some major species of songbirds (Blue-and-white Flycatcher (*Cyanoptila cyanomelana*), Red-billed Mesia (*Leiothrix lutea*), Eastern-crowned Warbler (*Phylloscopus coronatus*) and Japanese Bush Warbler (*Horornis diphone*)) and some cicadas (*Terpnosia nigricosta*) were singing around the microphone and sharing this frequency range. We adjusted the other parameters in HARKBird to localize these vocalizations as many as possible.

[2] `http://www.alife.cs.is.nagoya-u.ac.jp/~reiji/HARKBird/`

2.3 Soundscape visualization with false-color spectrograms

2.3.1 Acoustic index-based soundscape

Following Towsey et. al. [4], we create a false-color spectrogram based on three acoustic indices: acoustic complexity index (ACI) [13], temporal entropy in frequency bins (H[t]) and acoustic cover (CVR). Each original multi-channel recording (16 bit, 1.6 kHz) is mixed down to a single channel and its amplitude spectrogram (256 frequency bins for 8 kHz, 512 samples for each frame) is created using FFT, which is further divided into 10-second segments. The three types of the spectrum are calculated for each segment as follows:

H[t] spectrum: The temporal entropy of each frequency bin in the amplitude spectrogram. The amplitude values (overtime in a focal frequency bin) are normalized to the unit area and treated as a probability mass function. We calculate Shannon's entropy of this function, which is normalized by the maximum value. This index is useful for picking up infrequent vocalizations.

ACI spectrum: For each frequency bin, we calculate the average absolute fractional change in spectral amplitude from one spectrum to the next [13]. This index is proposed to estimate the abundance of bird vocalizations in a target soundscape.

CVR spectrum: For each frequency bin, we calculate the fraction of values where the spectral power exceeds the noise power (i.e., the average over the values in the frequency bins of 5-8 kHz).

We get a three spectrum matrix of the whole recording with three acoustic indices. Then, we create a false-color image by mapping the values of the three indices to the brightness of the RGB components of each pixel: red=ACI, green=1-H[t] and blue=CVR. The scaled values were assigned to each color in order to make the value differences clearer.

2.3.2 DOA-based soundscape

We create another false-color spectrogram that visualizes DOA-based soundscapes (Fig. 2), proposed in [2], according to the procedures as follows:

1. We generate a grayscale spectrogram of the whole original recording, where the (brighter) grayscale value of each pixel reflects the (higher) energy at the corresponding time and frequency.

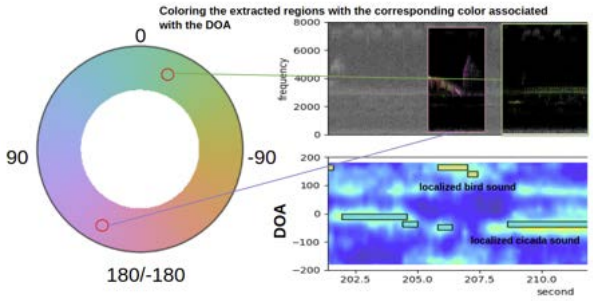


Figure 2: An overview of DOA-based spectrogram. (left) A circular color map, (top right) spectrogram, and (bottom right) MUSIC spectrum (the likelihood of sound existence in the space of time and DOA).

2. We generate a grayscale spectrogram of each separated sound, and extract pixels of which corresponding energy (dB normalized by the maximum value) is higher than 0.9 × the average value over the spectrogram.

3. We pick a color in a circular color map that corresponds to the DOA of each separated sound.

4. We assign the picked color in (3) to those of the pixels of the spectrogram in (1) that correspond to the extracted pixels in (2).

2.4 Classification of bird and cicada vocalizations and their interaction analyses

It is reported that CVR well responds to the continuous cicada chorus [4]. Our preliminary observations of the two spectrograms showed that CVR values around the frequency ranges of vocalizations on which we focus are significantly different between birds (low) and cicadas (high). We classified the localized sound sources into three classes (birds, cicadas, and noise) as follows:

1. We calculate the CVR values of 256 frequency bins of each separated sound file with HARK-Bird, and normalize these values so that their range is from 0 to 1.

2. We calculate the sum of the CVR values corresponding to the frequency range from 2.6 to 3.1 kHz, which is further divided by the sum of the entire values. We call this value the relative CVR (RCVR).

3. Each sound source is classified as a vocalization of cicada (or bird) if its RCVR is above (or below or equal to) the threshold value 0.2. The sound is regarded as noise if its RCVR is less than a small threshold (=0.0 in this case).

Note that used the normalized value in order to exclude the misclassification of cicadas due to other insect noises, and we adopted this threshold value because there were two peaks on both sides of the threshold in the frequency distribution of RCVR. While it is inevitable that this automatic but rough procedure can lead to misclassifications, we think that the results are enough to illustrate the basic tendency of their acoustic behaviors.

In order to investigate inter-specific interactions between birds and cicadas, we compared the temporal changes in vocal activities of birds and cicadas. Their activity in each 300-second time segment is calculated as the total duration of localized sounds in the segment, which is normalized by the maximum value overall segments.

3 Preliminary results

3.1 Soundscape analysis

Fig 3 shows (a) acoustic index-based and (b) DOA-based soundscapes. Each panel corresponds to a 13-minute recording. In (a), we can see regions colored with yellow (a mixture of red and blue) in the intermediate frequency range around 2.5-3 kHz. This means that ACI and 1-H[t] reflected similar sound events. These regions indicate bird vocalizations because high values of both indices reflect large temporal changes in the amplitude within short time periods. Actually, we see repetitions of short and high-frequency vocalizations around the corresponding time periods in (b). For example, we see some songs of Blue-and-white Flycatcher (purple), Red-billed leopard (orange), and Japanese Bush Warbler (green) in Fig. 4 (bottom, 14:16-14:19). The vocalizations were colored with similar colors among vocalizations of each species but they tended to be different between species. This implies that a single individual might be singing in a different direction for each species. However, their song colors tended to be biased strongly by simultaneously vocalizing songs of cicadas in other time periods, and thus the method needs further improvement.

We also see in (a) that there exist blue (CVR) narrow regions around 3 kHz. They reflect songs of cicadas as expected, and the corresponding clusters of songs were indicated with quite different colors in (b). This means that multiple individuals of cicada were expected to be singing in different directions alternately in this recording as illustrated in an example situation in Fig 4 (top, 13:49-13:52).

3.2 Vocal activity analysis

Fig. 5 shows the distribution of vocalizations in the space of time and direction of arrival, which were classified into bird and cicada vocalizations. The red and blue bars represent vocalizations of birds and cicadas, respectively. We found that multiple individuals of both birds and cicadas were vocalizing during the recording since their vocalizations were localized in various directions. We also see that there were time durations tended to be dominated by cicadas (e.g.

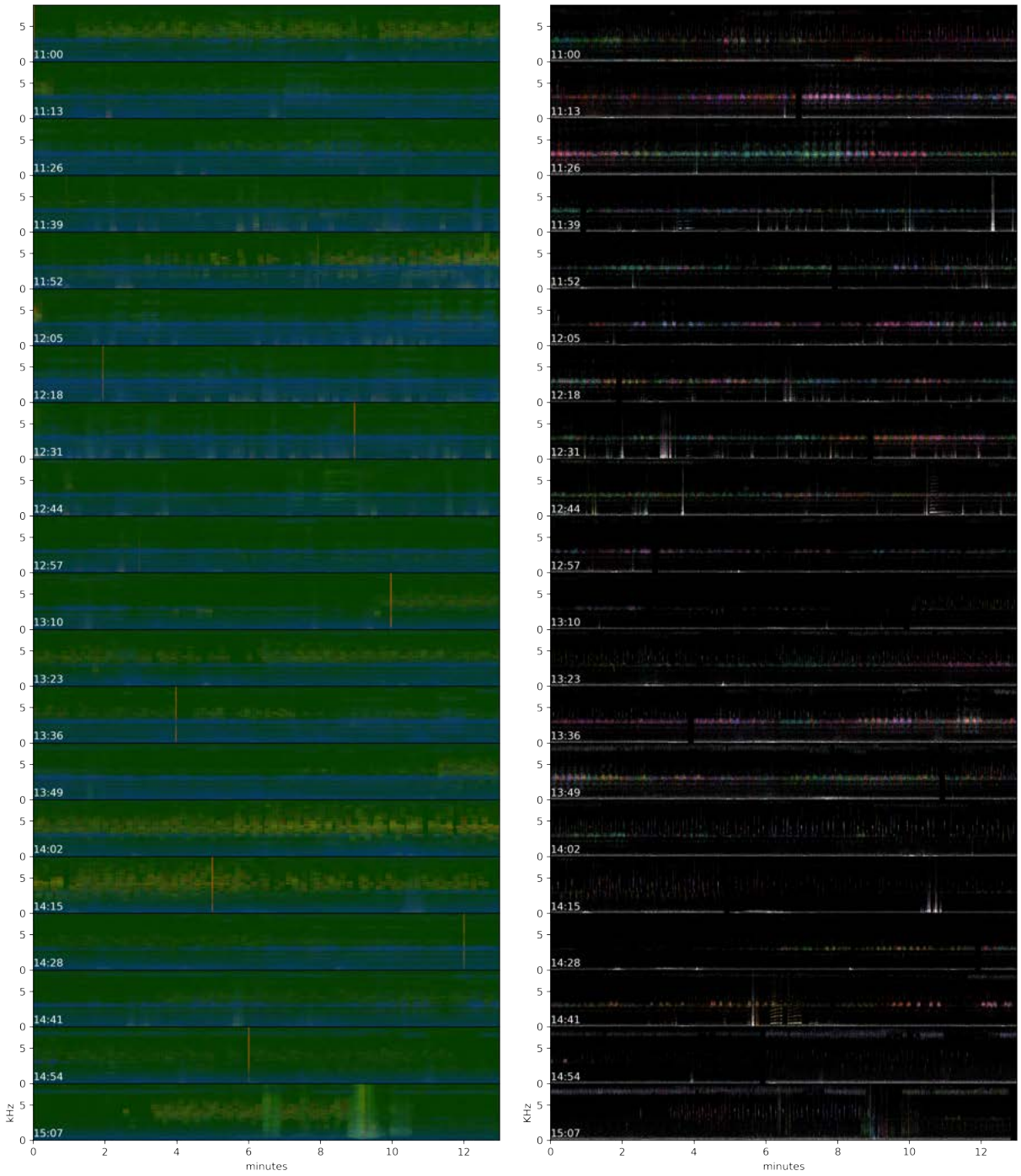


Figure 3: The acoustic index-based (left) and DOA-based (right) spectrograms for an about 4-hour recording.

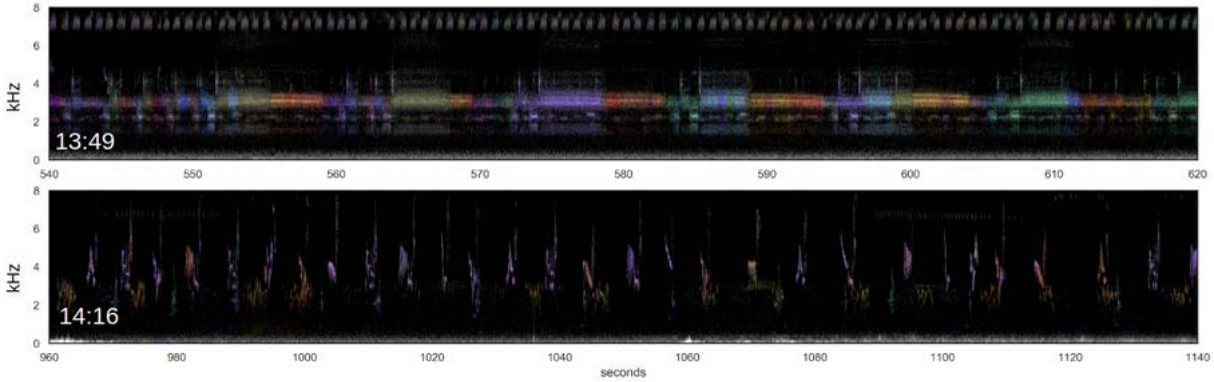


Figure 4: Examples of DOA-based spectrograms showing songs of birds (bottom) and cicadas (top).

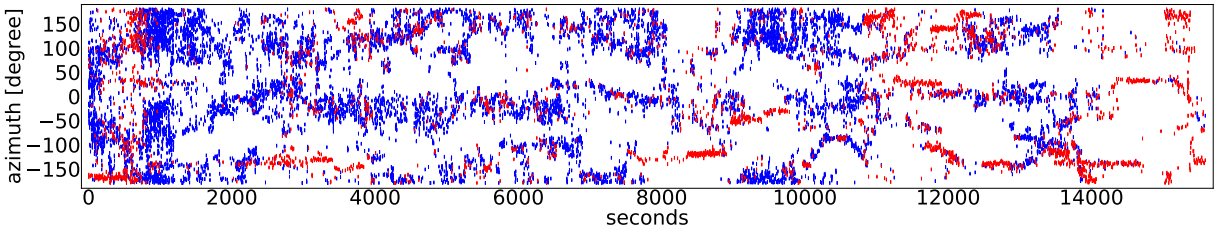


Figure 5: The directional distribution of localized vocalizations of birds (red) and cicadas (blue).

1000-2000 seconds) and one tended to be dominated by birds (e.g. 11000-12000 seconds).

Fig. 6 shows the changes in the vocal activity of birds and cicadas defined in Section 2.4. In the first half of the recording, the cicadas vocalized actively but the birds were relatively quiet, except for the first 900 seconds. On the other hand, the birds vocalized actively and the cicadas were gradually getting quiet in the latter half of the recording. In addition, it is suggested that there were vocal turn-taking between birds and cicadas at intervals of 5 to 15 minutes in that their activities repeated increased alternately. This could be an overlap avoidance of vocalizations between them because the frequency bands of vocalizations uttered by the birds and cicadas in this recording were relatively close. However, we need detailed analyses based on more sophisticated vocalization classification procedures.

We also observed intra-specific turn-takings between cicada individuals. Figure 7 shows an example of turn-taking situation. In this duration, multiple cicadas vocalized in some directions. The cicadas vocalized at -50 degrees and -100 degrees alternately in the first half. The cicada vocalized at -100 degrees and the positive directions (100 and 150) alternately. Both of them imply the occurrences of turn-takings among multiple individuals.

4 Conclusion

This paper discussed an application of robot audition techniques to a soundscape analysis of a complex situation of vocalizing birds and cicadas in early summer. We showed that two types of false-color spectrograms based on acoustic indices and direction of arrival can illustrate the overall dynamics of their acoustic behaviors. While the methods still need improvement, the preliminary quantitative analysis of their vocal activities implied that there might exist temporal avoidance behaviors among these birds and cicadas. We also found that there might also exist intra-specific turn-takings between cicada individuals.

Acknowledgements

We thank Naoki Takabe (Nagoya University) for supporting field recordings. This work was supported in part by JSPS/MEXT KAKENHI: 20H00475, 19KK0260, JP18K11467, and JP17H06383 in #4903 (Evolinguistics).

References

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

[2] R. Suzuki, Sumitani S. Zhao, H., S. Matsubayashi, K. Arita, T. Nakadai, and H. G. Okuno. Visualizing directional soundscapes of bird vocalizations using robot audition techniques. In *Proceedings of Proceedings of the 2020 IEEE/SICE International Symposium on System Integration (SII 2021)*, in press.

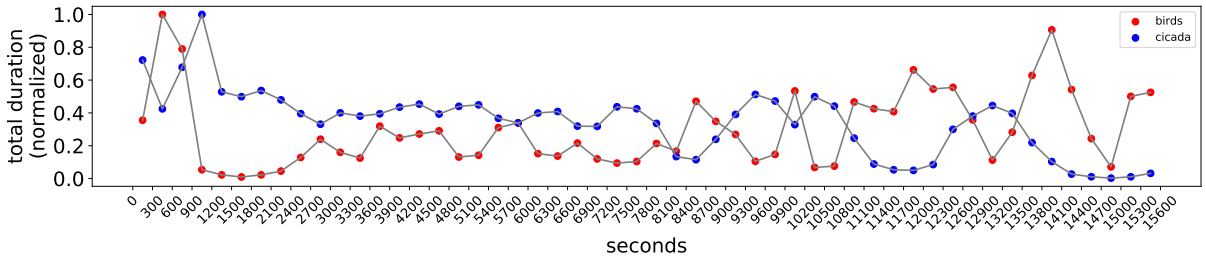


Figure 6: The changes in vocalization activities of birds (red) and cicadas (blue).

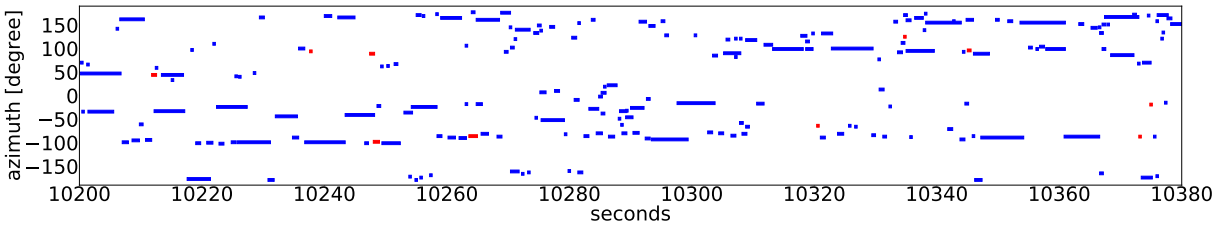


Figure 7: An example of turn-taking process of multiple cicada individuals.

[3] Shinji Sumitani, Reiji Suzuki, Shiho Matsubayashi, Takaya Arita, Kazuhiro Nakadai, and Hiroshi G. Okuno. Fine-scale observations of spatio-spectro-temporal dynamics of bird vocalizations using robot audition techniques. *Remote Sensing in Ecology and Conservation*, rse2.152, 2020.

[4] M. Towsey, L. Zhang, M. Cottman-Fields, J. Wimmer, J. Zhang, and P. Roe. Visualization of long-duration acoustic recordings of the environment. *Procedia Computer Science*, 29:703–712, 2014.

[5] M. Towsey, E. Znidersic, J. Broken-Brow, K. Indraswari, D. M. Watson, Y. Phillips, A. Truskinger, and P. Roe. Long-duration, false-colour spectrograms for detecting species in large audio data-sets. *Journal of Ecoacoustics*, 2:IUSWUI, 2018.

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

[7] H. Brumm. Signalling through acoustic windows: nightingales avoid interspecific competition by short-term adjustment of song timing. *Journal of Comparative Physiology A: Neuroethology*, 192:1279–1285, 2006.

[8] P. J. Hart, R. Hall, W. Ray, A. Beck, and J. Zook. Cicadas impact bird communication in a noisy tropical rainforest. *Behavioral Ecology*, 26:839–842, 2015.

[9] H. Slabbekoorn and M. Peet. Birds sing at a higher pitch in urban noise. *Nature*, 424, 2003.

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

[11] R. Schmidt. Bayesian nonparametrics for microphone array processing. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[12] 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*, pages 8246–8250, 2019.

[13] N. Pieretti, A. Farina, and D. Morri. A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). *Ecological Indicators*, 11:868–873, 2011.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-057-5 (11/20)

空間モデルを考慮した深層学習ベースの音源分離

DNN based speech source separation considering spatial model

戸上真人 [1*]

[1] LINE 株式会社
[1] LINE Corporation

Abstract: Recently, deep learning based speech source separation has been evolved rapidly. A neural network (NN) is usually learned independently of a spatial model. However, a research question remains whether the NN that is trained such as configuration is really optimal when speech source separation is performed with the spatial model. In this paper, I will introduce conventional statistical model based speech source separation and deep learning based speech source separation. After that, I will introduce four research directions which incorporate a spatial model into the NN structure.

1 はじめに

会議議事録の書き起こしシステムなどの音声認識システムを用いたシステム，及びテレビ会議システムなどの音声通話システムでは，マイクロホンに入ってくる音は様々な雑音・複数の人の話し声が混ざった音となる。このような様々な音が混ざったマイクロホン入力信号を音源毎に分離する音源分離技術に対する注目が集まっている。

これまで音源分離技術としてブラインド音源分離技術 (BSS) に関する検討が盛んになされてきた [2, 4, 7–10, 17, 24]。BSS では分離に必要なパラメータをマイク入力信号から求めるが，マイク入力信号以外に何も情報が無いとパラメータを一意に決定することが難しいため，追加の情報として音の伝搬過程および信号源に関する二つの手がかり・モデルを利用する手法が一般的に用いられている。音の伝搬過程に関するモデルは空間モデルと呼ばれ，マイク入力信号を音源と空間的なインパルス応答の線形畳み込み混合でモデル化する事が一般的である。一方，信号源に関するモデルとして音声の統計的性質（特に優ガウス性）に基づき，音源の原信号をラプラス分布 [7, 9] や時変ガウス分布 [4] でモデル化するような構成がよく用いられている。本稿ではこの信号源に関するモデルを統計モデルと呼ぶ。これら統計モデルを用いた音源分離技術はパラメータ最適化のための繰り返し計算に基づくアルゴリズムの検討 [16, 17, 25] と歩調を合わせて検討が進んでいる。

一方で，近年教師ありの音源分離手法として，深層学習に基づく音源分離方式に関する検討が広く進みつつある。例えば，ディープクラスタリング [6,23], パーミュテーション不変学習 [26, 27], ディープアトラクタネットワーク [1,11], BSS とのハイブリッド方式 [12,14,15] が提案されている。深層学習に基づく音源分離方式では，音源の周波数特性などの音源の特徴を従来の統計モデルに基づく音源モデルと比較し，より正確に捉えられる事が期待できる。加えて，音源分離のパラメータ最適化のために繰り返し計算に基づく方法を用いる必要が無いという利点も有する。こうしたことから深層学習に基づく音源分離方式の検討が飛躍的に進んでいる。特に，空間モデルを求めるための時間周波数マスクを深層学習により求める手法 [5, 6, 23, 26, 27] の検討が進んでいる。これらの手法では一般的に時間周波数マスクを学習するための教師データとして，時間周波数毎の S/N に基づく真の時間周波数マスクを定義し，その真の時間周波数マスクに近い時間周波数マスクをニューラルネットワークが出力するように学習を進める。これらの構成ではニューラルネットワーク学習時には，時間周波数マスクを用いて推定した空間モデルの精度を考慮することなく学習を行う。しかし，学習したニューラルネットワークを空間モデルを用いた音源分離に接続するとしたときには，時間周波数マスクを用いて推定した空間モデルの精度の影響を大きく受けるため，学習時においても時間周波数マスクを用いて推定した空間モデルの精度を考慮することが望ましいと考えられる。また時変ガウスモデルを用いた音源分離 [4] のように時々刻々フィルタの形状が変化する場合，時間周波数マスクと共に，時間周波数毎の音源の分散も求める必要がある。時間周波数マスクは空間モデルを推定するために用いる変数であり，一方で時間周波数毎の音源の分散は分離フィルタの形状を変化させるための変数であり，それぞれ役割が異なる。した

*連絡先：LINE 株式会社
E-mail: masahito.togami@linecorp.com

がって，空間モデルの影響を考慮し，それぞれの役割に適合した形で変数を推定することが望ましいと考えられる。

こうしたことから，著者らは，ニューラルネットワーク学習時において空間モデルの影響を考慮する方式として，次の４つの構成について検討を進めてきている。

1. 空間モデルの影響を考慮したニューラルネットワークの損失関数
2. ニューラルネットワークの構造の中に空間モデルを用いた音源分離を埋め込む方法
3. 所望音源の到来方向の情報をアトラクタとして用いて音源分離に必要なパラメータを推定するフレームワーク
4. 統計モデルに基づく音源分離法を疑似教師信号生成機として用いる教師無しニューラルネットワーク学習法

本稿では，これらの４つの方向性について紹介する。

2 空間モデルの影響を考慮したニューラルネットワークの損失関数 [18]

時間周波数領域でのマイク入力信号を $\boldsymbol{x}_{l,k}$ (l がフレームインデックス，k が周波数インデックス）とする。$\boldsymbol{x}_{l,k}$ は N_m 個の要素からなるベクトルとする。N_m はマイクロホン数である。時変ガウスモデルに基づく音源分離では複数チャネルのウィナーフィルタ (MWF)$\boldsymbol{W}_{i,l,k}$ (N_m 行 N_m 列) を使って，i 番目の音源信号 $\boldsymbol{c}_{i,l,k}$ を以下のように推定する。

$$\hat{\boldsymbol{c}}_{i,l,k} = \boldsymbol{W}_{i,l,k}\boldsymbol{x}_{l,k} \tag{1}$$

ここで MWF は

$$\boldsymbol{W}_{i,l,k} = v_{i,l,k}\boldsymbol{R}_{i,k}\left(\sum_{j=0}^{N_s-1} v_{j,l,k}\boldsymbol{R}_{j,k}\right)^{-1} \tag{2}$$

で求める事ができる。N_s は音源数，$v_{i,l,k}$ は i 番目の音源の時間周波数成分ごとの分散，$\boldsymbol{R}_{i,k}$ は空間共分散行列とする。空間共分散行列は時間周波数マスク $M_{i,l,k}$ を用いて以下のように推定される。

$$\boldsymbol{R}_{i,k} = \frac{1}{\sum_l M_{i,l,k}}\sum_l M_{i,l,k}\boldsymbol{x}_{l,k}\boldsymbol{x}_{l,k}^H \tag{3}$$

時間周波数マスク $M_{i,l,k}$ は音源の時間周波数成分毎の分散 $v_{i,l,k}$ と共にニューラルネットワークを介して推定される。これまで，一般的に時間周波数マスク $M_{i,l,k}$ を推定するためのニューラルネットワークの学習時には，時間周波数マスクの正解値を定義し，推定したマスクがその正解値に近づくようにニューラルネットワークを学習してきた。したがって，ニューラルネットワーク学習時には推定したマスクを用いて算出した空間モデルを通して音源分離した結果を評価してはいなかった。また，$M_{i,l,k}$ と $v_{i,l,k}$ は共に時間周波数毎の i 番目の音源の音量に関連する変数となるが，$M_{i,l,k}$ は空間共分散推定に用いられ，$v_{i,l,k}$ は時間毎のフィルタ形状を決めるために用いられるといったように，空間モデルを用いた音源分離における役割はそれぞれ異なる。したがって，$M_{i,l,k}$ と $v_{i,l,k}$ を推定するニューラルネットワークを i 番目の音源の音量を教師信号として学習することは必ずしも望ましくなく，空間モデルの影響を考慮した上で学習することが望ましいと考えられる。そこで我々は，$M_{i,l,k}$ と $v_{i,l,k}$ を推定するニューラルネットワークを音源分離信号 $\hat{\boldsymbol{c}}_{i,l,k}$ が真の音源信号 $\boldsymbol{c}_{i,l,k}$ に近づくように学習する手法を提案している [18]。提案法では，音源信号の事後確率を以下のように計算する。

$$p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k}) = \mathcal{N}(\boldsymbol{c}_{i,l,k}|\hat{\boldsymbol{c}}_{i,l,k}, \boldsymbol{V}_{i,l,k}) \tag{4}$$

ここで $\mathcal{M}$ はニューラルネットワークのパラメータとする。$\boldsymbol{V}_{i,l,k}$ は $\boldsymbol{c}_{i,l,k}$ の共分散行列であり $\hat{\boldsymbol{c}}_{i,l,k}$ と同様に $\boldsymbol{W}_{i,l,k}$，$v_{i,l,k}$ と $\boldsymbol{R}_{i,k}$ から算出することができる。$p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$ 算出のためのブロック構成を図 1 に示す。ニューラルネットワークの損失関数としては，負の対数事後確率 $-\log p_{\mathcal{M}}(\boldsymbol{c}_{i,l,k}|\boldsymbol{x}_{l,k})$ を用いる。提案する損失関数では，分散 $\boldsymbol{V}_{i,l,k}$ が一種の正則化効果を及ぼし $\hat{\boldsymbol{c}}_{i,l,k}$ が $\boldsymbol{c}_{i,l,k}$ から大きくずれている場合であっても，損失が過度に大きくなることを防ぐ効果があると期待できる。実際に二乗誤差と比較して提案する損失関数を用いて学習した結果，分離性能が向上することを確認している。また，残響除去と音源分離のためのニューラルネットワークを同時に学習する手法も提案されている [20]。

3 ニューラルネットワークの構造の中に空間モデルを用いた音源分離を埋め込む方法 [19]

空間モデルをニューラルネットワークに統合するための２つの構造として，ニューラルネットワークの構造の中に空間モデルを用いた音源分離を埋め込む方法を紹介する（図 2)。前章で紹介したようなニューラルネットワークで分離に必要なパラメータを推定しそのパラメータを使って音源分離を行うような構成の場合，図 2(a) で示すようにニューラルネットワークは順方向の計算中には空間モデルを参照しない。これに対して，空間モデルにより適合したパラメータをニューラルネッ

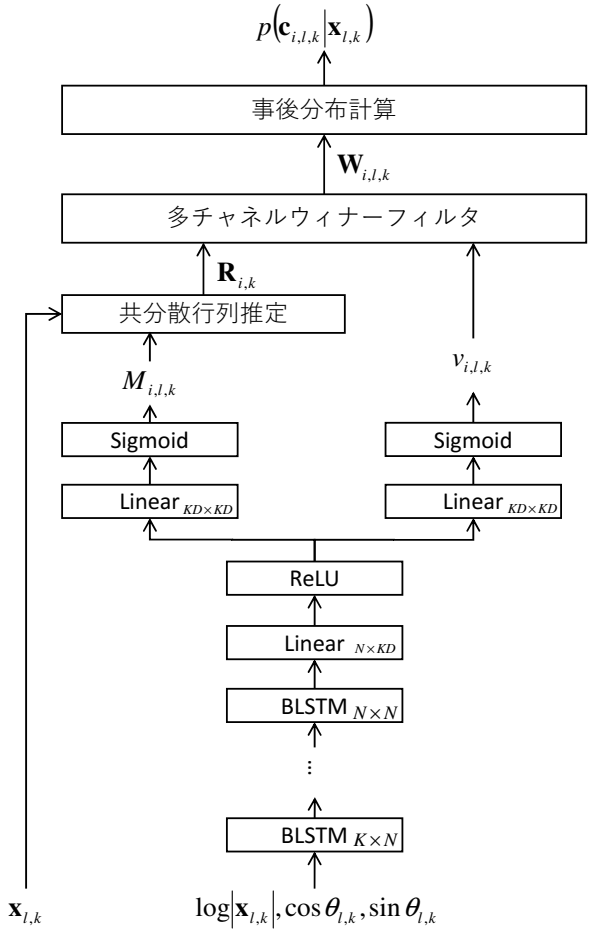


図 1: ニューラルネットワークを介した音源信号の事後確率算出ブロック

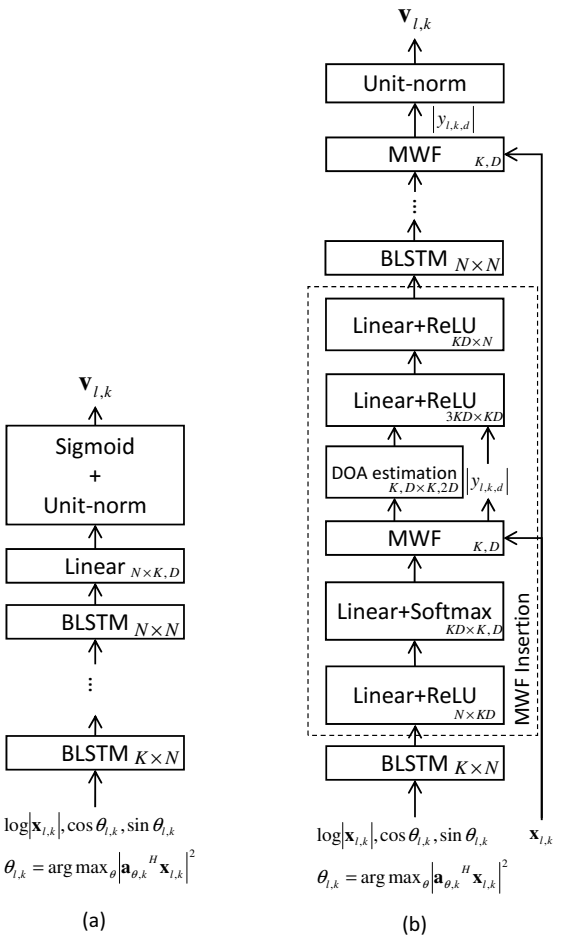


図 2: ブロック構成

トワークで推定可能となることを狙い，図 2(b) で示すように，順方向の計算時に空間モデルを参照する構成を提案する。本構成では，各 BLSTM の出力信号を時間周波数マスク $M_{i,l,k}$ に変換する。そして変換したマスクから共分散行列を構築し時不変の MWF に基づき音源分離実施する。その分離結果を次の BLSTM の入力信号として変換する。BLSTM の出力信号は時不変の MWF に変換され，BLSTM の出力信号の自由度よりも時不変の MWF の自由度が低いことから，音源分離を埋め込むことにより空間モデルに適合する形で自由度が落とすことが可能と考えられる。

ディープクラスタリング [6, 23] の構成に提案法の枠組みを適用し，音源分離性能が向上することを確認した。特に，空間モデルの適合度が向上することにより音源分離後の歪が減少することが確認された。

4 所望音源の到来方向の情報をアトラクタとして用いて音源分離に必要なパラメータを推定するフレームワーク [13]

所望音源の到来方向が分かっており，その方向の音のみを分離抽出したいというケースは多い。例えば，対話型ロボットでロボットの顔が向いている方向の音を取りたいというケースや，カメラ画像で人の顔を認識し人の方向の音だけを抽出したいというケースなどである。このような場合，音源分離した後に所望音源の到来方向から所望の分離音を選択するという構成よりも，音源分離の段階から到来方向の情報を利用した方が効率的に所望信号の情報と妨害音の情報を切り分けて推定できると推察される。そのような事から，所望音源の到来方向の情報を一種のアトラクタとして用いて，音源分離に必要なパラメータを推定するフレームワー

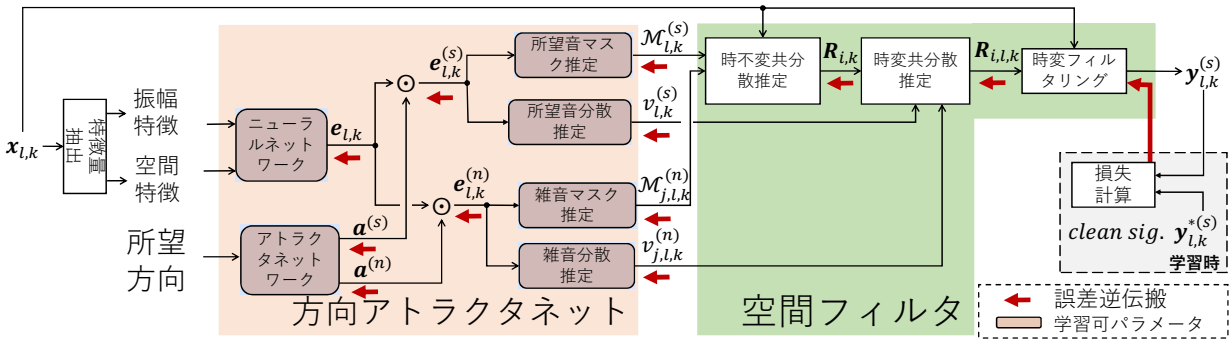


図 3: ブロック構成

クを提案した（図 3)。本構成は時変フィルタを用いた音源分離を実行するため，時間周波数マスクと共に，音源の時間周波数毎の分散を必要とする。これらのパラメータに対する教師信号を定義するのではなく，分離音に対する教師信号を定義することにより時間周波数マスクと音源の時間周波数毎の分散を推定するニューラルネットワークを統合的に学習することが可能となる。

5 統計モデルに基づく音源分離法を疑似教師信号生成機として用いる教師無し NN 学習法 [21]

深層学習を用いた音源分離で仮定する真の分離音は通常手に入らないことが多い。そのようなシーンでは，真の分離音がなくともニューラルネットワークを学習することが求められる。このようなニーズに対して，近年，教師無しのニューラルネットワーク学習法が提案されている [3,22]。これらの教師無しニューラルネットワーク学習法では，時間周波数マスクの教師信号が無くとも，各音源の時間周波数マスクを推定するためのディープクラスタリング構成のニューラルネットワークを学習することが可能な構成となっている。しかし，残響や背景雑音が存在する場合，各音源の時間周波数マスク以外にも様々な変数を推定することが必要になり，空間モデルを用いた音源分離の分離信号がより良くなるようにこれらの変数を推定することが望まれる。これに対して，真の分離音の代わりに，統計モデルに基づく音源分離法を疑似教師信号生成機として用いて，統計モデルに基づく音源分離法が出力する分離音に深層学習を用いた音源分離の分離音が近づくようにニューラルネットワークを学習する手法を提案する（図 4)。統計モデルに基づく音源分離法としては時変の MWF を用いる。共分散行列に各音源の直接音の共分散行列と共に残響と背景雑音の共分散行列も加えることにより，残響・背景雑音耐性を高める。統計モデルに基づく音源分離法の出力信号中に含まれる分離エラーに対して過度に追従することを防ぐために，Kullback Leibler Divergence (KLD) に基づく損失関数を用いる。

実験の結果，統計モデルに基づく音源分離法よりも高い分離性能を得ることが可能であることを確認している。

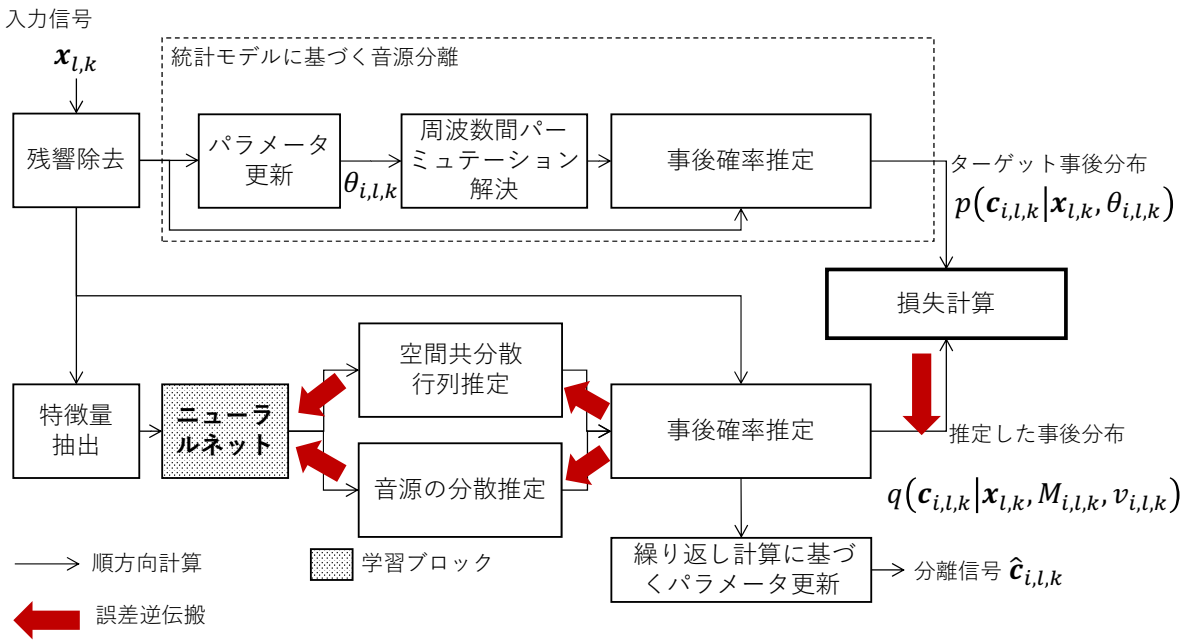


図 4: ブロック構成

6 まとめ

本稿では，空間モデルの影響を深層学習時にも考慮する 4 つの方向性を示した。深層学習に基づく方式の進展が著しいが，音源分離の空間モデルのような物理的な知識を活用可能なシステムにおいては，ニューラルネットワーク単体で全てを学習するのは望ましくなく，物理的な知識と深層学習に基づく方式をどう融合するかが今後の一つの重要な方向性と考えている。

参考文献

[1] Z. Chen, Y. Luo, and N. Mesgarani. Deep attractor network for single-microphone speaker separation. In *ICASSP 2017*, pp. 246–250, March 2017.

[2] P. Common. Independent component analysis, a new concept ? *Signal Processing*, Vol. 36, No. 3, pp. 287–314, April 1994.

[3] L. Drude, D. Hasenklever, and R. Haeb-Umbach. Unsupervised training of a deep clustering model for multichannel blind source separation. In *ICASSP 2019*, pp. 695–699, May 2019.

[4] N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Audio Speech Lang. Process.*, Vol. 18, No. 7, pp. 1830–1840, 2010.

[5] H. Erdogan, J.R. Hershey, S. Watanabe, M.I. Mandel, and J. Le Roux. Improved mvdr beamforming using single-channel mask prediction networks. In *Interspeech 2016*, pp. 1981–1985, 2016.

[6] J.R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP 2016*, pp. 31–35, 2016.

[7] A. Hiroe. Solution of permutation problem in frequency domain ica using multivariate probability density functions. In *Proceedings ICA*, pp. 601–608, Mar. 2006.

[8] N. Ito, S. Araki, and T. Nakatani. Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing. In *EUSIPCO 2016*, pp. 1153–1157, Aug 2016.

[9] T. Kim, H.T. Attias, S.-Y. Lee, and T.-W. Lee. Independent vector analysis: an extension of ica to multivariate components. In *Proceedings ICA*, pp. 165–172, Mar. 2006.

[10] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari. *Determined Blind Source separation with Independent Low-Rank Matrix Analysis*, chapter 6, pp. 125–155. Springer Publishing Company, Incorporated, 2018.

[11] Y. Luo, Z. Chen, and N. Mesgarani. Speaker-independent speech separation with deep attractor network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 4, pp. 787–796, April 2018.

[12] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari, and N. Ono. Independent deeply learned matrix analysis for multichannel audio source separation. In *EUSIPCO 2018*, pp. 1557–1561, Sep. 2018.

[13] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi. Deep speech extraction with time-varying spatial filtering guided by desired direction attractor. In *ICASSP 2020*, 2020.

[14] A.A. Nugraha, A. Liutkus, and E. Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.*, Vol. 24, No. 9, pp. 1652–1664, 2016.

[15] A.A. Nugraha, A. Liutkus, and E. Vincent. Deep neural network based multichannel audio source separation. In *Audio Source Separation*. Springer, March 2018.

[16] N. Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 189–192, Oct 2011.

[17] H. Sawada, H. Kameoka, S. Araki, and N. Ueda. Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Trans. Audio, Speech, and Language Process.*, Vol. 21, No. 5, pp. 971–982, May 2013.

[18] M. Togami. Multi-channel Itakura Saito distance minimization with deep neural network. In *ICASSP 2019*, pp. 536–540, May 2019.

[19] M. Togami. Spatial constraint on multi-channel deep clustering. In *ICASSP 2019*, pp. 531–535, May 2019.

[20] M. Togami. Joint training of deep neural networks for multi-channel dereverberation and speech source separation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3032–3036, 2020.

[21] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome. Unsupervised training for deep speech source separation with kullback-leibler divergence based probabilistic loss function. In *ICASSP 2020*, 2020.

[22] E. Tzinis, S. Venkataramani, and P. Smaragdis. Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information. In *ICASSP 2019*, pp. 81–85, May 2019.

[23] Z.Q. Wang, J. Le Roux, and J.R. Hershey. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *ICASSP 2018*, pp. 1–5, 2018.

[24] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, Vol. 52, No. 7, pp. 1830–1847, July 2004.

[25] K Yoshii, R Tomioka, D Mochihashi, and M Goto. Infinite positive semidefinite tensor factorization for source separation of mixture signals. *30th International Conference on Machine Learning, ICML 2013*, pp. 1613–1621, 01 2013.

[26] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva. Multi-microphone neural speech separation for far-field multi-talker speech recognition. In *ICASSP 2018*, pp. 5739–5743, April 2018.

[27] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP 2017*, pp. 241–245, March 2017.

バイナリマスク付き非負値行列因子分解に基づく発音時刻を用いた音源分離

Onset-informed Source Separation using Non-negative Matrix Factorization with Binary Masks

日下 湧太[1*]　糸山 克寿[1]　西田 健次[1]　中臺 一博[1, 2]
Yuta Kusaka[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 Insitute Japan Co.,Ltd.

Abstract: 本稿では，バイナリマスク付き非負値行列因子分解に基づき，目的音源の発音時刻を補足情報として利用する新しい音源分離手法について説明する．複数の楽器音により構成される混合音から特定の楽器音のみを分離するタスクにおいて，多くの既存手法は作成に時間と手間がかかる補足情報を利用していた．提案法では，楽曲を聴取しながらデバイスをタッピングするだけで容易に作成が可能な補足情報として，発音時刻を利用して音源分離を行う．NMF ベースの音源分離パイプラインに発音時刻を組み込むために，楽器音のオンとオフを制御するバイナリマスクを導入する．バイナリマスクは楽器音の連続性に関する仮定に基づき，マルコフ連鎖を用いてモデル化する．発音時刻はバイナリマスクがオンからオフに変化する時刻として扱う．提案モデルはギブスサンプリングによって推定され，推定時に発音時刻を活用することで効率的に目的音源を推定できる．提案法を用いて音楽音響信号からメロディを演奏する楽器音を分離する実験において，分離音と残留ノイズ比による評価で，2 から 10 dB の改善が確認できた．さらに，入力発音時刻の一部が欠落した場合や，時間方向のずれを含む場合の分離精度を評価し，提案法の頑健性を検証した．

1　はじめに

複数の楽器を含む音楽音響信号から目的の楽器音のみを分離する音源分離技術は，重要なトピックとして長年研究されている．分離によって得られる楽器音は，楽器練習や楽曲のリミキシングに有用であり，楽曲編集 [1]，カラオケ音源作成 [2]，自動採譜 [3] や楽器判別 [4, 5] といった音楽情報処理システムの改善にも活用できる．さらに，楽曲から分離したメロディラインの信号は，音楽検索システム [6, 7] のようなシステムにも利用可能である．

音源分離には非負値行列因子分解（non-negative matrix factorization; NMF）[8, 9] や独立成分分析（independent component analysis; ICA）[10] が提案されており，そのなかでも NMF はモノラル音響信号に対して有効な音源分離手法として長年研究されている．音楽音響信号に NMF を適用すると，信号に含まれる楽器音に対応する複数の基底に分解することができる．NMF によって混合音から目的の楽器音を分離するには，分解された基底から目的楽器に対応する基底の集合を選択する必要がある．しかし，NMF によって得られた基底と楽器音は基本的に一対一対応しないため，大量の基底から目的楽器の対応する基底を全て選択する操作は現実的には難しい．

NMF のような音源分離手法に分離したい音源に関する補足情報を入力することで，分離を補助したり分離精度を向上させたりするアプローチを informed source separation (ISS) [11] と呼ぶ．ISS で利用される情報の例として，目的楽器の音色のようなスペクトル的情報や，楽譜のような時間的情報などが挙げられる．ISS は楽器音や歌声分離に対して強力なアプローチであるが，補足情報の入手可能性の問題などにより適用不可能な場面も多い．ユーザが作成可能な補足情報を利用して

*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: kusaka@ra.sc.e.titech.ac.jp
本稿は DAFx2021 で採択された”ONSET-INFORMED SOURCE SEPARATION USING NON-NEGATIVE MATRIX FACTORIZATION WITH BINARY MASKS”を和訳したものである

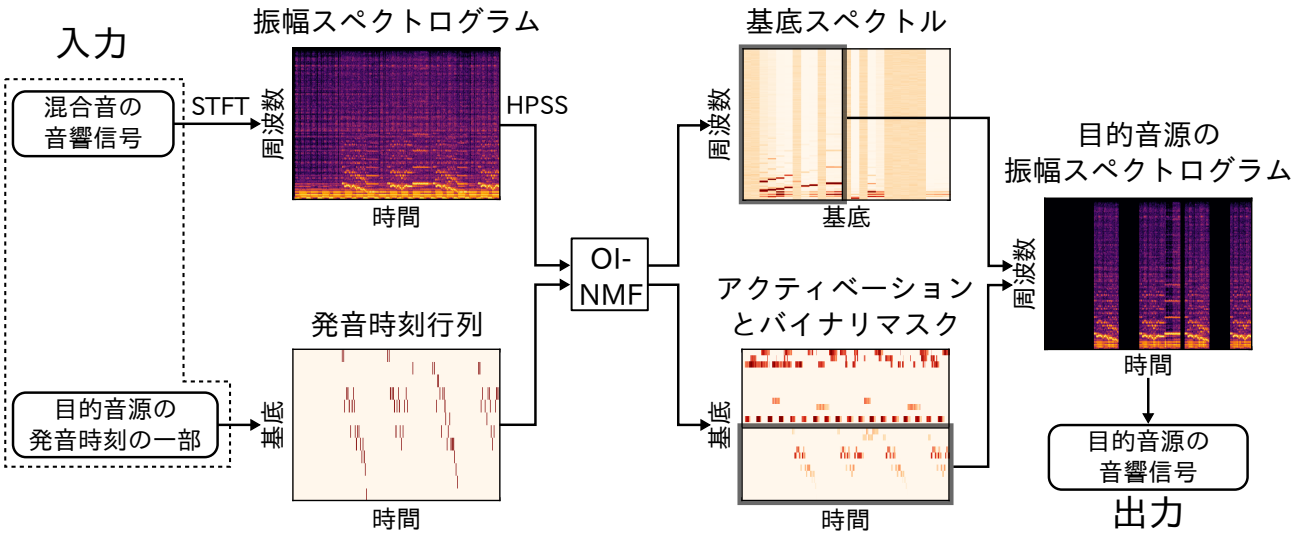


図 1: OI-NMF による目的音源分離の流れ．入力は混合音の音響信号と分離したい音源の発音時刻（の一部）である．入力は振幅スペクトログラムと発音時刻行列に変換され，OI-NMF によって基底スペクトル，アクティベーション，バイナリマスクに分解される．発音時刻を含めて推定を行った基底（図灰色の枠内の基底）が目的音源に対応しており，これらの基底を用いて信号を復元することで目的音源の音響信号を得ることができる．

音源分離を行うユーザガイドなアプローチ [12, 13] も提案されているが，情報の作成にはユーザの技量や手間が要求される．

そこで，本研究では分離したい楽器音の発音時刻を利用することでモノラルの音楽音響信号から目的の楽器音を分離可能な，新しい ISS 手法である *onset-informed NMF* (OI-NMF) を提案する．発音時刻は，ユーザが楽曲を聴取しながら分離したい楽器音の発音に合わせてキーボードやスマートフォンのようなデバイスをタッピングすることで容易に作成可能である．OI-NMF は既存の NMF ベースの音源分離モデルを，発音時刻を補足情報として入力できるよう拡張したモデルである．OI-NMF の特徴として，発音時刻は全てのタイミングで与えられる必要はなく，一部が欠落したものでも分離に利用できる点が挙げられる．これにより，既存の ISS 手法で利用されていた準備が難しい情報に比べ，簡単に作成できる情報に基づいて目的楽器音を分離できる．本研究の貢献を以下に示す．

- 既存の NMF モデルを拡張し，発音時刻を補足情報として入力できる OI-NMF モデルを開発した．NMF の変数として楽器音のオン/オフを表現するバイナリマスクを導入し，発音時刻をマスクがオンからオフに変化する時刻として扱う．OI-NMF のモデルを確率モデルとして定義し，バイナリマスクと発音時刻を含めてベイズ則によりモデルを推論することで目的の楽器音を効率的に推定することができる．
- OI-NMF を実装し，実楽曲から目的の楽器音を分離する実験とその分離精度を評価した．発音時刻は楽曲データセットに含まれる F0 アノテーションから作成したものを用いた．提案法と発音時刻を利用しないベースライン手法を比較し，OI-NMF と発音時刻の有効性を確認した．さらに，発音時刻の一部が欠落していたり，時間方向にずれを含んだりする場合の分離の頑健性を検証する実験を行った．

2 関連研究

OI-NNF は分離したい楽器の発音時刻を補足情報として利用するため，ISS の一種に含められる．基本的な ISS は利用する補足情報の種類によって次のように大別することができる．

- 目的音源のスペクトル的情報を利用するアプローチ．目的音源が楽器の場合，その音色や調波構造などが利用される．教師あり NMF [14, 15] は，分離したい楽器音を表す基底スペクトルを事前に用意した音源から学習して分離に利用する．用意した音源と目的楽器音が完全に一致しない場合に分離精度が劣化する問題点があるが，スペクトルに関する制約を加えることで精度劣化を抑えている．
- 目的音源の時間的情報を利用するアプローチ．OI-NMF で利用する発音時刻もこちらに該当する．音楽音響信号に関する典型的な時間情報に楽譜が

挙げられる．楽譜は楽器音の発音時刻，消音時刻という時間情報に加え，楽音の音高というスペクトル情報も持ち，これを利用する score-informed NMF [16, 17] は高精度な分離を実現している．また，近年盛んに研究されている深層学習を用いた手法 [18, 19] も，目的音源のみを含むスペクトログラムを教師としてモデルを学習するため，このアプローチに含めることができる．

これらの補足情報は分離に有効であるが，クリーンな目的音源の信号や楽譜は準備に手間がかかることやそもそも存在しないことがある．また，深層学習ベースの手法も適切な学習データを大量に用意する必要がある．そのため，これらの手法を実際に適用することは難しい場面も多く存在する．

この問題を解決するため，ユーザが楽曲を聴取して作成できるような補足情報を利用して分離するアプローチも提案されている．例えば，分離したい音源を真似た鼻歌 [12] や，スペクトログラム上の目的音源に対応する領域につけたアノテーション [13] などを分離に利用する．これらの情報は楽譜等の情報の準備が難しい楽曲に対しても適用可能である一方，その作成にはユーザの技能や時間を要求する．OI-NMF で利用する発音時刻は，これらの情報に比べて簡単に作成可能である．

また，発音時刻と類似した時間的情報を利用する手法として，非負値テンソル因子分解に基づきユーザが作成した楽器音の存在区間アノテーション [20] を利用する音源分離手法も提案されている．OI-NMF はこの手法と比較すると，モノラル音響信号にも適用可能である点や，存在区間で必要とされる消音時間を利用しないため情報作成が簡単という点で優れている．

3 非負値行列因子分解

非負値行列因子分解（non-negative matrix factorization; NMF） [8, 9] はモノラル音響信号の分離に有効なアルゴリズムである．もとは画像処理分野で提唱された [21] 手法であるが，音源分離 [22, 23] や自動採譜 [24] といった音声分野への応用も研究されている．音源分離における NMF は，入力である混合音の音響信号に短時間フーリエ変換（short-time Fourier transform; STFT）を適用して得られる振幅スペクトログラムを，音響信号の低ランク性に基づき 2 つの非負行列に分解する．

$$\boldsymbol{X} \sim \boldsymbol{W}\boldsymbol{H} \tag{1}$$

ここで，$\boldsymbol{X} \in \mathbb{R}_{\geq 0}^{F \times T}$ は振幅スペクトログラム（$\mathbb{R}_{\geq 0}$ は非負実数全体の集合），$f \in \{1, 2, \ldots, F\}$ は周波数ビン，$t \in \{1, 2, \ldots, T\}$ は時間フレームである．NMF の出力である $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{F \times K}$ は基底スペクトルと呼ばれ，振幅スペクトルに含まれる代表的なスペクトルパターンの基底 $k \in \{1, 2, \ldots, K\}$ から構成される行列である．もう一方の出力 $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{K \times T}$ はアクティベーションと呼ばれ，対応する基底の時間変化を表す行列である．K は基底数と呼ばれ，分解の基底の数を決めるハイパーパラメータである．

NMF によって得られた基底を用いて目的音源を分離するには，まず目的音源に対応する基底の組から選択して目的音源を表す基底スペクトル $\boldsymbol{W}_{\text{target}}$ とアクティベーション $\boldsymbol{H}_{\text{target}}$ を構成する．これらに対しウィナーフィルタを適用することで目的音源の振幅スペクトログラム $\boldsymbol{X}_{\text{target}}$ を得る．

$$\boldsymbol{X}_{\text{target}} = \frac{\boldsymbol{W}_{\text{target}}\boldsymbol{H}_{\text{target}}}{\boldsymbol{W}\boldsymbol{H}} \odot \boldsymbol{X} \tag{2}$$

ただし，$\odot$ は行列の要素ごとの積を表す．最後に，$\boldsymbol{X}_{\text{target}}$ と対応する位相スペクトログラムに逆短時間フーリエ変換（inverse STFT; ISTFT）を適用することで目的音源の音響信号を復元できる．ここで利用する位相スペクトログラムは，入力信号に STFT を適用して得られるもので十分であり，$\boldsymbol{X}_{\text{target}}$ から推定されたもの [25] を用いることで分離精度を向上させることも可能である．

4 Onset-informed NMF

本節では，OI-NMF とこれを用いた目的音源の分離について説明する．図 1 に OI-NMF による音源分離の流れを示す．入力は混合音の音響信号に STFT を適用して得られた振幅スペクトログラムと，分離したい音源の発音時刻の一部，出力は目的音源の振幅スペクトログラムである．発音時刻は発音時刻行列という時間周波数領域の行列に変換される．発音時刻行列を含めて OI-NMF モデルの推論を行うことで，分離したい音源が入力発音時刻から続くように推定される．最後に，発音時刻を入力した基底から $\boldsymbol{W}_{\text{target}}$ と $\boldsymbol{H}_{\text{target}}$ を構成し，NMF と同様にウィナーフィルタ (2) によって得られた振幅スペクトログラムに ISTFT を適用することで目的音源の音響信号を復元できる．

OI-NMF の一番の特徴は，NMF による音源分離において，発音時刻を補足情報として扱えるように導入したバイナリマスク $\boldsymbol{S} \in \{0, 1\}^{K \times T}$ にある．バイナリマスクはアクティベーションと同サイズの 2 値行列であり，アクティベーションと要素積をとる形で導入される．バイナリマスクを導入した NMF の拡張モデルとして beta process sparse NMF (BP-NMF) [26, 27] が提案されており，これに従って OI-NMF を次のように定義する．

$$\boldsymbol{X} \sim \boldsymbol{W}(\boldsymbol{H} \odot \boldsymbol{S}) \tag{3}$$

ここで，$\boldsymbol{X} \in \mathbb{Z}_{\geq 0}^{F \times T}$ は入力振幅スペクトログラム（$\mathbb{Z}_{\geq 0}$ は非負整数全体の集合），$\boldsymbol{W}, \boldsymbol{H}$ は通常の NMF(1) と同様の基底スペクトルとアクティベーションである．バイナリマスクは，既存の NMF におけるアクティベーションの値，つまり対応する楽器音の音量変化を，その 1/0 の値でオン/オフする機能を持つ．バイナリマスクが 1 のフレームはアクティベーションの値で楽器が発音しており，0 のフレームはアクティベーションの値によらず楽器の音量は 0 になる．このバイナリマスクにおいて，発音時刻はマスクが 0 から 1 に変化するフレームとして扱う．

OI-NMF のモデル (3) は，BP-NMF と同様に階層ベイズモデルとして定義され，モデルの変数を確率変数として次のように事前分布を導入する．

$$X_{f,t} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S} \sim \mathrm{Poisson}\left(X_{f,t} \,\middle|\, \sum_{k=1}^{K} W_{f,k} H_{k,t} S_{k,t} \right), \tag{4}$$

$$W_{f,k} \sim \mathrm{Gamma}\left(W_{f,k} \mid \alpha^W, \beta^W\right), \tag{5}$$

$$H_{k,t} \sim \mathrm{Gamma}\left(H_{k,t} \mid \alpha^H, \beta^H\right), \tag{6}$$

ここで，$\alpha^W, \beta^W, \alpha^H, \beta^H$ はガンマ分布のハイパーパラメータである．α^W, α^H はガンマ分布の形状パラメータであり，基底スペクトルに関する α^W は楽器音の調波構造におけるスパース性を誘導するため 1 より小さい値に設定する．一方，アクティベーションは 0 になるとバイナリマスクが機能しなくなるため，α^H を 1 より少し大きい値に設定することで，一定の大きさを持った値を誘導する．

4.1 OI-NMF の構造

提案法の新規性は，新しく導入したバイナリマスクと発音時刻の組み合わせにある．本節では，バイナリマスクと発音時刻のモデリングおよびこれらの変数を含めた OI-NMF モデルの推論方法について説明する．

4.1.1 バイナリマスク

バイナリマスクの事前分布をモデリングする際に，楽器音はその種類によって一定時間持続するという仮定を考える．つまり，現在発音している楽器音は次の時間フレームでも発音している確率が高く，発音していない楽器音は次のフレームも発音していない確率が高い．この仮定に基づき，バイナリマスクの事前分布をマルコフ連鎖によってモデル化する．バイナリマスク $\boldsymbol{S}$ のある基底 $\boldsymbol{S}_k = \boldsymbol{S}_{k,:}$ が従うマルコフ連鎖による事前分布は以下のように表される．

$$p(\boldsymbol{S}_k) = p(S_{k,1}) \prod_{t=2}^{T} p\left(S_{k,t} \mid S_{k,t-1}\right) \tag{7}$$

第 1 項 $p(S_{k,1})$ は最初の時間フレームの要素が従う確率分布であり，初期確率 $a_0 \in (0,1)$ をパラメータとするベルヌーイ分布によって定義される．

$$p(S_{k,1}) = \mathrm{Bernoulli}\left(S_{k,1} \mid a_0\right) \tag{8}$$

$p\left(S_{k,t} \mid S_{k,t-1}\right)$ はバイナリマスクの t が 2 以上のインデックスの要素が従う確率分布であり，ベルヌーイ分布の積によって定義される．

$$\begin{aligned} p\left(S_{k,t} \mid S_{k,t-1}\right) = \mathrm{Bernoulli}\left(S_{k,t} \mid a_{1\to 1}\right)^{S_{k,t-1}} \\ \cdot \mathrm{Bernoulli}\left(S_{k,t} \mid a_{0\to 1}\right)^{1-S_{k,t-1}} \end{aligned} \tag{9}$$

ここで，$a_{1\to 1}, a_{0\to 1} \in (0,1)$ はバイナリマスクがオン状態からオン状態，およびオフ状態からオン状態に遷移する確率である．これらの値は，楽器音の連続性の仮定の基づき，$a_{1\to 1}$ は 1 に近い値，$a_{0\to 1}$ は 0 に近い値に設定する．(7) より，バイナリマスク $\boldsymbol{S}$ 全体の事前分布は以下のように表すことができる．

$$p(\boldsymbol{S}) = \prod_{k=1}^{K} p(\boldsymbol{S}_k) = \prod_{k=1}^{K} p(S_{k,1}) \prod_{t=2}^{T} p\left(S_{k,t} \mid S_{k,t-1}\right) \tag{10}$$

4.1.2 発音時刻行列

分離したい楽器音は J 個（ただし $J < K$）の音高を持っており，音高 $j \in \{1, 2, \ldots, J\}$ に対して発音時刻の系列 $\boldsymbol{\tau}_j = (\tau_{j,1}, \ldots, \tau_{j,n}, \ldots, \tau_{j,N_j})$ が与えられると仮定する．ここで，N_j は音高 j に対して与えられる発音時刻の個数であり，発音時刻 $\tau_{j,n}$ は時間周波数領域の時間フレーム単位で表される．この発音時刻は，後述するモデル推論の際に扱いやすくするため，バイナリマスクと同サイズの発音時刻行列 $\boldsymbol{O} \in \{0,1\}^{K \times T}$ の形で次のように定義する．

$$O_{k,t} = \begin{cases} 1, & \tau_{k,n} \leq t \leq \tau_{k,n} + T_{\mathrm{onset}} \\ & \quad (k = 1, 2, \ldots, J, n = 1, 2, \ldots, N_j) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

ここで，T_{onset} は発音時刻の許容幅を表す．許容幅を設けることで，発音時刻が目的音源より前のタイミングに入力された場合でも分離を行うことができる．T_{onset} が大きすぎると目的音源以外の音源も参照される可能性があるため，T_{onset} を 1/16 拍, 1/8 拍, 1/4 拍 ... と変化させ，推論がうまく動作する下限である 1/8 拍に経験的に設定した．

4.2 提案モデルの推論

OI-NMF の出力変数である基底スペクトル $\boldsymbol{W}$，アクティベーション $\boldsymbol{H}$ およびバイナリマスク $\boldsymbol{S}$ を推定するためには，ベイズ則によりこれらの事後分布を推論すればよい．しかし，事後分布を解析的に計算することは困難なため，ギブスサンプリングによって期待値で近似的に求める．ギブスサンプリングでは，他の変数が与えられた条件付き分布に従ってサンプル列を生成し，サンプル列の平均を取ることで近似を行う．OI-NMF の i 番目のサンプリング式は次のように表される．

$$\boldsymbol{W}^{(i)} \sim p\left(\boldsymbol{W} \,\middle|\, \boldsymbol{H}^{(i)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}\right) \tag{12}$$

$$\boldsymbol{H}^{(i)} \sim p\left(\boldsymbol{H} \,\middle|\, \boldsymbol{W}^{(i+1)}, \boldsymbol{S}^{(i)}, \boldsymbol{X}\right) \tag{13}$$

$$\boldsymbol{S}^{(i)} \sim p\left(\boldsymbol{S} \,\middle|\, \boldsymbol{W}^{(i+1)}, \boldsymbol{H}^{(i+1)}, \boldsymbol{X}\right) \tag{14}$$

4.2.1 バイナリマスクのサンプリング

バイナリマスクの各要素は 0 か 1 の値をとるため，条件付き事後分布はベルヌーイ分布で表すことができる．

$$S_{k,t} \mid \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X} \sim \mathrm{Bernoulli}\left(S_{k,t} \,\middle|\, \frac{P_1}{P_1 + P_0}\right) \tag{15}$$

ここで，尤度 P_1, P_0 は $\boldsymbol{S}$ のインデックス k, t を除いた全ての要素 $S_{\neg k,t}$ を用いて次のように表される．

$$P_1 = p\left(S_{k,t} = 1 \mid S_{\neg k,t}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X}\right) \tag{16}$$

$$P_0 = p\left(S_{k,t} = 0 \mid S_{\neg k,t}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{X}\right) \tag{17}$$

尤度 P_1(16) は以下のように書き下される．

$$P_1 \propto p(S_{k,t} = 1) p\left(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, S_{k,t} = 1, S_{\neg k,t}\right) \tag{18}$$

(18) の第 1 項と第 2 項はそれぞれ次のように表すことができる．

$$p(S_{k,t} = 1) = \begin{cases} a_0, & t = 1 \\ a_{1\to1}^{S_{k,t-1}} a_{0\to1}^{1-S_{k,t-1}}, & t \geq 2 \end{cases} \tag{19}$$

$$p^1_{k,t} \triangleq p\left(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{H}, S_{k,t} = 1, S_{\neg k,t}\right) \tag{20}$$

$$\propto \prod_{f=1}^{F} (X^{\neg k}_{f,t} + W_{f,k} H_{k,t})^{X_{f,t}} \exp(-W_{f,k} H_{k,t}) \tag{21}$$

ここで，$X^{\neg k}_{f,t} = \sum_{l \neq k} W_{f,l} H_{l,t} S_{l,t}$ である．したがって，P_1 は次のように表すことができる．

$$P_1 = \begin{cases} a_0 p^1_{k,t}, & t = 1 \\ a_{1\to1}^{S_{k,t-1}} a_{0\to1}^{1-S_{k,t-1}} p^1_{k,t}, & t \geq 2 \end{cases} \tag{22}$$

同様に，P_0 は次のように表すことができる．

$$P_0 = \begin{cases} (1 - a_0) p^0_{k,t}, & t = 1 \\ (1 - a_{1\to1}^{S_{k,t-1}})(1 - a_{0\to1}^{1-S_{k,t-1}}) p^0_{k,t}, & t \geq 2 \end{cases} \tag{23}$$

ここで，$p^0_{k,t} \triangleq \prod_{f=1}^{F} (X^{\neg k}_{f,t})^{X_{f,t}}$ である．(22)，(23) および (15) に従って $t = 1$ から順にサンプリングすることで，バイナリマスク全体をサンプリングすることができる．

また，発音時刻が存在する部分は必ず楽器音はオン状態になっていると考え，発音時刻行列が $O_{k,t} = 1$ のインデックスの値をサンプリング結果にかかわらず $S_{k,t} = 1$ とする．それ以外のインデックスはサンプリング結果に従う．これにより，OI-NMF を推定する際に発音時刻を入力できる．

4.2.2 他の変数のサンプリング式

基底スペクトルとアクティベーションのサンプリング式は，BP-NMF のギブスサンプリング [27] と同様に以下のように導出できる．

$$W_{f,k} \mid \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{X} \sim \mathrm{Gamma}\left(\alpha^W + \sum_{t=1}^{T} X_{f,t} \phi_{f,t,k}\,,\ \beta^W + \sum_{t=1}^{T} H_{k,t} S_{k,t}\right) \tag{24}$$

$$H_{k,t} \mid \boldsymbol{W}, \boldsymbol{S}, \boldsymbol{X} \sim \mathrm{Gamma}\left(\alpha^H + \sum_{f=1}^{F} X_{f,t} \phi_{f,t,k},\ \beta^H + S_{k,t} \sum_{f=1}^{F} W_{f,k}\right) \tag{25}$$

4.2.3 OI-NMF のサンプリングアルゴリズム

アルゴリズム 1 に OI-NMF のギブスサンプリングアルゴリズムを示す．最初に各変数の初期化を行う．ギブスサンプリングで推論される確率分布は，初期値によらず定常分布に収束することが知られているが，どの音源がどの基底に推定されるかは初期値に大きく依存する．そのため，目的音源が発音時刻を与えた基底に出現するように誘導するため，確率変数は score-informed NMF [16, 17] を参考にして初期化する．

基底スペクトルは，ガンマ事前分布に従ってランダムに初期化する．アクティベーションは，発音時刻が与えられたフレームはガンマ分布で初期化する．また，発音時刻が与えられていない基底は全てのタイミング

Algorithm 1 OI-NMF のギブスサンプリング

1: $\boldsymbol{W}$, $\boldsymbol{H}$ および $\boldsymbol{S}$ を初期化
2: **for** $i = 1, 2, \cdots$ **do**
3: $\quad \phi_{f,t,k} = \frac{W_{f,k}H_{k,t}S_{k,t}}{\sum_l W_{f,l}H_{l,t}S_{l,t}}$ を計算
4: $\quad$ 式 (24) から $\boldsymbol{W}$ をサンプリング
5: $\quad$ 式 (25) から $\boldsymbol{H}$ をサンプリング
6: $\quad$ 式 (15), (22) および (23) から $\boldsymbol{S}$ をサンプリング
7: **end for**
8: サンプル列から $\boldsymbol{W}$, $\boldsymbol{H}$ および $\boldsymbol{S}$ の期待値を計算

で伴奏楽器が存在しうるため，同様にガンマ分布で初期化する．それ以外のフレームは 0 で初期化する．

$$H_{k,t} = \begin{cases} 0, & O_{k,t} \neq 1 \ (k = 1, 2, \ldots, L) \\ \frac{\alpha^H}{\beta^H}, & \text{otherwise} \end{cases} \tag{26}$$

バイナリマスクも同様のルールに従って初期化する．

$$S_{k,t} = \begin{cases} 0, & O_{k,t} \neq 1 \ (k = 1, 2, \ldots, L) \\ 1, & \text{otherwise} \end{cases} \tag{27}$$

その後，各変数のサンプリング式に従ってサンプル列を生成する．出力変数の値は，バーンイン後のサンプル列に対して平均をとったものとなる．ここで，バーンインとはサンプルが定常分布に達していないため破棄される期間を意味する．

4.3 目的音源の復元

ギブスサンプリングによって得られた出力変数を用いて，通常の NMF と同様に目的音源の音響信号を復元する．4.1.2 で述べたように，バイナリマスクに入力した発音時刻によって，基底 $k = 1, 2, \ldots, J$ に対応する音源が推定される．そのため，$\boldsymbol{W}_{\text{target}}$ と $\boldsymbol{H}_{\text{target}}$ は発音時刻を与えた基底を用いて次のように構成する．

$$\boldsymbol{W}_{\text{target}} = \boldsymbol{W}_{:,1:J}, \tag{28}$$

$$\boldsymbol{H}_{\text{target}} = \boldsymbol{H}_{1:J,:} \odot \boldsymbol{S}_{1:J,:}. \tag{29}$$

これに対してウィナーフィルタ (2) を適用し，得られた目的音源の振幅スペクトログラムに対して ISTFT を行うことで目的音源の音響信号を得ることができる．

5 評価実験

本節では，OI-NMF が発音時刻を利用して目的音源を分離できるか検証するために行ったメロディ分離実験について説明する．また，既存手法との比較による OI-NMF の有効性検証および頑健性評価についても説明する．

5.1 実験設定

入力楽曲には，音源分離実験用の実楽曲データセットである MedleyDB [28] から，ヴォーカルを含まずメロディのアノテーションが存在する楽曲として選択した，アーティストが MusicDelta であるジャズ楽曲 8 曲 (BebopJazz, CoolJazz, FreeJazz, FunkJazz, FusionJazz, LatinJazz, ModalJazz, SwingJazz) を利用した．これらの楽曲の wav ファイルから冒頭 20 秒を切り出し，22,050 [Hz] にダウンサンプリングした信号に対して，FFT サイズ 512 サンプル，オーバーラップ 50%，窓関数がハミング窓の STFT を適用することで得られた振幅スペクトログラムを OI-NMF の入力とした．なお，ドラムのような打楽器成分は，発音時刻が他の楽器と重複しやすいため，残存していると目的楽器音の分離が失敗しやすくなる．そのため，打楽器成分は予め調波・打楽器音分離 [29] によって除去した．

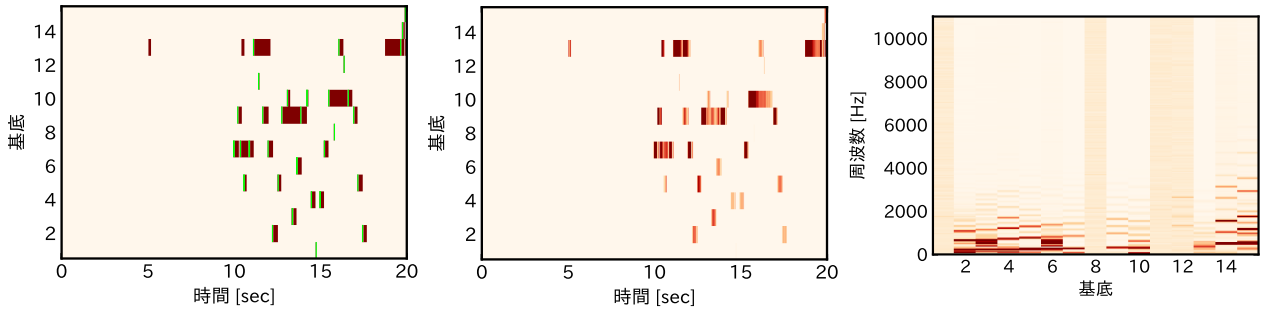
今回の実験では，メロディを演奏するアノテーションが付与された楽器を目的音源に設定し，これを分離する実験を行った．OI-NMF に入力するメロディ楽器の発音時刻は，MedleyDB データセットに含まれる F0 アノテーションから生成したものを用いた．F0 の値を MIDI ノート番号に変換し，ノート番号が変化する時刻を発音時刻とした．なお，ビブラートなどによる F0 の変化は発音時刻には含めない．

OI-NMF の基底数は十分大きい値として $K = 25$ とし，他のハイパーパラメータは $\alpha^W = 0.5$, $\beta^W = 1.0$, $\alpha^H = 1.1$, $\beta^H = 1.0$, $a_0 = 0.5$, $a_{1\to1} = 0.99$, $a_{0\to1} = 0.1$ に設定した．ギブスサンプリングは 200 回行い，得られたサンプルのうち開始から 100 サンプルはバーンインとして破棄して期待値を計算した．

5.2 分離精度評価指標

分離精度評価の指標には，signal-to-distortion ratio (SI-SDR), signal-to-interference ratio (SI-SIR), signal-to-artifacts ratio (SI-SAR) [30] を採用した．SI-SDR は推定された分離音と残差ノイズの比によって定義され，その値が大きいほど分離精度がよいことを表す．さらに，残差ノイズは目的音源以外の音源由来の干渉ノイズとアルゴリズム由来のノイズに分けられ，分離音とこれらのノイズの比によって SI-SIR と SI-SAR がそれぞれ定義される．SI-SIR と SI-SAR は互いにトレードオフの関係にあり，比較することでどちらのノイズ成分が支配的か調べることができる．

一般に，ブラインド音源分離の精度を評価するためには，上記の評価指標のスケール可変版である SDR, SIR および SAR [31] が用いられることが多い．しかし，OI-NMF のように分離音に無音区間が含まれると分離精度を正しく評価できなくなる．今回の実験設定



(a) $\boldsymbol{O}$ と $\boldsymbol{S}$　(b) $\boldsymbol{H} \odot \boldsymbol{S}$　(c) $\boldsymbol{W}$

図 2: OI-NMF による推定例（SwingJazz）．(a) 発音時刻行列（緑）とバイナリマスク（赤）．(b) アクティベーションとバイナリマスクの要素積．(c) 基底スペクトル．基底数 $K = 25$ のうち，発音時刻を与えた $J = 15$ 個の基底を表示している．

ではこれらの指標に代わり，スケール不変版を利用することでより正確な分離精度比較が可能となる．

5.3　分離例

まずはじめに，OI-NMF によるメロディ楽器分離の例を示す．ここでは，SwingJazz からクラリネットを分離した例を挙げる．発音時刻はすべてのタイミングで欠落なく与えられたものを利用した．

図 2 に推定された OI-NMF の各変数のヒートマップを示す．図 2(a) に入力発音時刻と推定されたバイナリマスクを示す．バイナリマスクが発音時刻から続いてピアノロールのように推定されている．基底 $k = 1, 8, 11, 12$ にはクラリネット以外の音源が推定されているが，発音時刻が存在しないフレームのマスクは 0 になっている．また，基底 $k = 13$ にはクラリネットとそれ以外の楽器音が同時に推定されてしまっている．図 2(b) に推定されたバイナリマスクとアクティベーションの要素積を示す．図 2(c) に推定された基底スペクトルを示す．クラリネットの調波構造が現れていることが確認できる．基底 $k = 1, 8, 11, 12$ にはクラリネットではない非調波成分が推定されているが，対応するアクティベーションと共に小さな値をとっているので信号復元時には打ち消される．この例の SI-SDR 改善率は約 4dB であった．ほかの楽曲の分離例と音源は次のリポジトリで確認できる[1]．

5.4　発音時刻の有効性検証

OI-NMF における発音時刻の貢献を示すため，OI-NMF と発音時刻を利用しない分離手法の精度を比較する実験を行った．比較対象には，発音時刻を入力しない OI-NMF と Bayesian NMF (BNMF) [32] を採用した．BNMF は通常の NMF を確率モデルとして推定を行う手法である．これらの発音時刻を利用しない手法により推定された基底は，目的楽器音とそれ以外の楽器音の基底で分類されていない．そのため，OI − NMF のように基底 $k = 1, 2, \ldots, J$ を利用して復元した信号は OI-NMF の分離精度の下限を与える．この下限と OI-NMF の分離精度を比較することで，発音時刻の有効性を確認することができる．

表 1: 発音時刻を利用する手法（発音時刻あり OI-NMF）と利用しない手法（発音時刻なし OI-NMF と B NMF）の SI-SDR 改善率 [dB] の平均．カッコ内の値は標準偏差を表す．

	OI-NMF		BNMF
	発音時刻あり	発音時刻なし	
Bebop	**5.62** (2.46)	−2.75 (3.32)	−4.42 (5.80)
Cool	**4.84** (2.75)	−0.56 (3.10)	−0.83 (5.38)
Free	**4.49** (1.63)	−3.37 (5.24)	−8.44 (14.7)
Funk	**9.86** (0.97)	−3.69 (3.99)	−4.84 (7.32)
Fusion	**7.09** (1.21)	−1.54 (3.33)	0.33 (3.22)
Latin	**5.77** (0.37)	−4.58 (11.9)	−6.67 (10.3)
Modal	**4.52** (1.92)	−5.60 (4.05)	−1.77 (5.52)
Swing	**4.29** (1.09)	−2.38 (2.36)	−6.08 (1.82)

各楽曲に対して 10 回分離を行ったときの，SI-SDR 改善率の平均と標準偏差を表 1 に示す．全ての楽曲において，発音時刻を入力した OI-NMF では平均が 0 以上であり，分離精度が改善していることが確認できる．一方，発音時刻を利用しない手法では平均は 0 未満であり，目的音源の分離ができていないことを示している．この結果より，OI-NMF は目的音源の分離に発音時刻を活用していることが確認できる．

[1] https://github.com/YutaKusaka/onset-informed-NMF-example

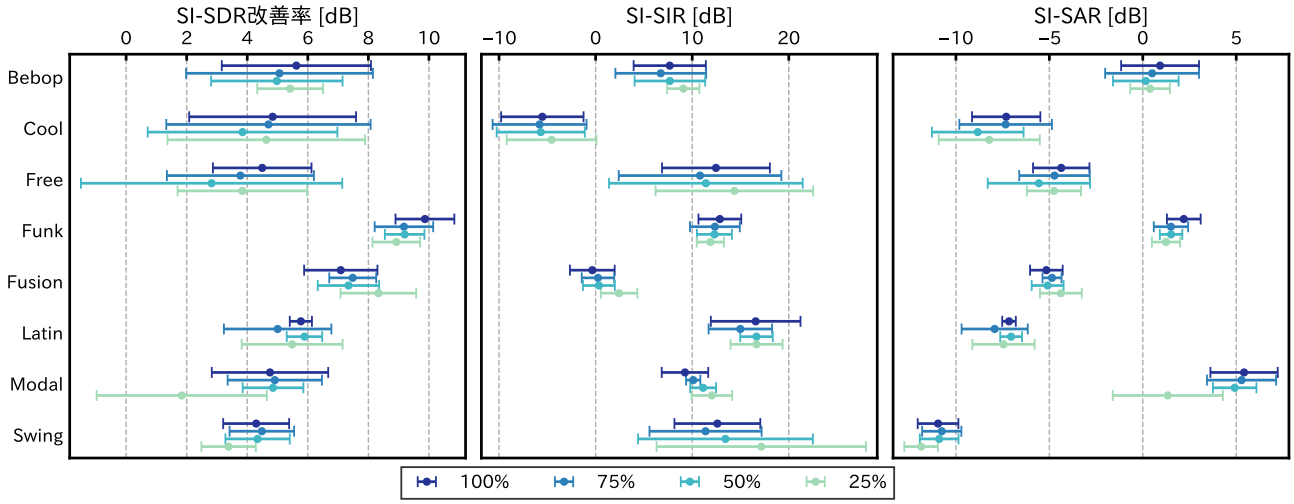


図 3: 発音時刻の存在割合を変化させたときの分離精度変化．ドットは 10 回分離を行った平均，エラーバーは標準偏差を表す．凡例は入力発音時刻の割合を示す．

5.5 発音時刻の欠落に対する頑健性評価

ユーザが楽曲を聴取しながら発音時刻を入力する際には，聞き逃しなどによる発音時刻の一部欠落が予測される．この一部が欠落した発音時刻が入力された場合でも OI-NMF による目的音源の分離は可能か検証する実験を行った．．この実験では，欠落がない発音時刻（100%）に対して存在割合が 75%, 50%, 25%になるよう各基底からランダムに欠落させた発音時刻を作成し，これらを入力した場合の分離精度を比較した．他の実験設定は 5.4 節と同様である．評価には SI-SDR 改善率に加え，どの種類のノイズが支配的か調べるために SI-SIR，SI-SAR も指標として利用した．

図 3 に評価結果を示す．全ての楽曲，発音時刻割合で SI-SDR の改善率の平均が 0 以上となり，発音時刻の割合が減少するにつれて，平均は減少し，標準偏差は増加する傾向にあることが確認できる．各試行で分解された基底を確認すると，発音時刻の割合が減少するにつれて発音時刻に対応する音の推定に失敗する基底が増加しており，このような傾向が現れていると考えられる．また，SI-SIR と SI-SAR に関しても SI-SDR と同様の変化傾向がみられ，これらを比較すると，OI-NMF においてはアルゴリズム由来のノイズのほうが支配的であることが確認できた．

また，発音時刻割合が 50%や 25%のように少ないとき，いくつかの試行で SI-SDR の改善率が 0 以下になり，分離に失敗していることが確認できた．さらなる考察を行うため，OI-NMF に入力される発音時刻が最悪の場合を想定し，各基底に対して 1 つだけ発音時刻を与えて分離する実験を各楽曲 10 回行った．その結果，多くの楽曲で分離に失敗している試行がみられ，FreeJazz では 4 回，LatinJazz では 6 回，SwingJazz では 8 回失敗していることが確認できた．これらの結果より，入力発音時刻が少なすぎる場合は分離に失敗すると予想される．一方で，50%以上の発音時刻が与えられる場合は分離精度の劣化は小さく抑えられており，発音時刻の一部の欠落を許容して分離ができると期待される．

5.6 発音時刻のずれに対する頑健性評価

ユーザが入力した発音時刻には，欠落だけでなく真の位置からのずれも含むことが予測される．この時間方向にずれを含む発音事項が入力された場合の，OI-NMF の分離の安定性を検証する実験を行った．発音時刻が含むずれは，実際にユーザに発音時刻する操作を行った研究で報告された統計値に基づき，平均が真の位置から 10 ms 後ろ [33], 標準偏差が 100 ms [34] の正規分布でモデル化した．つまり，ずれを含まない発音時刻を τ とすると，ずれを含む発音時刻 $\tilde{\tau}$ は次のように表すことができる．

$$\tilde{\tau} = \tau + \epsilon \tag{30}$$

$$\epsilon \sim \mathcal{N}(0.01, 0.1^2) \tag{31}$$

欠落のない 100%の発音時刻と，これに (30), (31) を適用して作成したずれを含む発音時刻の 2 種類を入力した際のそれぞれの分離精度を比較した．この実験も，5.4 節や 5.5 節の実験と同様のパラメータで，各楽曲に 10 回分離を行った．

図 4 に評価結果を示す．CoolJazz と FreeJazz を除く全ての楽曲で，ずれによって分離精度が劣化していることが確認できる．BebopJazz と ModalJazz の SI-SDR 改善率は，ずれがない場合に比べて大きく劣化しているが，他の楽曲では劣化幅が小さく抑えられてい

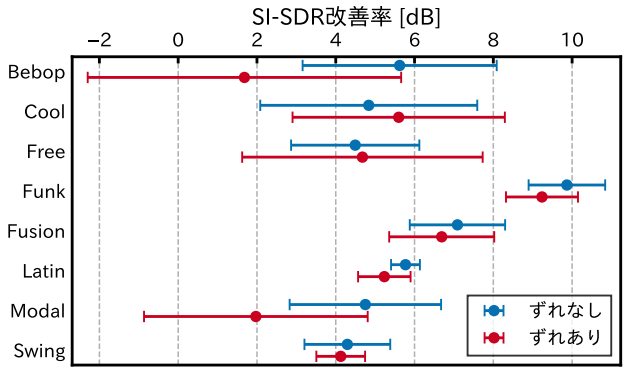


図 4: 発音時刻にずれを含む場合の分離精度の比較．ドットは 10 回分離を行った平均，エラーバーは標準偏差を表す．

る．BebopJazz と ModalJazz の大きな分離精度劣化の原因として，ノイズ性を持つ伴奏楽器が楽曲全体に存在しており，これを目的音源と誤って推定していることが考えられる．また，CoolJazz と FreeJazz においては，ずれを含まない場合よりも高い分離精度を示している．これは，データセットに含まれる F0 アノテーションから作成した発音時刻よりも，ずれを含んだ発音時刻のほうがより適しているためと考えられる．そのため，今後は入力発音時刻の作成方法による分離精度の変化なども考慮する必要がある．

以上の実験により，OI-NMF は発音時刻を利用して分離したい楽器音を効率的に推定できることが示された．さらに，人間が作成する際に想定されるレベルでの発音時刻の欠落やずれといった外乱をある程度許容した分離ができることを確認した．

6 おわりに

本稿では，分離したい音源の発音時刻を補足情報として利用する，新しい音源分離手法である onset-informed NMF を提案した．発音時刻を NMF ベースの音源分離フレームワークへ組み込むために，NMF のアクティベーションにマルコフ連鎖に基づくバイナリマスクを導入し，マスク上で発音時刻を扱った．さらに，バイナリマスクと発音時刻も含めてモデルを推論するアルゴリズムを導出した．分離精度を検証する実験で，発音時刻の一部が欠落したり，時間方向にずれを含むような現実的な設定においても，安定した分離を実現することが期待できる結果を示した．

現在の問題として，目的音源と伴奏音源が同時に発音しているような場合，NMF の性質上，OI-NMF ではこれらを分離することは難しいと考えられる．そのため，基底スペクトルに対して目的音源の調波構造に関する制約を取り入れるなどして，目的音源推定の精度を高めることを考えている．さらに，目的音源以外に発音時刻が与えられてしまった場合も，分離精度が劣化すると予想されるため，これに対する検証実験も行う予定である．

さらに，現在は入力発音時刻は楽器の音高ごとにグルーピングされて与えられる仮定をおいている．この仮定は，実際にユーザが発音時刻を入力する際には手間がかかる操作になると予想される．そのため，発音時刻を音高に依存しない単一の時系列としてモデルに入力できるように拡張することも考えている．

謝辞

本研究は JSPS 科研費 JP19K12017, JP19KK0260 および JP20H00475 の助成を受けた．

参考文献

[1] Kazuyoshi Yoshii et al. Drumix: An audio player with real-time drum-part rearrangement functions for active music listening. *Information and Media Technologies*, 2(2):601–611, 2007.

[2] A. J. Simpson et al. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *LVA/ICA*, pages 429–436, 2015.

[3] E. Benetos et al. Automatic music transcription: challenges and future directions. *J. Intell. Inf. Syst.*, 41(3):407–434, 2013.

[4] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *ICASSP*, volume 2, pages II753–II756, 2000.

[5] B. Kostek. Musical instrument classification and duet analysis employing music information retrieval techniques. *Proc. IEEE*, 92(4):712–729, 2004.

[6] M. Marolt. A mid-level representation for melody-based retrieval in audio collections. *IEEE Trans. Multimedia.*, 10(8):1617–1625, 2008.

[7] S. S. Shwartz et al. Robust temporal and spectral modeling for query by melody. In *SIGIR*, pages 331–338, 2002.

[8] D. D. Lee and H. S. Seung. Algorithms for Non-negative Matrix Factorization. In *NIPS*, pages 556–562, 2001.

[9] C. Févotte et al. Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis. *Neural Comput.*, 21(3):793–830, 2009.

[10] A. Hyvräinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Netw.*, 13(4):411–430, 2000.

[11] A. Liutkus et al. An overview of informed audio source separation. In *WIAMIS*, pages 1–4, 2013.

[12] P. Smaragdis and G. J. Mysore. Separation by “humming”: User-guided sound extraction from monophonic mixtures. In *WASPAA*, pages 69–72.

[13] A. Lefèvre et al. Semi-supervised NMF with time-frequency annotations for single-channel source separation. In *ISMIR*, pages 1–6, 2012.

[14] D. Kitamura et al. Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing. In *ISSPIT*, pages 392–397, 2013.

[15] D. Kitamura et al. Music signal separation by supervised nonnegative matrix factorization with basis deformation. In *DSP*, pages 1–6, 2013.

[16] S. Ewert and M. Muller. Using score-informed constraints for NMF-based source separation. In *ICASSP*, pages 129–132, 2012.

[17] J. Fritsch and M. D. Plumbley. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *ICASSP*, pages 888–891, 2013.

[18] S. Uhlich et al. Deep neural network based instrument extraction from music. In *ICASSP*, pages 2135–2139, 2015.

[19] P. Chandna et al. Monoaural audio source separation using deep convolutional neural networks. In *LVA/ICA*, volume 10169, pages 258–266, 2017.

[20] A. Ozerov et al. Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In *ICASSP*, pages 257–260, 2011.

[21] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[22] B. Wang and M. D. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *DMRN*, pages 1–5, 2005.

[23] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio Speech Lang. Process*, 15(3):1066–1074, 2007.

[24] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *WASPAA*, volume 3, pages 177–180, 2003.

[25] D. W. Griffin and J. S. Lim. Signal Estimation From Modified Short-Time Fourier Transform. In *ICASSP*, volume 2, pages 804–807, 1983.

[26] D. Liang et al. Beta Process Sparse Nonnegative Matrix Factorization for Music. In *ISMIR*, pages 375–380, 2013.

[27] D. Liang and M. D. Hoffman. Beta Process Non-negative Matrix Factorization with Stochastic Structured Mean-Field Variational Inference. In *arXiv:1411.1804*, pages 1–6, 2014.

[28] R. Bittner et al. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *ISMIR*, pages 155–160, 2014.

[29] D. FitzGerald. Harmonic/Percussive Separation Using Median Filtering. In *DAFx*, pages 1–4, 2010.

[30] J. L. Roux et al. SDR - half-baked or well done? In *ICASSP*, pages 626–630, 2019.

[31] E. Vincent et al. Performance Measurement in Blind Audio Source Separation. *IEEE Trans. Audio Speech Lang. Process*, 14(4):1462–1469, 2006.

[32] A. T. Cemgil. Bayesian Inference for Nonnegative Matrix Factorisation Models. *Comput. Intell. Neurosci.*, 2009:1–17, 2009.

[33] P. Leveau et al. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *ISMIR*, pages 1–4, 2004.

[34] M. Cartwright et al. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proc. ACM Hum. Comput. Interact.*, 1:1–21, 2017.

マイクロホン位置と音源スペクトルの確率モデルに基づくマイクロホンアレイのキャリブレーション

Calibration of a microphone array based on stochastic model of microphone position and sound source spectrum

段 雄啓[1]　糸山 克寿[1*]　西田 健次[1]　中臺 一博[1,2]
Katsuhiro Dan[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，マイクロホンアレイを構成するマイクロホンの位置ずれを校正（キャリブレーション）する手法について述べる．提案手法は，マイクロホンの位置，音源のスペクトル，録音された混合音のスペクトルに関する確率的生成モデルに基づいている．これにより，従来の手法では困難だった混合音を入力としたキャリブレーションを実現する．このモデルに基づき，観測した混合音に対する最大事後確率推定問題の解としてマイクロホン位置のキャリブレーションを実現する．3種類のマイクロホンアレイを用いたシミュレーション実験により，提案手法は混合音に対してマイクロホン位置のキャリブレーションを行えることが示された．

1　はじめに

近年，マイクロホンアレイはロボットなどの様々なデバイスに搭載されるようになっており，マイクロホンアレイの収録音を用いた音源定位や音源分離などの音響信号処理技術などが研究されている [1–4]．具体的な適用分野としては，スマートスピーカーやドローンを用いた災害救助などが挙げられる．前者は実際に存在する商品の一機能として既に実装されており，話者の方向をデバイスのランプで示す技術として用いられている．後者は災害時における要救助者発見のための技術として研究されており，暗所や瓦礫の中に要救助者がいる場合での運用デモンストレーションが既に行われている．上記のシステムのように，「現実世界の聴覚機能をロボットに対して実装すること」を目指しているロボット聴覚と呼ばれる分野が近年研究者の注目を集めている [5].

実際にマイクロホンアレイが実社会で活用される際，マイクロホンアレイ内のパラメタや外部環境を知ることは非常に重要な課題である．外部の環境は音源信号がマイクロホンアレイに収録するまでの過程に影響を及ぼし，内部のパラメタは音源定位などのアルゴリズムに既知として用いられるためその誤差が性能に大きな影響を及ぼすためである．主に音源定位などの研究分野では外部環境に対してロバストな推定を行うことができるようなアルゴリズムの開発が試みられている．一方で，マイクロホンアレイ内のパラメタ推定では主に以下の 2 つの分野が研究されている．

1. マイクロホン位置もしくは伝達関数の推定
2. 非同期信号の時刻ずれの推定

図 1 のように，マイクロホン位置や伝達関数が事前測定された値とずれてしまった場合，音源信号がマイクロホンアレイに収録される時間がずれてしまい，音響信号処理の性能に対して影響を与えることがある．信号が同期されていない場合も同様で，収録される時間のずれが以降の信号処理に対して悪影響をもたらす．同時に上記 2 点の同時解決を試みる手法などの提案も行われているが，音源信号の性質に関して制約を設けていることが多い．具体的には，単一音源であり，且つ，拍手音や TSP 信号などの立ち上がりの明確な音が主に用いられることが多い．立ち上がりの明確な音を用いることによって，音源信号がマイクロホンアレイに収録されるまでの時間が一意に定まるようになるためである．

*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 W8W-W310
E-mail: itoyama@ra.sc.e.titech.ac.jp
本稿は IEA/AIE2020 で採択された “Calibration of a Microphone Array Based on a Probabilistic Model of Microphone Positions” を和訳・一部修正したものである

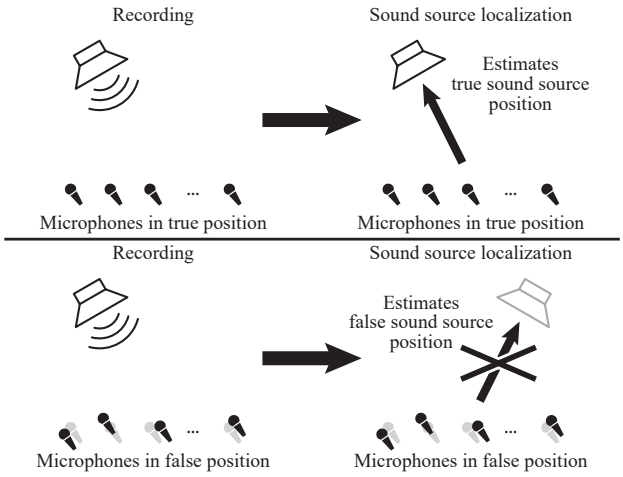


図 1: 真の位置にある各マイクロホンによって収録された音響信号を利用した場合，正しい音源定位結果が得られる．一方，誤った位置にあるマイクロホンによって収録された音響信号を利用した場合，音源定位結果は不正確なものとなる．

使いやすいキャリブレーションには，大きく分けて 2 つの条件が必要である．

1. 複数の音源が同時に収録された信号を用いたキャリブレーション：環境音などの音源は音の発生タイミングをコントロールできないため，同時収録された信号を利用できるキャリブレーションが必要である．
2. 任意の音源信号を使用したキャリブレーション：キャリブレーションの実施のために特定の音源を準備することなく実施できる手法が必要である．しかし，これまでのほとんどの研究においては，マイクロホンアレイ信号処理の前に面倒なキャリブレーション手順を必要とするキャリブレーション手法が提案されていた [6–9].

本論文では上記の 2 条件を実現したモデルを提案し評価することを目的とし，確率的生成モデルを用いた最大事後確率推定によるマイクロホン位置のキャリブレーション手法を提案する．従来の手法とは異なり，マイクロホン位置のずれや収録時に混ざるノイズ，未知の音源信号などの複数の乱雑さに対応するため，確率的なフレームワークからのアプローチを試みる．キャリブレーションには、環境音 (ホワイトノイズ) や音声などの複数の同時音源の収録信号を使用し，環境音などの立ち上がりが良くない音源信号を使用してもキャリブレーションを行うことができることを確認する．

2 関連研究

マイクロホンアレイの伝達関数の推定には主に二つの流れが存在する．同期マイクロホンアレイを対象とした手法と，非同期分散マイクロホンアレイを対象とした手法である．同期マイクロホンアレイの最初のテストは Thrun S [10] によって報告された．彼は音源信号の開始タイミングを使用したオンラインキャリブレーション手法を提案し，実際のマイクロホン装置を使用してその有用性を実証した．しかし，音源の位置が事前に決定されており，マイクロホンが完全に同期されているという制約が存在した．これらの制約を克服するために，三浦 *et al.* はロボットの自己位置とマップの同時推定を行う手法である SLAM(Simultaneous Localization And Mapping) に基づく非同期オンラインマイクロホン位置の推定法を提案した [7]．SLAM のロボット位置とマップ位置を，音源位置とマイクロホン位置に置き換えた手法である．拍手音を収録することにより，8ch マイクロホンアレイのマイクロホン位置を段階的にキャリブレーションすることに成功した．また，アドホックマイクロホンアレイについては，Raykar *et al.* と Ono *et al.* が別々に提案を行っている．彼らは，録音と再生が可能なデバイスを用意し，到来時間差 (TDOA: Time difference Of Arrival) [8,9] を使用して距離測定を行った．デバイス間の時間同期を使用し，機器間でお互いに発信した音を録音することによりマイクロホン位置のキャリブレーションを実現した．また，到来時間 (TOA: Time Of Arrival) でのキャリブレーションにおいて，近年提案された双線形アプローチなども応用されるようになった [11]．また，[8,12,13] で示されているように，MDS (Multidimentional Scaling) アルゴリズムを用いることで，マイクロホン位置を推定することをマイクロホン間の距離行列を推定することに帰着する手法も提案されている．[12] で提案されている手法は，Basis-point MDS と呼ばれる修正 MDS アルゴリズムを使用しており，推定すべき距離の数を減らすことができる．

上記のキャリブレーション方法では，いずれも音源信号に対して

- 音の鳴り始めの時刻が明確に分かる (すなわち TOA もしくは TDOA を波形から得ることができる)
- 音源信号のスペクトルがスパースである

という仮定が与えられていた．すなわち，鳴り始めの時刻が明示的には分からない音や複数の音源が同時に鳴動するような音などといった音響信号処理に用いる音でキャリブレーションを行うことができないという課題が存在している．

第 1 章で述べたように，使いやすいキャリブレーショ

ンには満たすべき 2 つの要件がある．本論文では，複数のランダムな要因を同じフレームワークで扱うために，確率を用いてマイクロホン位置を推定することを試みる．具体的には，マイクロホン位置と収録スペクトル，音源スペクトルに事前分布を仮定することによって確率的生成モデルを構築する．そして，構築した確率的生成モデルを用いて，最大事後確立推定によりマイクロホン位置のキャリブレーション手法の提案を行う．

3 提案手法

マイクロホンアレイを用いた録音には複数の乱雑さや未知の要因が存在する．提案手法では以下の 3 点を単一の尺度で扱うべく，確率的なフレームワークにおいて確率的生成モデルの構築を行った．

- 音源信号が未知
- マイクロホン位置の想定位置からのずれ
- 観測ノイズがランダムに発生

データの流れは以下のようになる．

1. 収録音を短時間フーリエ変換 (short-time Fourier transform, STFT) し，観測スペクトルを取得
2. 観測スペクトルとマイクロホン位置の基準位置を用いて推定音源スペクトルを導出
3. 観測スペクトルと推定音源スペクトルを用いて推定マイクロホン位置を導出
4. 観測スペクトルと推定マイクロホン位置を用いて推定音源スペクトルを導出
5. 3. に戻り，推定マイクロホン位置が収束するまで反復

3.1 問題設定

$\boldsymbol{x}_m \in \mathbb{R}^d$ を d 次元空間（$d = 2$ or 3）で M チャネルマイクロホンアレイを構成する m 番目のマイクロホン位置の座標とする．これらのマイクロホンは所与の基準位置 $\bar{\boldsymbol{x}}_m \in \mathbb{R}^d$ に従って配置されるが，実際の位置は基準位置からずれている．提案手法の目標は，マイクロホンアレイによって収録された音響信号を利用して，各マイクロホン位置 $\boldsymbol{x}_m$ を推定することである．提案手法では以下の 2 点を仮定する．

1. マイクロホン位置および音源信号の位置は時不変である．
2. 伝達関数は時不変でマイクロホン位置の関数として与えられる（例えば [3, 14]）．

N を音源数とし，$s_{nft} \in \mathbb{C}$ を n 番目の音源信号を STFT して得られる，f 番目の周波数ビン（$f = 1, \ldots, F$），t 番目の時間フレーム（$t = 1, \ldots, T$）における複素スペクトルとする．$z_{mft} \in \mathbb{C}$ をマイクロホンアレイを構成する m 番目のマイクロホンで録音された混合音の f 番目の周波数ビン，t 番目の時間フレームにおける複素スペクトルとする．以下で表される伝達関数

$$\boldsymbol{r}_{nf} = (r_{n1f}, \ldots, r_{nMf})^T \tag{1}$$

によって，n 番目の音源の音源スペクトルと m 番目のマイクロホンの観測スペクトルの関係は以下の式で表現される．

$$z_{mft} = \sum_n r_{nmf} s_{nft} + \epsilon_{mft} \tag{2}$$

ϵ_{mft} は m 番目のマイクロホン，f 番目の周波数ビン，t 番目の時間フレームにおける観測ノイズを表す．

3.2 確率的生成モデル

音源信号の生成や伝達過程に関する複数の乱雑さを扱うために，乱雑さが確率的に生成されると考え，混合音スペクトルが観測される過程をモデル化する．以下の 3 つの乱雑さを含む要因を考える．

- マイクロホンの位置 $\boldsymbol{x}_m \in \mathbb{R}^d$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$
- 音源スペクトル $s_{nft} \in \mathbb{C}$, $\boldsymbol{S} = (s_{nft})_{n=1,\ldots,N,\, f=1,\ldots,F,\, t=1,\ldots,T}$
- 観測ノイズを含む混合音スペクトル $z_{mft} \in \mathbb{C}$, $\boldsymbol{Z} = (z_{mft})_{m=1,\ldots,M,\, f=1,\ldots,F,\, t=1,\ldots,T}$

キャリブレーションの問題は同時事後確率 $p(\boldsymbol{X}, \boldsymbol{S}|\boldsymbol{Z})$[1] の MAP 推定として表現できる．提案手法の目的はマイクロホン位置のキャリブレーションであるため，音源スペクトル $\boldsymbol{S}$ の推定は必ずしも必要ではないが，MAP 推定によって副次的に得られる．すなわち提案手法では，キャリブレーションと同時に音源分離も行われることになる．

$p(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{S})$ を全てのランダム要素の同時確率分布とする．音源スペクトルとマイクロホン位置は独立な事象であるため，同時確率は以下のように分解することができる．

$$p(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{S}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{S}) p(\boldsymbol{X}) p(\boldsymbol{S}). \tag{3}$$

第 1 項 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{S})$ は，観測スペクトル $\boldsymbol{Z}$ の条件付確率を表す．各マイクロホンでの観測ノイズは時間，周波数，チャネルにおいて独立であるという仮定の下で，$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{S})$ を以下のように分解する．

$$p(\boldsymbol{Z}|\boldsymbol{S}, \boldsymbol{X}) = \prod_m \prod_f \prod_t p(z_{mft}|s_{1ft}, \ldots, s_{Nft}, \boldsymbol{X}) \tag{4}$$

[1] $\boldsymbol{S}$ を周辺化し，$p(\boldsymbol{X}|\boldsymbol{Z})$ の MAP 推定を行うことがより望ましい．

観測ノイズ ϵ_{mft} は平均が 0 で分散が σ_z^2 の円対象複素ガウス分布に従うとする．式 (2) より，観測スペクトル z_{mft} は平均 $\sum_n r_{nmf} s_{nft}$，分散 σ_z^2 の複素ガウス分布に従う．すなわち，以下のように表される．

$$p(z_{mft}|s_{1ft},\ldots,s_{Nft},\boldsymbol{X}) \propto \exp\left(-\frac{|z_{mft}-\sum_n r_{nmf}s_{nft}|^2}{\sigma_z^2}\right) \quad (5)$$

$(\cdot)^\star$ は複素共役を表す．

第 2 項 $p(\boldsymbol{X})$ はマイクロホン位置の事前分布を表す．マイクロホンは所与の基準位置に配置されるが，実際の位置には製造誤差や取り付ける際の誤差などの不確実性による基準位置からのずれが含まれる．このずれは各マイクロホンに独立であるとすると，$p(\boldsymbol{X})$ は以下のように分解できる．

$$p(\boldsymbol{X}) = \prod_m p(\boldsymbol{x}_m) \quad (6)$$

各マイクロホンの位置のずれは等方向的な正規分布 (分散が σ_x^2) に従うとすると，各マイクロホン位置の事前分布は以下で表現される．

$$p(\boldsymbol{x}_m) \propto \exp\left(-\frac{\|\boldsymbol{x}_m-\bar{\boldsymbol{x}}_m\|_2^2}{2\sigma_x^2}\right) \quad (7)$$

第 3 項 $p(\boldsymbol{S})$ は音源スペクトルの分布を表す．音源スペクトルは音源ごとに独立しているとする．単一音源のスペクトルの事前分布の表現に関しては，様々な表現が提案されてきた．例えば，非負値行列因子分解を用いた低ランク表現 [15,16] や，深層学習 [17,18] を使用した非線形表現が提案されている．ここでは，時間周波数平面での音源スペクトルの独立性を仮定し，$p(\boldsymbol{S})$ を以下のように分解する．

$$p(\boldsymbol{S}) = \prod_n \prod_f \prod_t p(s_{nft}) \quad (8)$$

各時間周波数スロットにおける音源スペクトル s_{nft} は等方的な複素ガウス分布に従うとすると，その分布は以下のように表現される．

$$p(s_{nft}) \propto \exp\left(-\frac{|s_{nft}|^2}{\sigma_s^2}\right) \quad (9)$$

3.3 キャリブレーションアルゴリズム

前節で述べた確率的生成モデルに基づく，マイクロホンアレイのキャリブレーションアルゴリズムを構築する．与えられた観測スペクトル $\boldsymbol{Z}$ の最適なマイクロホン位置 $\boldsymbol{X}$ は対数事後確率の最大化により得られる．

$$\hat{\boldsymbol{X}},\hat{\boldsymbol{S}} = \arg\max_{\boldsymbol{X},\boldsymbol{S}} p(\boldsymbol{X},\boldsymbol{S}|\boldsymbol{Z}) = \arg\max_{\boldsymbol{X},\boldsymbol{S}} \log p(\boldsymbol{Z},\boldsymbol{X},\boldsymbol{S}) \quad (10)$$

マイクロホン位置 $\boldsymbol{X}$ と音源スペクトル $\boldsymbol{S}$ の事後確率は独立ではないため，事後確率を最大化する $\boldsymbol{X}$ と $\boldsymbol{S}$ を同時に求めるのは困難である．本稿では $\boldsymbol{X}$ と $\boldsymbol{S}$ に関して反復的に事後確率を最大化することで，式 (10) を近似的に実現する．

マイクロホン位置 $\boldsymbol{X}$ に関する事後確率最大化はグリッドサーチによって実現する．未知の関数によって $\boldsymbol{X}$ から伝達関数 r_{nmf} への変換はなされると想定しているため，対数事後確率の $\boldsymbol{X}$ に関する導関数も得られない．グリッドサーチの範囲とグリッドの間隔は事前に定義する．

一方, 音源スペクトル $\boldsymbol{S}$ に関する事後確率最大化は解析的に導出可能である．対数事後確率は $\boldsymbol{S}$ に関して上に凸であるため, 偏導関数の零点を解くことで最適な音源スペクトル $\hat{s}_{nft}$ は得られる．観測スペクトルが $\boldsymbol{z}_{ft} = (z_{1ft},\ldots,z_{Mft})$，伝達関数が $\boldsymbol{r}_f = (\boldsymbol{r}_{1f},\ldots,\boldsymbol{r}_{Nf})$ と表されるとき，推定音源スペクトルは以下のように表される（$\boldsymbol{I}$ は $N\times N$ の単位行列）．

$$\hat{s}_{ft} = \left\{\left(\boldsymbol{r}_f^T\boldsymbol{r}_f^\star\right)^T + \frac{\sigma_z^2}{\sigma_s^2}\boldsymbol{I}\right\}^{-1}\left(\boldsymbol{z}_{ft}^T\boldsymbol{r}_f^\star\right)^T. \quad (11)$$

構築したアルゴリズムを Algorithm 1 に示す．

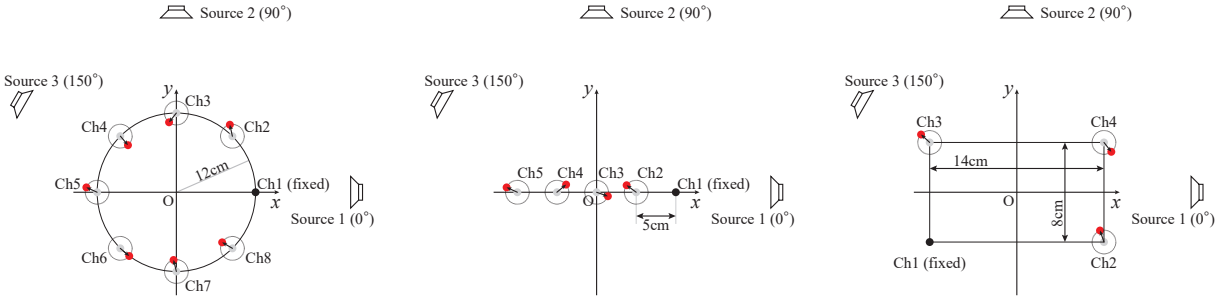
Algorithm 1 Iterative Estimation of $\boldsymbol{X}$ and $\boldsymbol{S}$

Initialize $\boldsymbol{X}^{(0)}$ and set $t \leftarrow 0$
repeat
  $\boldsymbol{S}^{(t+1)} \leftarrow \arg\max_S \log p(\boldsymbol{X}^{(t)},\boldsymbol{S}|\boldsymbol{Z})$ using Eq. (11)
  $\boldsymbol{X}^{(t+1)} \leftarrow \arg\max_X \log p(\boldsymbol{X},\boldsymbol{S}^{(t+1)}|\boldsymbol{Z})$ using grid search
  $t \leftarrow t+1$
until convergence

4 評価実験

提案手法を用いてマイクロホンアレイを構成するマイクロホンの位置のずれに対するキャリブレーションを行いその性能を評価した．性能評価の尺度には，キャリブレーションで推定された位置と真の位置の誤差を用いる．

実験はシミュレーション環境で行った．収録音のサンプリング周波数は 24 kHz，STFT の窓幅は 1024 点，シフト幅は 256 点とした．また，$\sigma_z^2 = 5\times10^{-10}$，$\sigma_s^2 = 5\times10^{-8}$ とした．σ_x^2 は与えたズレの大きさの二乗とした．グリッドサーチにおけるグリッドの大きさは 0.1 cm，グリッドサーチを行う範囲は与えた変位の大きさに比例して変化させた．音源信号については，ホ



(a) 8 チャネル円形アレイ (C8)　(b) 5 チャネル線型アレイ (L5)　(c) 4 チャネル長方形アレイ (R4)

図 2: シミュレーション実験に用いたマイクロホンアレイの形状と音源の配置

ワイトノイズ，音声データを用いた．音声は JVS corpus [19] からランダムに選択した．

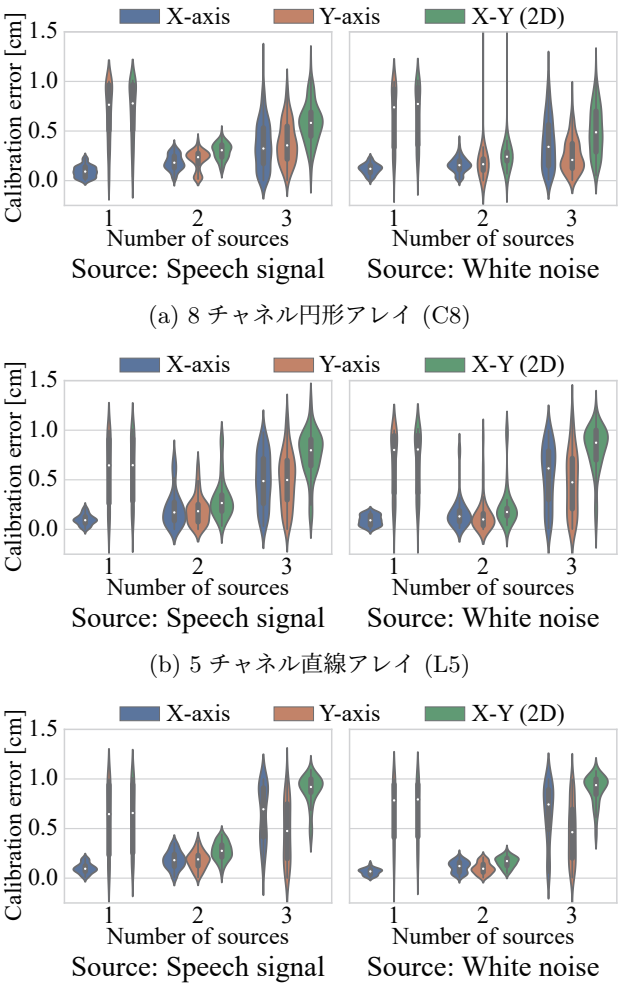
マイクロホンアレイは 8 チャネル円形，5 チャネル直線，4 チャネル長方形の 3 種類を用いた．以下では，それぞれのマイクロホンアレイを **C8**, **L5**, **R4** と呼ぶ．マイクロホンアレイの形状を図 2 に示す．マイクロホンアレイ C8 は半径 12 cm の円形で，45° おきにマイクロホンを配置した．マイクロホンアレイ L5 は長さ 20 cm の直線形で，5 cm おきにマイクロホンを配置した．マイクロホンアレイ R4 は 14 cm × 8 cm の長方形で，各頂点にマイクロホンを配置した．いずれのマイクロホンアレイも 2 次元平面上で構築され，音源もマイクロホンアレイと同一の平面上に配置したため，次元数は $d = 2$ と設定した．音源信号は平面波であると仮定し，到来方向は 0°, 90°, 150° とした．

本性能評価では，以下の二つの要素を変化させる．

- 音源数：ズレの大きさを 1 cm に固定した上で音源数を 1 (0°), 2 (0°, 90°), 3 (0°, 90°, 150°) とした際の性能の評価を行った．
- 与えるズレの大きさ：音源数は 2 つに固定し，Ch1 を除くマイクロホンに対して，1 cm, 2 cm, 3 cm のズレを与えた際の性能の評価を行った．

4.1　音源数の変化に対する評価

図 3 に音源数の変化に対するキャリブレーション誤差を示す．いずれのマイクロホン形状，音源の種類に関しても，2 音源を用いた場合が最もキャリブレーション誤差が小さかった．1 音源のみを用いた場合は，X 軸方向のキャリブレーション誤差は小さいものの，Y 軸方向のキャリブレーションがほとんど行われていない．この場合は音源が X 軸上に存在するため，その方向のキャリブレーションのみが行われ，直交する Y 軸方向のキャリブレーションは行われないためだと解釈できる．3 音源を用いた場合は，2 音源を用いた場合に比べ



(a) 8 チャネル円形アレイ (C8)

(b) 5 チャネル直線アレイ (L5)

(c) 4 チャネル長方形アレイ (R4)

図 3: 音源数の変化に対するキャリブレーション誤差．X 軸方向，Y 軸方向，X-Y 平面上での誤差の分布を表す．

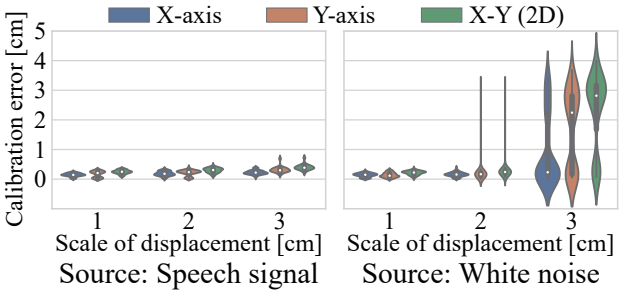


図 4: 8 チャネル円形アレイ (C8) を用いた場合のマイクロホンのずれの大きさに対するキャリブレーション誤差.

てキャリブレーション誤差が増大する傾向にある．音源数が増加すると，キャリブレーションに利用できる音源方向に関する情報は増えるものの，音源分離（音源スペクトルの推定）が困難になるため，誤差が増大したと考えられる．ただし，本実験ではすべての音源が常に発音していたが，実際には音源の発音区間は時間的にスパースになることが多いため，実環境である程度長時間の音響信号を用いればこの問題は解決される可能性がある．

マイクロホンアレイの形状を変化させても，キャリブレーション誤差の傾向に大きな違いは確認されなかった．したがって，提案手法は様々なマイクロホンアレイ形状に適用可能であるといえる．

4.2 ずれの大きさに対する評価

図 4 にマイクロホンアレイ C8 を用いた場合のマイクロホン位置のずれの大きさに対するキャリブレーション誤差を示す．音源信号に音声を用いた場合は，ずれの大きさ 1 cm, 2 cm, 3 cm に対してキャリブレーション後の誤差の中央値が 0.23 cm, 0.35 cm, 0.37 cm であり，ずれを 77 %, 83 %, 88 % 減少させることができた．一方ホワイトノイズを用いた場合は，ずれの大きさが 1 cm–2 cm のときは音声を用いた場合と同様の傾向を示しているが，ずれの大きさが 3 cm のときにキャリブレーション誤差が大きく増大した．図 4 を見ると，0.5 cm 付近と 3 cm 付近に誤差の分布が 2 極化していることが分かる．複数のホワイトノイズが重複するとスペクトログラムの時間周波数上でのスパース性がほぼ成り立たず，音源分離（音源スペクトルの推定）が困難になるため，誤差が増大したと考えられる．

4.3 音源定位性能の評価

マイクロホンアレイ C8 を用いてキャリブレーション前後で音源定位を行った際の推定誤差を図 5 に示す．

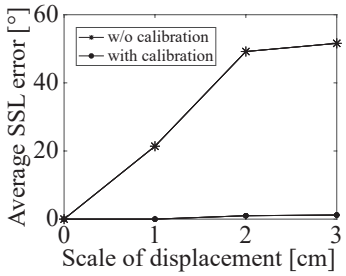


図 5: キャリブレーション前後の音源定位誤差の変化.

音源数は 2 に固定し，音源信号には音声を用いた．キャリブレーションを行わない場合はマイクロホン位置ずれの増大に伴い音源定位誤差も増大するが，キャリブレーションにより定位誤差は 2° 以下，キャリブレーション前の 4 % 以下に抑制された．上述の実験で示したように，提案手法はマイクロホン位置の真値を完璧に推定するわけではないが，音源定位性能を十分に改善することが示された．

5 まとめ

本論文では，マイクロホン位置に関する確率的生成モデルを用いたマイクロホンのキャリブレーションを提案した．確率的生成モデルを構築し，MAP 推定によるキャリブレーション手法を構築し，提案手法の性能評価のため，評価実験を行った．評価の結果，複数の同時音源を用いたキャリブレーションが実行可能であることが確認された．また，提案法によるキャリブレーションの結果，音源定位の性能向上を確認し，有効性を示した．

本論文ではグリッドサーチしか試していないため，他の最適化手法も検討してみる必要がある．また，マイクロホン位置の推定から伝達関数の推定にまで手法を拡張する必要性もある．

謝辞

本研究は JSPS 科研費 JP19K12017, JP19KK0260 および JP20H00475 の助成を受けた．

参考文献

[1] Zhang, C., Florencio, D., Ba, D. E. and Zhang, Z.: Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings, *IEEE Trans. Multimedia*, Vol. 10, No. 3, pp. 538–548 (2008).

[2] Nishiura, T., T., Y., S., N. and K., S.: Localization of multiple sound sources based on a CSP analysis with a microphone array, in *ICASSP 2000*, Vol. 2, pp. 1053–1056 (2000).

[3] Nakamura, K., Nakadai, K., Asano, F., Hasegawa, Y. and Tsujino, H.: Intelligent sound source localization for dynamic environments, in *IROS 2009*, pp. 664–669 (2009).

[4] Nakadai, K., Nakajima, H., Hasegawa, Y. and Tsujino, H.: Sound source separation of moving speakers for robot audition, in *ICASSP 2009*, pp. 3685–3688 (2009).

[5] Nakadai, K., Okuno, H. G. and Mizumoto, T.: Development, deployment and applications of robot audition open source software HARK, *J. Robot. Mechatron.*, Vol. 29, No. 1, pp. 16–25 (2017).

[6] Su, D., Vidal-Calleja, T. and Miro, J. V.: Simultaneous asynchronous microphone array calibration and sound source localisation, in *IROS 2015*, pp. 5561–5567 (2015).

[7] Miura, H., Yoshida, T., Nakamura, K. and Nakadai, K.: SLAM-based online calibration of asynchronous microphone array for robot audition, in *IROS 2011*, pp. 524–529 (2011).

[8] Raykar, V. C., Kozintsev, I. V. and Lienhart, R.: Position calibration of microphones and loudspeakers in distributed computing platforms, *IEEE Trans. Speech and Audio Process.*, Vol. 13, No. 1, pp. 70–83 (2005).

[9] Ono, N., Shibata, K. and Kameoka, H.: Self-localization and channel synchronization of smartphone arrays using sound emissions, in *APSIPA ASC 2016*, pp. 1–5 (2016).

[10] Thrun, S.: Affine structure from sound, in *NIPS'05*, pp. 1353–1360 (2005).

[11] Crocco, M., Del Bue, A. and Murino, V.: A bilinear approach to the position self-calibration of multiple sensors, *IEEE Trans. on Signal Process.*, Vol. 60, No. 2, pp. 660–673 (2012).

[12] Birchfield, S. T. and Subramanya, A.: Microphone array position calibration by basis-point classical multidimensional scaling, *IEEE Trans. on Speech and Audio Process.*, Vol. 13, No. 5, pp. 1025–1034 (2005).

[13] Birchfield, S. T.: Geometric microphone array calibration by multidimensional scaling, in *ICASSP 2003*, Vol. 5, pp. 157–160 (2003).

[14] Valin, J.-M., Rouat, J. and Michaud, F.: Enhanced robot audition based on microphone array source separation with post-filter, in *IROS 2004*, Vol. 3, pp. 2123–2128 (2004).

[15] Févotte, C., Bertin, N. and Durrieu, J.-L.: Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis, *Neural Computation*, Vol. 21, No. 3, pp. 793–830 (2009).

[16] Smaragdis, P., Févotte, C., Mysore, G. J., Mohammadiha, N. and Hoffman, M.: Static and dynamic source separation using nonnegative factorizations: A unified view, *IEEE Signal Process. Mag.*, Vol. 31, No. 3, pp. 66–75 (2014).

[17] Uhlich, S., Giron, F. and Mitsufuji, Y.: Deep neural network based instrument extraction from music, in *ICASSP 2015*, pp. 2135–2139 (2015).

[18] Nugraha, A., Liutkus, A. and Vincent, E.: Multichannel audio source separation with deep neural networks, *IEEE/ACM Trans. Audio, Speech, Language Process.*, Vol. 24, No. 9, pp. 1652–1664 (2015).

[19] Takamichi, S., Mitsui, K., Saito, Y., Koriyama, T., Tanji, N. and Saruwatari, H.: JVS corpus: free Japanese multi-speaker voice corpus, *arXiv:1908.06248 [cs.SD]* (2019).

テニスにおける打球音を用いた球種識別の検討

Detection of Ball Spin Direction Using Hitting Sound in Tennis

山本修己 [1*] 西田健次 [1] 糸山克寿 [1] 中臺一博 [1,2]

[1] 東京工業大学工学院システム制御系
[1] Department of Systems Controlling and Engineering, School of Engineering, Tokyo Institute of Technology
[2] ホンダ・リサーチ・インスティテュート・ジャパン
[2] Honda Research Institute Japan Co., Ltd.

Abstract: This paper describes the detection of rotation direction using the hitting sound of tennis balls. Since each ball rotation direction has a slightly different rotation direction and trajectory, there should be a difference in the hitting sound. To distinguish the characteristics of ball rotation direction, a database was constructed that combines the hitting sound recorded experimentally with ball rotation direction. Since it is difficult to distinguish audible differences in hitting sounds by ear, it is necessary to identify them using measuring instruments. For this purpose, after extracting the amplitude spectrum by fast Fourier transform of the shot sound, the entire data was normalized and classified by a support vector machine. As a result of evaluating this method, a high accuracy was obtained in every rotation direction. The proposed method also evaluated the hitting sound from a YouTube video in an unknown environment and showed the effectiveness.

1 はじめに

近年，世界的にスポーツに科学技術を取り入れる動きが活発化している．複数のカメラを搭載し，コンピュータビジョン技術を用いて選手やボールの動きを追跡できるスマートコート [1, 2] は，サッカーやバスケットボールなど様々なスポーツで活用されている [3]．テニスの審判判断の精度を高めるために開発されたスマートコートシステム「ホークアイ」は，8 台の超高速カメラを搭載し，ボールの軌道や着地点を特定して，プロの試合で審判に情報を伝え，試合の円滑な進行を支援できる [4]． しかし， スマートコートシステムは大規模であり，設置の労力やコストを考えると個人での利用は難しい．また，スポーツではコンピュータビジョン技術を用いたフォーム解析が重要な課題となっており [5–8]，これまでに様々なスポーツを対象に多くの研究が報告されている．

テニスは広いコートを一人ないし二人でカバーする必要があるため，ボールの軌道を予測することが重要になる．ボールの軌道は回転の影響を大きく受けるため，プレーヤーには相手が打った球種を判別する能力が求められる．

図 1: スマートテニスセンサ：ラケットのグリップエンドに取り付けることができ，ストロークの球種，速度，回転数などを測定することができる．

Canal-Bruland らは，被験者にプロテニスの試合をビデオで見てもらい，その時のボールの軌道を予測してもらう実験を行い，その結果ボールの軌道を予測する上で，打球音が重要な要素であること報告している [10]．ボールの軌道を予測する際には，球種がスピン，フラット，スライスいずれであるかを認識することで，ボールの大まかな軌道を予測することができる．なお，スピ

*連絡先： 東京工業大学工学院中臺研究室
東京都目黒区大岡山 2-12-1 西 8 号館 W 棟 W310 号室
E-mail: yamamoto@ra.sc.e.titech.ac.jp

本稿は ic-sports で採択された”Detection of Ball Spin Direction Using Hitting Sound in Tennis” を和訳したもの（一部改変）です.

ンは，図 2(a) のようにボールの上面をこするように打つことで，ボールの軌道と同方向の回転をかける．これにより，ボール下面の抗力が低下し，ボールが下に落ちる軌道となる．フラットは，図 2(b) のように，ボールを回転させないように打つ．実際には，軽いスピンがかかることが多い．スライスは，図 2(c) のようにラケットを傾けてボールの下面をスライドさせるように打つ．スピンと反対に，ボールの上向き方向に力が働くのでボールが落ちにくくなる．一方で，ボールの軌道と逆方向の回転がかかるため，スピンやフラットよりも速度が低くなる傾向がある．

実際に球種判別が可能な製品として，図 1 に示すスマートテニスセンサが開発されている [9]．スマートセンサは，加速度センサと 3 軸ジャイロセンサを用いて，ボールの速度，球種，回転数を測定することができる．プレーヤは，スマートテニスセンサを用いて様々な球種を打ち方を効率的に学習することができるが，相手の打った球種に関する情報は得ることができない．さらにセンサをグリップエンドに装着する必要があるため，試合で邪魔になったりプレーに影響が出るといった欠点がある．

浅野らはボールにマーカーを取り付け，高速度カメラを用いた 3 軸の回転角度と回転数を計測する手法を報告した [11]．2 台のカメラでカメラパラメータからボール中心の 3 次元位置を求め，ボールの軌道を推定しているが，カメラの設置が必要であり，オクルージョンが発生する場合は計測が困難である．

本稿では，この問題を解決するために打球音から 3 種類の球種を識別する技術の構築を目指す．特に，プレーヤのパフォーマンス向上に寄与するのは，サーブよりもストロークであることから，ストローク打球の球種識別にフォーカスする．このために，まず，ストローク打球とその球種を対応づけたデータセットを設計・構築した．次に，打球音から球種を識別する手法として SVM を用いた球種識別器を提案する．提案手法は，学習データが少なく，クラス間のデータ数のバランスが悪くても識別学習が可能な点が特長である．最後に提案手法を用いて，構築したデータセット，ならびに YouTube から抽出した打球音に対して識別実験を行った．結果として，チャンスレートを大きく超える精度が得られ，テニスの球種判別に有効であることを示すことができた．

以下，第 2 節では関連研究，第 3 節では構築したデータベース，第 4 節では球種の識別手法提案，第 5 節では提案法を用いた評価実験の結果と考察についてそれぞれ延べ，第 6 節でまとめる．

2 関連研究

本節では，本稿で着目する打球音を扱った関連研究について述べる．

視覚障がい者用の卓球であるサウンドテーブルテニスでは打球音が聞こえなければファウルとするルールが存在している． この判定には，審判の聴覚を頼りに行っているため，ヒューマンエラーを避けることができないという問題があった．小薬ら [12] はこの問題の解決を目指し，デジタルオーディオテープレコーダーで打球音を騒音計を介して録音し，ウェーブレット変換解析を行い，周波数領域の成分に着目して打球音の存在判定を行う手法を提案している．

Zhang らは，テニスの打球音の特性調査を行った [13]．この研究では，サービス打球音をデュース側とアドバンテージ側からそれぞれ 15 サンプルずつ抽出し，時間領域で波形を重ね合わせて特徴を比較している． 具体的には，テレビ映像を録画して打球音を抽出し，各音波形の最初のピークを選手ごとに重ね合わせて，最初のピークの平均振幅と最初のピークから最後のピークまでの到達時間から選手ごとの音の特徴を分析した．結果として，球速と打球音に相関があることを報告している．

これらの研究を総括すると，打球音解析の重要性は広く認識されており，テニスに限らず，研究が行われている．一方でその内容は，打球そのものの判定や音量と球速との相関の解析にとどまっており，球種判定は十分に取り組まれていない．

3 打球音データベースの構築

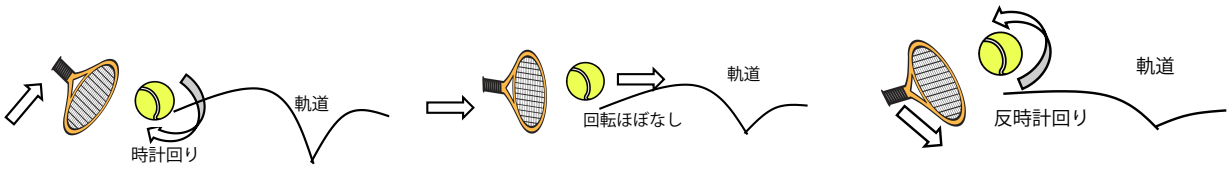
本稿では，従来研究で解決されていない課題として，打球音から球種を識別する技術を扱う．打球音から球種を識別する上で，打球音と球種が紐づいたデータセットが必要である．上述したように，筆者らの知る限り，打球音から球種を判別するタスクは扱われていないため，これを行うためのデータセットも存在しない．この問題を解決するため，テニスの球種判定用データセットの設計および構築を行った．

3.1 録音実験

球種識別用のテニス打球音データセットを構築するため，スピン，フラット，スライスの打球音を球種ラベル付きで収録した．

収録条件は以下の通りである．

- 日時：2019/12/10 11:00〜13:00



(a) スピンの回転と打球方向　(b) フラットの回転と打球方向　(c) スライスの回転と打球方向

図 2: 3 球種の回転と打球方向

表 1: 録音に使用した設備

機材	型番・設置位置・個数
マイクロホン	TAMAGO-03
マイクロホン位置	ポール近傍（各ポールに 1 個，計 2 個）
テニスボール	20 球（新球）
ラケット	SRIXON REVO CV3.0 (SR21802)
PC	16 kHz，16-bit 録音

- 場所：つくば市二宮公園テニスコート（ハードコート，屋外）
- 天気：晴れ，ほぼ無風.
- 打者：テニス歴 15 年の男性一名

図 3 に収録の状況を示す．打球音データベースは 8*ch* 音響信号 $\times 2$ が収録されており，解析にはその中から任意の 1 チャネルを選択した．また，表 1 に収録に使用した機器の仕様を示す．打球収録は，以下の手順で行った．

1. 球出し役が打者のためにボールを投げる．
2. 打者は，打者が決めた球種と力でボールを打つ．
3. 打者はそれらを記録係に伝える．

計 92 回の試行を行った．

3.2　打球音クリップの抽出

各録音について，インパクトの瞬間を含むように 50 ms のクリップを抽出した．この作業は Audacity を使用して，92 個の録音された全ての打球音データに対して手動で行った．このようにして，92 個の打球音とそれに対応する球種ラベルからなる打球パターンデータセットを収集した．92 個の内訳はスピン 46 個，フラット 16 個，スライス 30 個である．

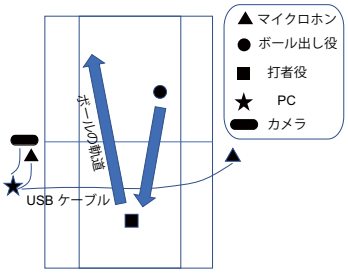


図 3: 実験設定．矢印は打球の軌道を示す.

4　球種識別手法

この節では打球音から球種を識別する手法を提案する．提案手法は，図 4 に示すように周波数解析，データ正規化，次元削減，2 クラス SVM からなる．以下，各処理について説明する．

4.1　周波数解析

入力音は前節で述べたように 50ms で切り出しているので，この切り出した信号に対してフーリエ変換を行う．フーリエ変換は，複雑な音を分解する周波数解析手法である．フーリエ変換は，対象となる信号が周期関数であることを求めるため，一般的な非周期信号に対してフーリエ変換を行う際に，ハニングやハミングといった窓関数を適用することが多い．今回収録した打球音信号は，50ms の中にほぼすべての成分が含まれるので，切り出した信号の両端ははぼ振幅が 0 となっている．このため，本稿では窓関数には方形窓を用いた（窓関数を用いない）．また，高速フーリエ変換 (Fast Fourier Transform, FFT) を用いれば，フーリエ変換を高速に実現できるため，FFT を用いた周波数解析を行った．このためには FFT 長を 2 のべき乗に設定する必要がある．構築した打球音データベースのサンプリング周波数は 16kHz であるため，切り出した信号の信号長 50ms は 800 サンプルに相当する．1,024 に足らない分については 0 埋め（0 パディング）を行った．結果としては 0–8kHz（ナイキスト周波数）の周波数成分から

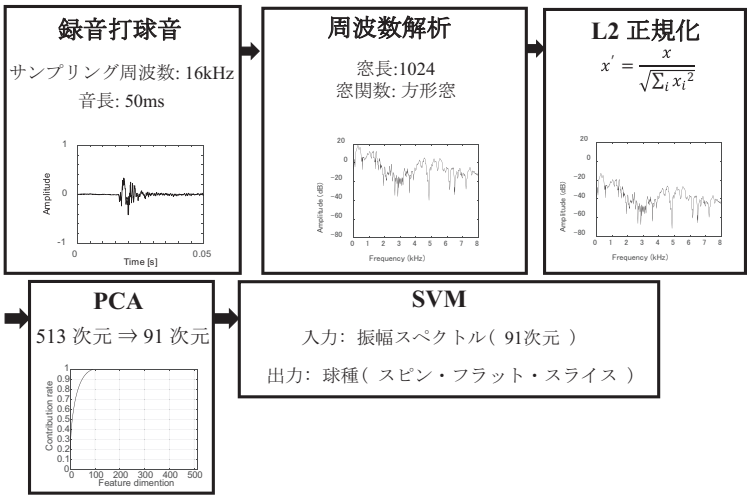


図 4: 提案手法のフローチャート

なる 513 次元の特徴量を各打球音サンプルごとに得た．解析には MATLAB を用いた．

4.2 データ正規化

次に，各打球音に対して得られる 513 次元の特徴量 $\mathbf{x} = \{x_i | i = 1 \cdots 513\}$ に対して，インパクト位置の違いによる打球音のばらつきを防ぐために以下の式により L2 ノルム正規化を行い，$\bar{\mathbf{x}}$ を求めた．

$$\bar{\mathbf{x}} = \left(\sum_{i=1}^{513} x_i^2 \right)^{-\frac{1}{2}} \mathbf{x} \tag{1}$$

4.3 主成分分析による次元削減

汎化性能を確保するため主成分分析 (Principal Component Analysis, PCA) [14] を利用した特徴量の低次元化を行った．PCA は互いに直交する主成分を推定する手法で，大きい寄与率を持つ主成分はそれだけよくデータセットを説明することができる．大きい寄与率を持つ主成分を選択することで，低次元でデータセットを表現することができる．得られた 92 個の 513 次元特徴量に対して，主成分分析 (PCA) を適用した．PCA の手順は以下の通りである．$\mathbf{x}_i$ $(i = 1, \ldots, N)$ は i 番目の D 次元データを表す．まず $\mathbf{x}_i$ から全データ平均を引くことで i 番目のデータの偏差を求める．

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \frac{1}{N} \sum_{j=1}^{N} (\mathbf{x}_j). \tag{2}$$

次に分散共分散行列 $\mathbf{X}$ は以下のように計算できる．

$$\Sigma_{\mathbf{X}} = \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T. \tag{3}$$

以下の式を用いて，得られた共分散行列 $\mathbf{X}$ に対して固有値分解を行う．

$$\Sigma_{\mathbf{X}} \mathbf{U} = \mathbf{U} \Lambda, (\mathbf{U}^T \mathbf{U} = \mathbf{I}), \tag{4}$$

ここで $\mathbf{U}$ は i 列目が固有ベクトルである $D \times D$ の正方行列であり，Λ は対角成分が固有値に対応する対角行列である．このとき，最大固有値の固有ベクトルが最も寄与率の大きい第一主成分に対応する．

図 5 は，513 次元の特徴量に対する累積寄与率を示したものである．第 91 主成分までの累積寄与率が 1 となったため，91 次元に特徴量を低次元化することにした．

4.4 SVM

提案手法では，2 クラス分類を行う．例えば，対象がスピンの場合，その音がスピンの打球音であるかスピン以外かを識別する． これにより，スピン，フラット，スライスの 3 種類の 2 クラス識別器を構築する．

PCA により得られた低次元入力ベクトル $\mathbf{s}$ に対し，識別関数 y は以下の式で定義される．

$$y = \mathrm{sign}\left(\mathbf{w}^T \mathbf{s} - h \right), \tag{5}$$

$\mathbf{w}$ は入力に対する重みベクトル，h は閾値を示す．関数 $\mathbf{sign}(u)$ は符号関数であり， $u > 0$ のとき，1 を出力し，$u \leq 0$ のとき，-1 を出力する．言い換えれば，式 (5) は

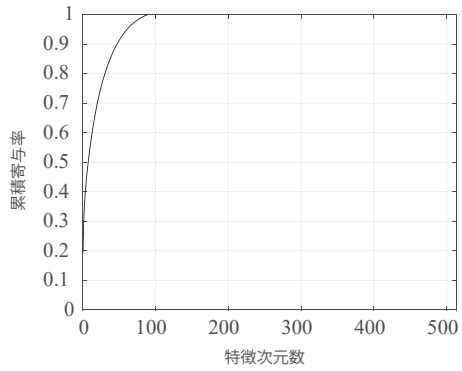


図 5: 構築されたデータベースに対する PCA の累積寄与率．横軸は特徴次元，縦軸は選択された特徴次元における累積寄与率.

$\mathbf{w}$ により定義される分離超平面を用いて，$\mathbf{s}$ で表現される空間を 2 つの部分空間に分離している．SVM [15,16] は，分離超平面と最も近いサンプルとの距離（マージン）が最大となる分離超平面を決定する方法である．しかし，従来の SVM ではすべての入力サンプルは以下の式のように線形分離可能でなければならない.

$$(\mathbf{w}^T\mathbf{s}_i - h)\cdot t_i \geq 1,\ i = 1,\ldots,N, \tag{6}$$

ここで t_i は，i 番目の入力ベクトル $\mathbf{s}_i$ に対する正解クラスラベル（(1 または -1）である.

これは以下の式のようにサンプルは 2 つの超平面により分離されることを示し，超平面の間にはサンプルは存在しないことを意味している.

$$H1 \quad : \quad \mathbf{w}^T\mathbf{s}_i - h = 1, \tag{7}$$

$$H2 \quad : \quad \mathbf{w}^T\mathbf{s}_i - h = -1, \tag{8}$$

ここで，各分離超平面間の距離は $1/\|\mathbf{w}\|$ で定義される.

線形分離可能の制約を緩和するために，H1 と H2 の間に訓練サンプルが入り込むことを許すソフトマージンを導入する．ソフトマージンを導入することで $\mathbf{s}_i$ に対する距離パラメータ ξ_i が導入される．このパラメータは $t_i = 1$ における i 番目のサンプルを以下のように示す.

$$\xi_i = \begin{cases} -\mathbf{w}^T\mathbf{s}_i + h + 1 & (\mathbf{w}^T\mathbf{s}_i - h < 1) \\ 0 & (otherwise) \end{cases} \tag{9}$$

$t_i = -1$ における i 番目のサンプルは以下のように示す.

$$\xi_i = \begin{cases} \mathbf{w}^T\mathbf{s}_i - h + 1 & (\mathbf{w}^T\mathbf{s}_i - h > -1) \\ 0 & (otherwise) \end{cases} \tag{10}$$

ソフトマージン SVM は $\xi = \{\xi_i | i = 1,\cdots,N\}$, が，式 (11) を満たすという条件のもと，式 (12) を最小化する

表 2: データ数とクラス重み

識別器	スピン	フラット	スライス
正解	46	16	30
不正解	46	76	62
q_i	1	4.75	2.07

最適化問題として，定義される.

$$\xi_i \geq 0,\ t_i\cdot(\mathbf{w}^T\mathbf{s}_i - h) \geq 1 - \xi_i,\ (i = 1,\ldots,N), \tag{11}$$

$$L(\mathbf{w},\xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} q_i\xi_i \tag{12}$$

ここで C は ξ に対するコストパラメータである．また q_i は以下の式で定義される i 番目サンプルの重みである.

$$q_i = \begin{cases} 1 & \mathbf{s}_i \in C_l \\ |x \in C_l|/|x \in C_s| & \mathbf{s}_i \in C_s \end{cases} \tag{13}$$

C_s と C_l はそれぞれデータ数の多いクラスと少ないクラスを示す.

5 評価

提案手法は，構築された打球音データベースと YouTube 動画から抽出された打球音を用いて検証する.

5.1 打球音データベースの識別

構築したデータセットを提案手法で評価した．データセットのサンプル数は 92 と少ないため，LOOCV で評価を行った．LOOCV とはサンプルを 2 つに分割し，サンプル中の 1 つのデータで構成される検証データとサンプル中の残りのデータで構成される訓練データに分けて検討する手法である．この手法には *N-1* 個のサンプルで観測を行うため，少ないデータでも過適合を防ぐことができるという利点をもつ．ソフトマージンなどのハイパーパラメータは精度が最大化されるように最適化した.

表 3 は 91 次元の特徴量を用いた 2 クラス識別の混同行列を示している．各表は，識別の真陽性，真陰性，偽陽性，偽陰性を示している．この表を見ると，3 球種とも 65%以上の精度が得られているが，精度にはそれぞれ特徴がある．フラットの適合率や再現率はスピンとスライスより低くなっている．これはフラットのデータ数が他の 2 種類よりも，少ないこと，フラットはスピンとスライスの中間的な回転数を持っているため判別が難しいことがその理由であると考えられる.

表 3: 打球音データベースを用いて識別した際の混合行列．92 サンプルは全て 91 次元特徴量で識別.

(a) スピン識別に対する混同行列

	スピン (正解)	その他 (正解)
スピン (予測)	28	13
その他 (予測)	18	33

(b) フラット識別に対する混同行列

-	フラット (正解)	その他 (正解)
フラット (予測)	8	6
その他 (予測)	8	70

(c) スライス識別に対する混同行列

	スライス (正解)	その他 (正解)
スライス (予測)	24	11
その他 (予測)	6	51

表 4: YouTube 打球音識別に対する混合行列．91 次元の特徴量を使用.

(a) スピン識別に対する混同行列

	スピン (正解)	その他 (正解)
スピン (予測)	2	5
その他 (予測)	8	15

(b) フラット識別に対する混同行列

-	フラット (正解)	その他 (正解)
フラット (予測)	2	1
その他 (予測)	8	19

(c) スライス識別に対する混同行列

	スライス (正解)	その他 (正解)
スライス (予測)	7	9
その他 (予測)	3	11

5.2 YouTube クリップの解析

YouTube から現世界 4 位のロジャー・フェデラー選手が 2020 年 1 月の全豪オープン (ハードコート) で練習している動画*を選定し，その動画の打球音から球種の識別を試みた．抽出した打球音はスピン，フラット，スライス各 10 本ずつの計 30 本である．構築したデータベースと同様の処理を行うため，YouTube のサンプリングレートである 44.1 kHz を 16 kHz にリサンプリングし，その後インパクト音を含む 50 ms の打球音を手動で抽出した．今回は，この 30 の打球音に対して先程の 92 個のデータを学習したモデルを使って予測を行い，その評価を行った．表 4(a)–(c) にその結果を示す．精度はどの球種も 60%程度であった．再現率では，スライスが 70%と最も高い値を示した．これは，スライスは他球種と回転方向が反対であることから他球種とは識別しやすいためと考えられる．しかし，スライスも適合率は低く他の球種も適合率や再現率は低いため，高い評価指標を得ることができていないのが現状である．この問題は今後の課題としたい．

*https://youtu.be/hTn42aJIhk8

6 むすび

本稿は，テニスの打球音から球種識別を行う手法を主成分分析とサポートベクトルマシンを用いた手法として提案した．構築した打球音データベース，YouTube 動画から抽出したプロテニス選手の打球音に対して提案手法を適用し，その有効性を示した．用いたデータ数が限られているため，今後は大量のデータを用いて手法の一般性や頑健性を確認する必要がある．実応用に向けては入力音に混入する雑音への対策も課題である．

謝辞

本研究は JSPS 科研費 19K12017, 19KK0260 および 20H00475 の助成を受けた.

参考文献

[1] SecondSpetrum, The next way of seeing sports, https://www.secondspectrum.com/index.html (2020)

[2] Playsight: Smartcourt, https://www.playsight.com (2020)

[3] Seo, S.-W., Kim, M., and Kim, Y.: Optical and acoustic sensor-based 3D ball motion estimation for ball sport simulators, *Proceedings of the 2017 International Conference on Information and Communication Technology Convergence*, Vol. 18, No. 1323. (2018)

[4] Baodon, Y.: Hawkeye technology using tennis match, *Computer Modelling & New Technologies*, Vol. 18, No. 12, pp. 400–402. (2014)

[5] Cust, E. E., Sweeting, A. J., Ball, K., and Robertson, S.: Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance, *Journal of Sports Sciences*, Vol. 37, No. 5, pp. 568–600. (2019)

[6] Appelbaum, L. G., Erickson, G.: Sports vision training: A review of the state-of-the-art in digital training techniques, *International Review of Sport and Exercise Psychology* ,Vol. 11, No. 1, pp. 160–189. (2018)

[7] Okamoto, H., Moro, A., Yamashita, A., and Asama, H.: Toward sports training service with the interactive learning platform, In Sawatani, Y., Spohrer, J. C., Kwan, S. K., and Takenaka,T., editors, *Serviceology for Smart Service System, Selected papers*

of the 3rd International Conference of Serviceology, ICServ 2015, San Jose, CA,USA, 7-9 July 2015, pp. 231–236. Springer. (2015)

[8] Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y. A.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1. (2019)

[9] Zepp: Smart tennis sensors, https://www.secondspectrum.com/index.html (2020)

[10] Canal-Bruland, R.: Auditory contributions to visual anticipation in tennis. *Psychology of Sport and Exercise*, Vol. 36, pp. 100–103. (2018)

[11] Toshiro,Asano., Yuji,Serikawa., Kazuyuki,Ishiguro., Seiji,Itoh.: Analysis of Tennis Ball Rotation and Trajectory by Image Proccessing, *Journal of the Japan Society for Precision Engineering*; Vol. 82, No. 2, pp. 168–174. (2016)

[12] Kogusuri, Y., Sato, T., Toyoda, K., Miyato, S.: Developement of the holding judgement technology using batted ball sound of sound table tennis, *The Proceeding of the Conference on Information, Intelligence and Precision Equipement : IIP*, No. 8, pp. 49–52. (2008)

[13] Zhang, D., Yokohama, K., and Yamamoto, Y.: Characterisitics of impact sound in tennis service among top-level players, *Nogoya J. Health, Physical, Fittness, Sports*, Vol. 40, No. 1, pp. 37–43. (2017)

[14] Diamantaras, K. I., Kung, S. Y.: Principal compornent neural networks: Theory and applications, In Karhunen, J., editor, *Pattern Analysis and Applications*, pp. 74–75, John Wiley & Sons. (1998)

[15] Scholkopf, B., Burges, C. J. C., and Smola, A. J.: In *Advances in Kernel Methods - Support Vector Learning*. The MIT Press, USA. (1999)

[16] Vapnik, V. N.: In *Statistical Learning Theory*. John Wiley and Sons. (1998)

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-057-9 (11/20)

表情による感情推定と音声による感情推定手法の検討

Examination of voice-based sentiment estimation method using facial expression-based sentiment estimation

西田健次[1*]　山田亨[2]　糸山克寿[1]　中臺一博[1,3]
Kenji Nishida[1], Toru Yamada[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: 表情，声などから人間の感情推定を行う需要は年々高まってきており，特に脳卒中以後のリハビリテーションや認知症進行抑制のための認知活性化療法での介入効果推定手法に取り入れられ，その効果が確認されつつある．表情からの感情推定に関して，単純な線形識別器による表情識別器により，個人内での感情推定が有効なことを示してきた．音声情報による感情推定も需要が高まってきているが，リハビリテーションや認知活性化療法の効果推定手法としての有効性は確認されていない．本稿では，音声情報に関しても，単純な線形識別器による感情推定手法を適用し，その有効性を確認した．また，音声付き動画に表情識別による感情推定を適用し，表情検出によって音声情報への感情のアノテーションの可能性を探った．音声情報での感情推定に関しては，24 人の被験者に対して 24-way（23 人のデータで学習し，残る 1 人のデータで検証する）交差検証を行った結果，75%から 85%の汎化性能を達成することができた．また，この動画に対して，異なるデータセットで学習した表情識別器を適用したところ，喜び（笑顔）に対する検出性能が特異的に高く，笑顔検出による喜びの感情を表す音声情報へのアノテーションの有効性が示された．

1　はじめに

人間の感情推定は，昨今の人工知能技術において重要な課題であり，多くの分野でその適用が提案されている．特に，脳卒中後のリハビリテーションなどの，脳機能傷害に対するリハビリテーションにおいては，身体的なリハビリテーションに比べ，その有効性を評価する指標を得ることが難しかったが，笑顔検出による「快」の感情を推定することによる客観的な評価手法が提案され，その有効性が示されてきている [1]．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 [2]，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．音声情報から感情推定を行う手法も提案されており，表情検出手法と同様にリハビリテーションへの適用が期待されている．

*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

脳機能障害患者の感情推定手法を確立する際の大きな課題は，個人情報保護などのために実際の脳機能障害患者から収集された公開データセットの入手が困難であること，また，入手できたとしても充分なデータ数とは言えないことが挙げられる．そのため，公開されたデータセットによる感情推定器（感情識別器）を構成し，脳機能障害患者の感情推定に援用する手法が提案されている．西田らは，表情が乏しくなることが多い脳機能障害患者 [3, 4] に対して，健常者のデータセットで学習した線形識別器の推定値を用いることで乏しい表情変化に対応する手法を提案した [5]．この手法では，一人の被験者に対する感情推定器の最大値と最小値によって感情推定値を正規化することで，個人内での感情の変化を捉えることに成功している．そして，正規化された感情推定値を用いることで感情の変化を個人間でも比較できる可能性を示した．音声情報による感情推定においても，脳機能障害患者から収集され，感情に対するアノテーションの行われたデータは乏しく，表情検出による手法と同様に健常者のデー

タセットで学習した結果を援用する必要があると考えられる．また，感情推定の精度を向上するためには，より多くの脳機能障害患者からデータを収集する必要があり，データ収集を容易にするためにはアノテーションの自動化が必要となってくる．

本稿では，Ekman の基本感情 [6] にアノテーションされた音声付き動画データセットを用いて，音声情報による感情推定器を構成した．音声情報による感情推定では，24-way Cross-Validation（24 人分のデータセットに対して 23 人分のデータで学習し，残る 1 人のデータで評価を行う）において，77%から 85%の精度を得ることができ，十分な汎化性能があることが示された．動画部分に対して，[5] で述べた学習済み表情識別器による感情推定を適用し，音声情報に対して表情識別器を用いた感情のアノテーションの可否を検証した．その結果，喜びの表情（笑顔）検出精度 (precision) が充分に高く，笑顔（喜び）表情識別を用いた音声情報へのアノテーションが有効なことが示された．

2　感情推定手法

感情推定は，人工知能技術において重要な課題であり，これまで多くの研究がなされてきている．コンピュータビジョンを用いた感情推定は，基本感情を代表する表情を識別し，その強度の推定値を感情の推定値として用いている．音声情報を用いて感情推定を行うためには，基本感情を代表する音声情報から，感情の推定値をエう必要がある．本稿では，[5] と同様に，基本感情にアノテーションされたデータ（音声付き動画）を用いて線形識別器の学習を行い，閾値処理を行う前の識別器出力を感情の推定値としても用いることとした．

2.1　線形識別器による感情推定手法

Ekman は，基本感情を,「怒り (anger)」,「嫌悪 (disgust)」,「恐怖 (fear)」,「喜び (happiness)」,「悲しみ (sadness)」,「驚き (surprise)」の 6 種に類型化したが，其々の基本感情は有る無しの二値ではなく，その感情を持たない状態から，その感情を強く抱く状態までの連続値として考える方が妥当である．また，6 種の基本感情は，必ずしも排他的なものではなく，複数の基本感情が組み合わさった状態もあると考えられる．したがって，感情推定は，6 クラス（あるいは，感情的に中立を含めて 7 クラス）の多クラス識別器を構成するより，それぞれの感情強度を推定する識別器を構成する方が妥当である．その一方で，感情に対するアノテーションを行う際に，その強度まで指定するのは難しく，高い精度は期待できない．そこで，[1, 5] と同様に，学習データは感情の有無を示す 2 値のラベルを付け線形識別器の学習を行い，閾値処理前の識別器出力を感情強度の推定値として扱うこととした．表情識別による感情推定を例にとると，ある個人の感情（例えば,「喜び」）を代表する表情が検出できるとし，さらに，無表情からその感情（「喜び」）を抱く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推定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．本手法は，この表情の強度を，感情推定値として用いるものである．

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

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，$\boldsymbol{h}$ はバイアス値を示す．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．

2.2　感情推定のための学習・評価用データセット

音声情報による感情推定器の学習用データセットとして，The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[7] を使用した．RAVDESS は，24 人の話者 (12 人が男性，12 人がによる怒り，嫌悪，恐怖，喜び（笑顔），悲しみ，驚きの 6 種の感情に，中立 (neutral) と平静 (calm) の 2 種を加えた 8 種の感情にアノテーションされた音声付き動画に分類されており，2 種の文章（”Kids are talking by the door”と”Dogs are sitting by the door”）の読み上げを 2 回繰り返し，中立以外の感情では 2 種の強度（中立は強度は 1 種のみ）の一人当たり 60 本の動画が含まれ，総計 1440 本の動画が含まれている（図 1)．これに加え，同様の構成で 2 種の文章を，読み上げでなく歌ったデータセットも含まれているが，本稿では読み上げの部分のみを使用した．

表情識別器の学習用のデータセットとして The FaceGrabber Database and Software [8]（図 2） を使用した．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種（怒り，嫌悪，恐怖，喜び，悲しみ，驚き）の訓練を行った．この表情

図 1: RAVDESS DB 顔画像の例

図 2: FaceGrabber DB 顔画像の例

識別器を RAVDESS データセットの動画に適用し，表情識別による感情の推定値とした．

2.3 音声情報による感情推定器

RAVDESS の動画に含まれる約 2 秒の読み上げ音声に対して短時間フーリエ変換を行ったスペクトログラムを特徴量として用いた．RAVDESS 動画の音声は，48,000Hz でサンプリングされていたため，8,000Hz のローパスフィルターを通した後，16,000Hz でリサンプリングを行った．リサンプリングされたデータに対して，周波数ビン 513，時間方向ビン 766 のパワースペクトログラムを生成し，特徴量とした（図 3）．このスペクトログラムを元に，怒り，嫌悪，恐怖，喜び，悲しみ，驚きの 6 種の 2 クラス線形サポートベクトルマシン（SVM）の訓練を行ったが，汎化性能を向上するため，23 人のデータで学習し，人分を未学習データとして扱うことで，24-way Cross Validation を行いソフトマージンに対するコスト係数を決定した．

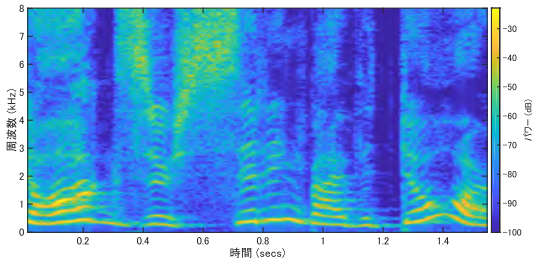


図 3: RAVDESS 動画音声のパワースペクトログラムの例

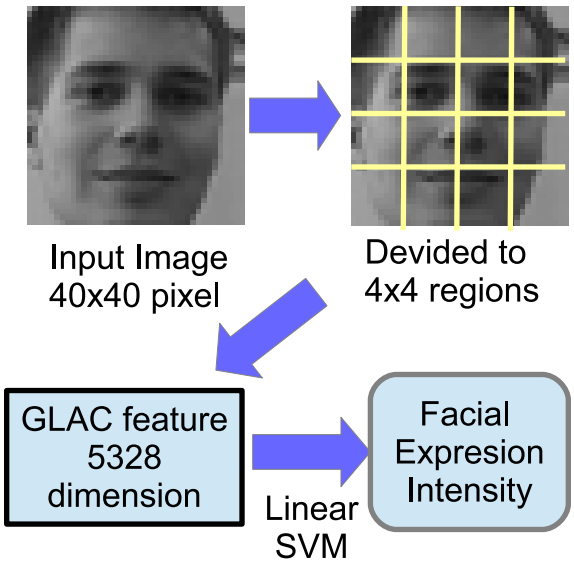


図 4: 表情識別器の構成

2.4 表情識別器による感情推定器

表情識別器は，特徴量として位置ずれや照明条件に頑健な GLAC (Gradient Local Auto-Correlation)[9] 特徴を採用した．検出された顔画像は 40×40 のグレースケール画像に変換され，4×4，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 4 ）．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 実験結果

3.1 音声情報による感情推定結果

24-way Cross Validation による平均精度を，表 1 に示す．男性，女性双方が含まれるデータセットにおい

表 1: 音声スペクトログラムによるクロスバリデーション結果

感情種別	平均精度
怒り	85.3%
嫌悪	77.5%
恐怖	81.3%
喜び	77.1%
悲しみ	74.3%
驚き	77.5%

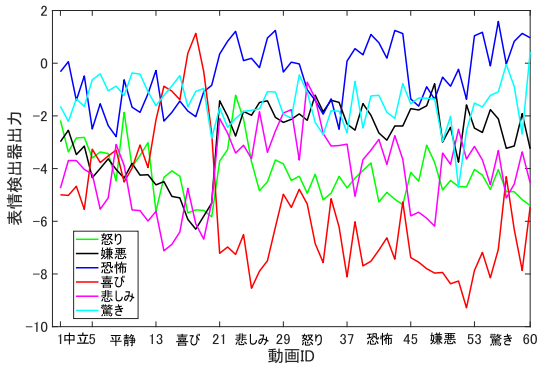


図 5: RAVDESS 話者 1 に対する表情識別器出力

て，一人分の未学習データに対して高い識別精度を達成している．これは，音声情報による感情推定器は，高い汎化性能が持つことが示唆されている．

3.2 表情識別器による感情推定結果

FaceGrabber データセットで学習した表情識別器を RAVDESS データセットに適用した結果を，図 5 に示す．動画 ID の 1 から 4 が中立（neutral），5 から 12 が平静（calm），13 から 20 が喜び，21 から 28 が悲しみ，29 から 36 が怒り，37 から 44 が恐怖，45 から 52 が嫌悪，53 から 60 が驚きにアノテーションされたもので，それぞれの線は対応する表情識別器の出力を示す．喜び識別器の出力は，喜びにアノテーションされた動画に対して高い推定値を出力しており，喜び以外の感情にアノテーションされた動画に対しては低い推定値を出力している．喜び以外の識別器は，対応する感情にアノテーションされた動画だけでなく，他の感情にアノテーションされた動画に対しても高い推定値を出力しているものが多く，必ずしも一つの感情に対する推定器とはなりえていないと考えられる．

次に，表情識別器による感情への自動アノテーションの可能性を検討した．中立と平静は基本感情に含まれていないため，怒り，嫌悪，恐怖，喜び，悲しみ，驚きの 6 種にアノテーションされた動画について検討を行った．

代表的な表情識別器，喜び識別器と怒り識別器の出力を，話者間で比較してみる．図 6 に話者 6 人分の喜び識別器の出力を示す．実際には 24 人分のデータに対して処理を行ったが，表示が煩雑になるため，6 人分だけを図示する事とした．喜び識別器出力は，喜びにアノテーションされた動画に対して高い数値を示し，それ以外の感情にアノテーションされた動画に対しては相対的に低い値を出力していることがわかる．しかし，しかし，喜び表情検出器の出力は，話者間での絶対値には差があるため，単純な閾値では喜びの感情推定とすることはできない．そこで，一人の話者に対する一つの表情識別器出力の最大値と最小値により，表情識別器の出力を正規化する．正規化は，式（2）にしたがって行った．

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

$\tilde{y}$ 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．

図 7 は，それぞれの喜び表情識別器の出力を最大最小値で正規化し，話者 6 人分を重ねてプロットしたものである．全ての話者で喜びの動画に対する出力が高くなっているが，その他の感情の動画に対する出力との差は，話者によって異なっている．そこで，適合率 (Precision) 最大となる閾値を求めることとした．適合率 Pr は式 (3) に従って計算される．

$$Pr = \frac{TP}{TP + FP} \tag{3}$$

TP は True Positive，FP は False Positive となったデータ数を示す．

また，再現率 (Recall) も同時に求めることとした．再現率 $Recall$ の定義を，式 (4) に示す．

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

FN は，False Negative となったデータ数を示す．

表 2 に，他の表情識別器に対しても，喜び識別器と同様の正規化処理と適合率最大の閾値を行った結果の適合率，再現率，閾値を示す．喜び表情識別器の喜び動画に対する適合率が 0.94 と，他の表情識別器に比べて高い値を示している．再現率は 0.44 と決して高くはないが，自動アノテーションを行える可能性の高い性能である．嫌悪表情識別器は，適合率は 0.79 と決して低い値ではないが，再現率が 0.10 と低い．その他の表情識別器は，適合率最大の閾値を求めたにもかかわら

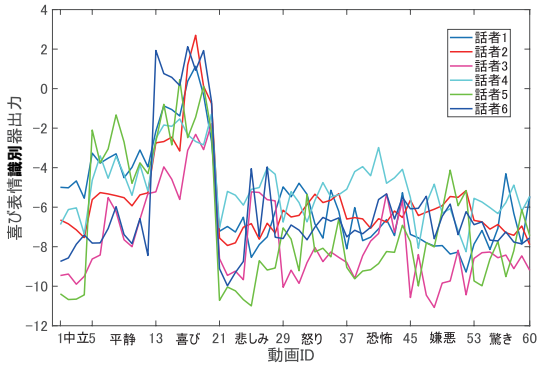


図 6: 話者 6 人に対する喜び識別器の出力

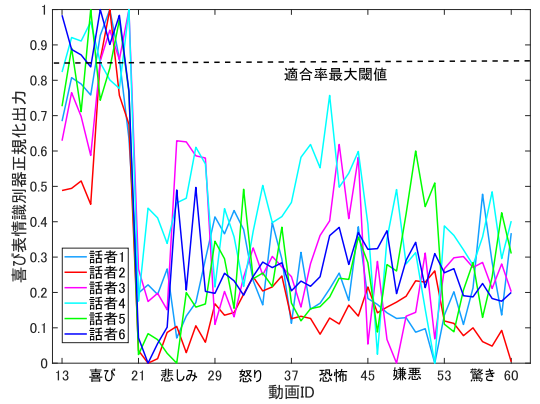


図 7: 話者 6 人の喜び表情検出器の正規化出力

ず適合率が低く，アノテーションに使用できるレベルではない．これは，[5] で確認された，喜びは他の感情に比べて特異的に判別性が高いことを再現したと考えられる．

4 結論

RAVDESS データセットを用いて，音声情報による感情推定器を構成し，汎化性の高い感情推定が可能なことを示した．また，FaceGrabber データセットで訓練した表情識別器を RAVDESS データセットの動画に適用し，表情検出による感情アノテーションの可能性を探った．その結果，喜びに関しては表情検出によるアノテーションの可能性を示すことができた．本稿での用いた音声データセットは，決められた文章を読み上げるものであり，音声の持続時間もほぼ揃っているものであり，一般的な音声情報での感情推定にはなっていない，今後は，よりバリエーションの大きな音声データセットをもちいて提案手法の有効性を検証していきたいと考えている．また，喜びの表情識別は未学習データに対しても有効なことが示されたので，これを用いて喜び表情 (笑顔) データの収集を行い，同時に音声情報に対するアノテーションの有効性を確認していきたいと考えている

表 2: 表情識別器の適合率，再現率，閾値

識別器	適合率	再現率	閾値
怒り	0.23	0.47	0.63
嫌悪	0.79	0.10	0.99
恐怖	0.58	0.18	0.94
喜び	0.94	0.44	0.85
悲しみ	0.50	0.12	0.94
驚き	0.44	0.06	0.99

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．本研究は JSPS 科研費 20H01765 の助成を受けた．

参考文献

[1] 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).

[2] 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).

[3] 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).

[4] 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).

[5] 西田健次, 山田亨, 糸山克寿, 中臺一博: 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, 人工知能学会 AI チャレンジ研究会, SIG-Challenge-055-8, pp. 41–47, (2019).

[6] Ekman, P., Davidson, R. J. (Eds.). (1994). Series in affective science.The nature of emotion: Fundamental questions. Oxford University Press.

[7] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions, *North American English. PLoS ONE* 13(5): e0196391.

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

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

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-057-10 (11/20)

複雑なニューラルネットワークを対象とした ノードプルーニングベースのモデル圧縮の検討

Model Compression of Non-uniform Neural Networks with Non-Linear Functions Based on Node Pruning

中臺 一博 [1,2*]　福本 陽典 [1]　武田 龍 [3]
Kazuhiro NAKADAI[1,2]　Yosuke FUKUMOTO[1]　Ryu TAKEDA[3]

[1] （株）ホンダ・リサーチ・インスティチュート・ジャパン
[1] Honda Research Institute Japan Co., Ltd.
[2] 東京工業大学 工学院 システム制御系　[3] 大阪大学 産業科学研究所
[2] Tokyo Institute of Technology　[3] Osaka University

Abstract: 本稿では，近年複雑化しているニューラルネットワークに対応するため，ノード枝刈りに基づいた不均一で複雑な深層学習ネットワークのモデル圧縮を扱う．ノード枝刈りによるモデル圧縮は，比較的古くから研究されているが，そのほとんどは，活性化関数としてシグモイド関数を用い，バイパス接続のない均一で単純な全結合ニューラルネットワークであることを前提としている．近年は，活性化関数として ReLU など非シグモイド関数を用いることが一般的であるため，こうした非線形活性化関数に対応するため，ノードエントロピー法を拡張したノード活性度推定法を提案する．さらに，バイパス接続など不均一なトポロジを持つネットワークに対応するため，層間ペアリング，バイパス接続の枝刈り，ネットワーク全体に対する枝刈りポリシーに関する規範を提案する．これらを組み合わせた手法を提案手法とし，ReLU，バイパス接続を有する TDNN ベースのニューラルネットワークである，音声認識システム Kaldi 用の音響モデルの圧縮に適用を試みた．結果として，音声認識性能を維持しつつ，音声認識システム全体として 31% の速度向上を達成することができた．

1　はじめに

近年，最適な数以上のパラメーターを使用すれば，深層学習のニューラルネットワークモデルを問題なく学習できることが報告されており，過剰パラメータ化 (over-parameterization) [1, 2, 3] と呼ばれている．一方，過剰パラメータ化によるモデルは規模が大きいため，組み込みデバイス，モバイルコンピューティング，車載情報 (In-Vehicle Information, IVI) システムへの応用など，実用向けには，モデル圧縮によるコンパクト化が不可欠である．また，モデル圧縮により，ネットワークの複雑さや過剰適合問題を軽減することもできる [4]．このため，因子分解 (Factorization), 知識蒸留 (Knowledge Distillation)，枝刈り (Pruning) といったモデル圧縮手法が研究されてきた [5].

因数分解はモデルのスパース性を利用しており，特異値分解 [6]，低ランク行列因子分解 [7]，ベクトル量子化 [8] を利用したモデル圧縮手法が報告されている．また，テプリッツ行列を使用した分解 [9] と，小さい行列の組合せで置換する手法 [10] も，因数分解ベースのモデル圧縮手法と捉えることができる．これらは，画像分類と音声認識 (ASR) を対象に，元のモデル性能を維持または改善しつつ，モデルサイズを大幅に圧縮している．因子分解は，パラメータ行列として表せるネットワークに適用できるので，一般に，比較的単純で均一な全結合型のニューラルネットワークに適用される．複雑なネットワークに適用する場合には，パラメータ行列に変換できる部分を抽出しての適用する必要があるため，限定的な適用となる．

知識蒸留は，主に TS (Teacher-Student) 学習を適用することでモデル圧縮を行う [11, 12]．まず，パラメータ数の多い大規模ニューラルネットワークモデルを，教師モデルとして学習する．次に，パラメータ数が少ない小規模のニューラルネットワークを生徒モデルとして，教師モデルと同等の性能となるように学習を行う．この際の教師モデルと生徒モデルの比較には，KL(Kullback-Liebler) 距離ベースの出力分布 [13] や

*（株）ホンダ・リサーチ・インスティチュート・ジャパン
〒 351-0188 埼玉県和光市本町 8-1
E-mail: nakadai@jp.honda-ri.com

シーケンスレベルの出力分布 [14] などが使用される．知識蒸留では，ラベルを平滑化するような正則化により，効果的なモデル圧縮が実現されているという報告もあるが [15]，教師が与えられた際に，どのようなトポロジーを持った生徒モデルが最適なのかといった問題について十分な議論されているわけではない．

枝刈りは，音声認識 [16, 17, 18]，画像分類 [19, 20, 21, 22, 23]，翻訳 [24] などを対象に広く研究されている．主に，枝刈りの鍵として重みパラメータやノード活性度を用いており，対象となるニューラルネットワークの層，チャネル，ノード，リンクの寄与に応じて，各々が枝刈りもしくは共有化される．性能を維持しつつ，処理速度を向上するために，ビット量子化の併用も試みられている [18, 21, 25]．武田ら [25] は，重みパラメータ，ノード活性度，ビット量子化の組合せにより，ノード枝刈り率 30% で，音声認識精度を維持しつつ，デコード処理の 5 倍高速化を実現している．これらの手法は，枝刈り後に再学習が必要ではあるが，ニューラルネットワークのサイズを大幅に縮小することができる．さらに，因子分解や TS 学習で必要なトポロジーに対する仮定が不要であるため，適用が容易であるという利点がある．しかし，これらの研究では，これまで，DNN や CNN といった単純なニューラルネットワークを対象に，活性化関数には伝統的なシグモイド関数を想定して研究が行われてきた．近年では，バイパス接続など，ネットワークトポロジーは複雑化する傾向にあり，活性化関数にも ReLU [26] など，他の非線形関数が使用されることが一般的になってきている．

つまり，従来のアプローチの問題は，モデルが複雑化しているにもかかわらず，単純で均一なニューラルネットワークを対象としており，不均一かつ複雑なネットワークに対するモデル圧縮のガイドラインが十分に議論されていないことであると言える．そこで，本稿では，バイパス接続や非シグモイド活性化関数を備えた不均一で複雑なネットワークをモデル圧縮する方法を提案する．モデル圧縮のアプローチとしては，因子分解や知識蒸留のようなトポロジーに対する仮定が少なく，局所的にも適用できるノードの枝刈りを対象とする．具体的には，非シグモイド活性化関数を扱うため，以前提案したノードエントロピーに基づくノード枝刈り法 [25] を拡張する．また，不均一なニューラルネットワークを扱うため，層間ペアリング，バイパス接続プルーニング，ネットワーク全体に対する枝刈り率設定に関する 3 つの規範を提案する．提案した方法と規範を TDNN (Time Delay Neural Network), TDNN-F (Factorized TDNN)，バイパス接続を有する，オープンソースの音声認識システム Kaldi の音響モデルに対して適用し検証を行った．

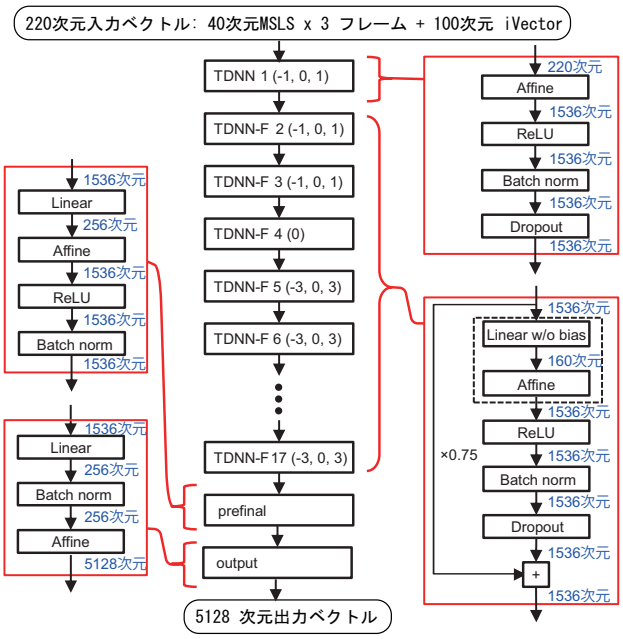


図 1: Kaldi の Nnet3-Chain 音響モデルの構成 (独自レシピ)

2 Kaldi のネットワーク構成

Kaldi の音響モデル学習には，日本語話し言葉コーパス (CSJ) レシピ [27] をベースとして，性能向上のために以下の変更を行った独自レシピを使用した．

- LibriSpeech レシピ [28] を参考に，nnet3-chain モデル [29] をサポートするための拡張．
- MFCC 特徴量の代わりに，40 次元 MSLS (Mel-Scale Log Spectrum) 特徴量 [30] を使用．これにより，入力には，3 フレーム分の MSLS 特徴量と 100 次元 iVector で構成される 220 次元のベクトルを使用．
- 言語モデル学習には，CSJ レシピで使用される SRILM [31] に代え，pocolm [32] を使用．

図 1 に，独自レシピで得られる全 19 層からなる音響モデルの構造を示す．第 1 層は TDNN で，フレームごとに 220 次元の入力ベクトルを受け取り，1,536 次元のベクトルを出力する．次の 16 層は TDNN-F で，各層は 160 次元のボトルネック層とバイパス接続を内包している．pre-final 層は バイパス接続はなく，256 次元のボトルネック層のみで構成される．出力層は，全結合ボトルネック層であり，5,128 次元ベクトルを出力する．バッチ正規化はすべての層で，ドロップアウトは最後の 2 層を除くすべての層で行われる．TDNN は，第 1 層から第 4 層目までは，連続 3 フレームを使用する．これを $(-1, 0, 1)$ と記述する．第 5 層では，現在のフレームのみが使用される．第 6 層から第 17 層目ま

では，高速な処理を実現しつつ，より長いコンテキストを扱うため，3 フレームおきに 3 フレーム分，つまり $(-3, 0, 3)$ が使用される．出力層を除き，活性化関数には，ReLU [26] が使用される．各層は，因子分解用の線形サブレイヤとアフィンサブレイヤ，ReLU 用の非線形サブレイヤ，バッチ正規化用サブレイヤ，ドロップアウト用サブレイヤなど 5 – 8 のサブレイヤで構成される．

3　提案するノード枝刈り手法

提案するノード枝刈り法，および，不均一なネットワークに枝刈りを適用するための規範を説明する．

3.1　ノード活性度の定義

l 番目の層の i 番目のノード $x_{l,i}$ に対するノードエントロピー q_e は，次式で定義できる [16, 25]．

$$q_e(l,i|D) = -\frac{N_0}{N_{0+1}} \log \frac{N_0}{N_{0+1}} - \frac{N_1}{N_{0+1}} \log \frac{N_1}{N_{0+1}}, \quad (1)$$

ここで，D はデータセット，N_0 と N_1 は，D に対して，閾値よりも低い値および高い値となったシグモイド活性化関数の出力数である．N_{0+1} は D 内のサンプル数であり，$N_0 + N_1$ と等しい．

ノードエントロピーは，ノードごとに異なる活性度を表現できるため，[16] らが提案している出力重みノルムよりもモデル圧縮性能がよいことが報告されている [25]．ただし，活性化関数としてシグモイドが前提であるため，N_0 と N_1 を判定するための閾値は，シグモイド関数の値域 (0–1) の中間値 0.5 であった．しかし，Kaldi では，値域が 0 から無限大である ReLU が活性化関数として用いられている．このため，閾値を 0.5 から 0 に近い ϵ に変更した．これは，閾値を単に変更したというよりも，ノードが活性化しているかどうかを活性化関数の出力の高低ではなく，0 かそれ以外かによって判断するように考え方を変更したと言える．この定義が有効かどうかを検証する意味で，比較のため，これに加え，頻度ベースのノード活性度 q_f と分散ベースのノード活性度 q_v という 2 つのノード活性度の定義を行った．

$$q_f(l,i|D) = \frac{N_0}{N_{0+1}}, \quad (2)$$

$$q_v(l,i|D) = \frac{1}{N_{0+1}} \sum_{t=1}^{N_{0+1}} x_t^2 - \bar{x}^2, \quad (3)$$

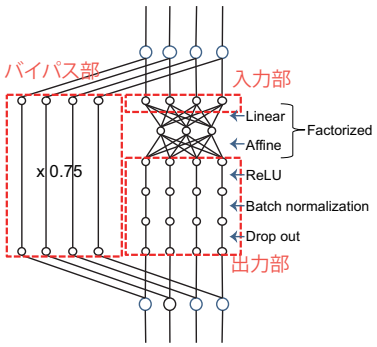


図 2: A TDNN-F layer

ここで，x_t は D における t 番目のサンプルの ReLU 出力値であり，$\bar{x}$ はデータセット内のすべてのサンプルの平均を示す．

3.2　ネットワークトポロジーの考慮

近年のニューラルネットワークのトポロジーは，TDNN，バイパス接続，attention モデルの登場により，加速度的に複雑化しており，層内でも複数のブロックで構成される不均一な構造となっている．図 2 は，Kaldi の典型的な TDNN-F 層の構造を示している．入力部，出力部，およびバイパス接続部と 3 つのブロックからなる複雑なトポロジーを有していることがわかる．

まず，出力部について考える．出力部は，ReLU サブレイヤーが含まれている．このため，ReLU の直後 (バッチ正規化の前) の値を用いて，式 (1) の ReLU バージョンに基づき，各ノードのノード活性度を求める．得られるノード活性度の小さいものから，枝刈りを行う．枝刈り率が与えられた際の具体的な枝刈り法については，後述する．

次に入力部について考える．図 2 のボトルネック部のみに注目すると，出力部とは独立して枝刈りができるように見える．しかし，実際には，入力部の入力は前の層の出力部に直接，接続されている．このため，前の層の出力部で枝刈りされたノードに直接接続されている入力部のノードを，枝刈りするものとした．これを「層間ペアリング」と呼ぶことにする．一方で，同じ層内の入力部と出力部はバイパス部を介して接続されている．この点に注目すれば，同じ層内の対応する入力部と出力部のノードを一緒に枝刈りする必要があると考えることもできる．これを「層内ペアリング」と呼ぶことにする．評価実験では，入力部と出力部を独立に枝刈りする場合を加えて，提案する層間ペアリングが良好な結果となるか比較実験を行うものとする．

バイパス部については，ResNet [33] に代表されるように，勾配消失を回避するのに役立つため，入力部と出力部の枝刈り結果に関係なく，枝刈りを行わないものとした．入力部，または出力部，あるいはその両方に

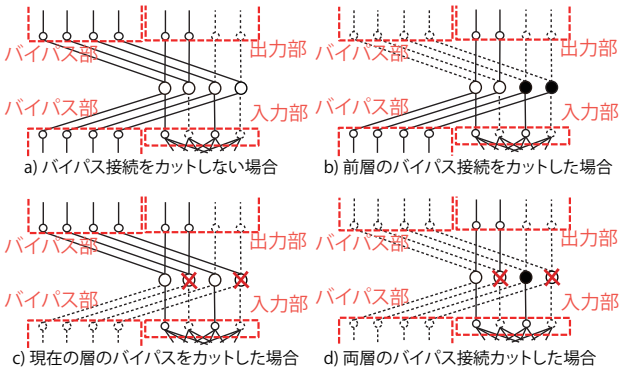
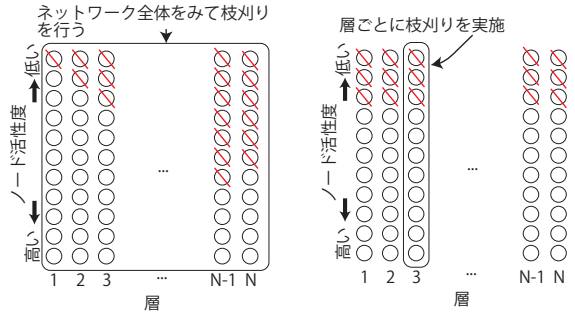


図 3: 層間接続問題．点線と点線の丸は枝刈りされたことを示す．層間の接続を保つため，上下の層の枝刈りの状況に応じて，常に 0 を出力するノード (黒丸) や終端ノード (赤×印) を設ける．



a) ネットワークベース　　b) 層ベース

図 4: 2 種類の枝刈りポリシー． 枝刈り率 30% の例:「ネットワークベース」は，ネットワーク全体を考慮して，活性度の小さいノードから順に 30% を枝刈りする．一方,「層ベース」は，層ごとに，活性度の小さいノード 30% を枝刈りする．

対応するノードが枝刈りされた場合に，バイパス接続を枝刈りの是非は，評価で比較実験を行うものとする．

枝刈り後，図 3 に示すように，入力部，出力部，バイパス部の接続を見ると，どちらか一方のノードだけが枝刈りされる接続がある．このようなオープンエンドの接続を扱うため，[33] に倣い，図 3b) および d) に示すように常にゼロ出力を行うノード (黒丸) や図 3c) および d) に示すよう終端ノード (赤×印) を設けるものとした．

以上，一つの層に対する枝刈りについて考察したが，実際のネットワークは多層に渡るため，枝刈り率[1]が与えられた際に，ネットワーク全体を考慮して枝刈りポリシーを決める必要がある．大きく 2 つのポリシーが考えられる．一つは，図 4a) に示すような「ネットワークベース枝刈り」ポリシーである．これは，層ごとの枝刈り率は一定でなくてよく，ネットワーク全体で平均して目的の枝刈り率になっていればよいとするポリシーである．実際，強化学習の結果として，枝刈り率は層ごと変えた方がよいとする報告もある [34]．また，以下の 3 点が冒頭に述べた過剰パラメータ化の知見として報告されている [1, 2, 3]．

1. 学習後，複数の中間層を初期化しても性能は維持される．
2. 入力層，または出力層が，1〜2 エポック後の対応する層と置き換えられると，性能が低下する．
3. 一般に，上位層はネットワークに対する貢献度が低くなる傾向がある．

これらの知見は，層ごとに異なる枝刈り率を許す，つまりネットワークベース枝刈りポリシーを支持しているといえる．

一方で，我々は，これに反して，図 4b) に示すように，すべての層で同じ枝刈り率となる必要がある「層ベース枝刈り」ポリシーを用いることを提案する．この背景となる考え方は，以下の通りである．

1. 上位層は下位層よりも出力層に近いため，最終出力により寄与するべきである．
2. ネットワークベース枝刈り率を適用するには，異なる層のノード間の活性度の違いを公正に評価する必要があるが，これは困難であるため，層ベースの枝刈り率で十分である．
3. ネットワークベース枝刈り率を適用すると，上位層の貢献度が低くなるため，積極的に上位層のノードが枝刈りされる．モデルの圧縮率を上げるには，枝刈り率を高くする必要があるが，この場合，上位層のノード数が少なくなり，モデルを十分表現できなくなる可能性がある．

これらの枝刈りポリシーについても評価実験で比較実験を行う．

表 1: **C1** – **C5** の実験条件．提案手法は太字

実験	ノード活性度	枝刈り率	入出力ペアリング	バイパス枝刈り	枝刈りポリシー
C1	**エントロピー**	0-70%	出力部のみ	**なし**	**層ベース**
	頻度				
	分散				
	ランダム				
C2	**エントロピー**	50%	**層間**	**なし**	**層ベース**
			層内		
			独立		
			出力部のみ		
C3	ランダム	0-70%	**層間**	**なし**	**層ベース**
				あり	
C4	**エントロピー**	0-70%	出力部のみ	**なし**	**層ベース**
					ネットワークベース
C5	**エントロピー**	50%	**層間**	**なし**	**層ベース**
	N/A	0%	N/A	**なし**	N/A

[1]本稿では，枝刈り率は，出力部内のノード総数と枝刈りされたノード数の比として定義するものとする．

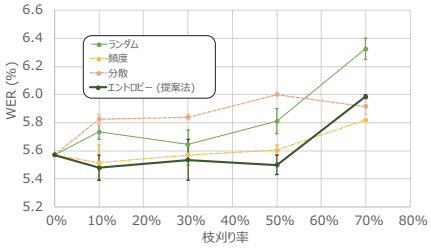


図 5: **C1**:ノード活性度比較結果 (JNAS)

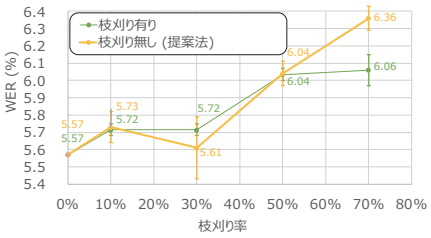


図 7: **C3**: バイパス部枝刈り比較結果 (JNAS)

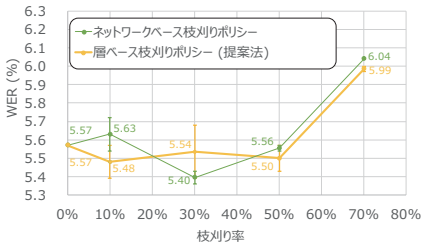


図 9: **C4**:枝刈りポリシー比較結果 (JNAS)

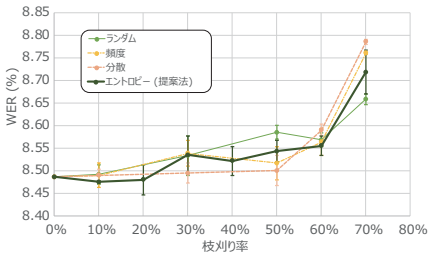


図 6: **C1**:ノード活性度比較結果 (CSJ)

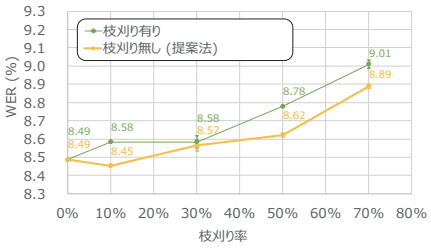


図 8: **C3**: バイパス部枝刈り比較結果 (CSJ)

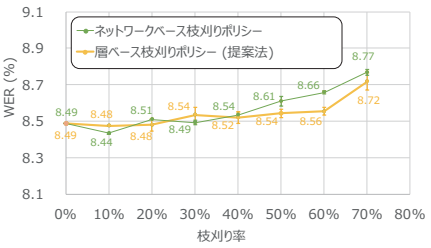


図 10: **C4**:枝刈りポリシー比較結果 (CSJ)

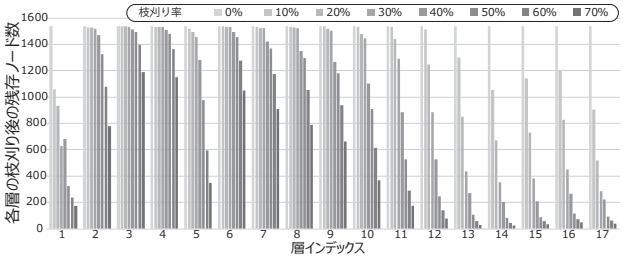


図 11: ネットワークベース枝刈りポリシー適用後の各層のノード数．層ベース枝刈りポリシー適用後は，各層のノード数はすべて同数．

表 2: **C2**:入出力間ペアリング比較結果．枝刈り率 50%．

コーパス	出力部のみ (上限値)	独立	層内	層間 (提案法)
JNAS	5.50	5.72	5.88	**5.70**
CSJ	8.54	8.69	8.61	**8.56**

表 3: 音声認識速度と WER．提案枝刈り法は，ノードエントロピーベースのノード活性度推定，層間ペアリング，バイパス接続の枝刈りなし，層ベース枝刈りポリシーの組合せ．

	音声認識速度 (s)	WER (%)
枝刈りなし	3,036	8.49
提案枝刈り法	2,388 (31% up)	8.57

4 評価

提案手法の有効性を検証するため，**C1** – **C5** の 5 種類の実験を行った．評価指標には，単語誤り率 (WER) を使用した．**C5** では，併せて音声認識システムの速度の計測も行った．表 1 に詳細な実験条件を示す．

C1: ノードエントロピーベースのノード活性度の検証．「ノードエントロピー (提案法)」,「頻度ベース」,「分散ベース」,「ランダム」の 4 種類のノード活性度を比較した．他の条件は以下の通り．枝刈り率は 0%，10%，30%，50%，60%，70%．枝刈りは出力部に対してのみ実施．バイパス部と入力部は，枝刈りなし．層ベース枝刈りポリシーを使用．

C2: 入出力層のペアリング基準の検証．「層間ペアリング (提案法)」と「層内ペアリング」に加えて，「独立」と「出力部のみ (オラクル)」の 2 つも比較に加えた．「独立」は，入力部のノードを出力部と無関係にランダムに枝刈りすることを意味する．「出力部のみ」は，入力部の枝刈りを行わないため，**C1** と同じ条件となり，原理上の性能上限値と言える． なお，**C1** の結果から，枝刈り率 50%までは同等の WER が得られることが判明したため，枝刈り率は 50% に固定した (層ベース枝刈りポリシー)．

C3: バイパス部の枝刈り是非の検証．バイパス接続の枝刈りを行う場合と行わない場合を比較した．層ベース枝刈りポリシーを用いて，枝刈り率 0 –

70% を検証．ノード活性度の定義とは無関係にこの要因を調査するため，出力部の枝刈りにおけるノード活性度には **C1** のランダム条件を用いた．また，入出力間ペアリングには層間ペアリングを用いた．

C4: 枝刈りポリシーの検証．「層ベース (提案法)」と「ネットワークベース」の枝刈りポリシーを比較した．枝刈り率は 0%–70%に変更した．また，ノードエントロピーベースの枝刈りと層間ペアリングを使用した．

C5: 上記の 4 提案手法，つまり，エントロピーベースのノード活性度，層間ペアリング，バイパス接続枝刈りなし，層ベース枝刈りポリシーを組み合わせた提案手法による音声認識速度向上の検証．枝刈り率を 50% に設定し，枝刈りなしの場合と比較を行った．

4.1 実験条件

学習とテスト用のデータセットには，JNAS (新聞記事読み上げ音声コーパス) と CSJ (日本語話し言葉コーパス) の 2 種類のサイズの異なる日本語コーパスを用いた．JNAS [35] は，60 時間の学習用音声データセットと，男女それぞれ 23 名，50 文からなるテストデータセットからなる小規模のコーパスである．CSJ [36] の学習用音声データセットは，JNAS の 10 倍強の 660 時間の音声データからなっている．また，それぞれ約 1,300 の発話からなる 3 種類のテスト用データセット (eval1–3) を含んでいる．テスト用データセットの総発話数は，3,949 で，計 18,376 秒である．式 (1)–(3) でノード活性度を求めるためのデータセット D は，すべての実験で共通の 300 発話 (19 分) を使用した．式 (1) および式 (2) の，ϵ は 0.001 に設定した．また，いずれの実験でもノード枝刈りを行った後，1 エポック分の再学習を行った． 実験は同じ条件下で 2 回行い，平均 WER を計算した．CSJ では，3 つのテストデータセットの平均 WER として算出した．

4.2 結果

図 5 と図 6 に，JNAS と CSJ に対する **C1** の結果を示す．図の横軸と縦軸は，それぞれ枝刈り率と WER を示す．JNAS では，ノードエントロピーベースのノード活性度が最もよい性能を示した．また，頻度もランダムを上回り，有効であることがわかる．CSJ では，ノードエントロピーは，枝刈り率が 0% から 60% まではランダムに上回り，枝刈り率変化に対して最も安定した性能を示したが，3 つの手法に明確な違いは認められなかった．全体としては，提案するノードエントロピーが最もよい性能を示したといえる．これは，ノード活性度を求める際に，ReLU の閾値として *epsilon* を使用するという考え方が正しいことを指示するといえる．提案したノードエントロピーベースのノード活性度が，他よりも基本的に良好な結果となったという事実は，1）分散ベースの方法は，ガウス分布を想定しているのに対し，ReLU の出力は非線形かつ非対称な分布を持っているため，2）頻度ベースの方法（式 (2)）は，式 (1) の右辺の最初の項のみを考慮しており，第 2 項を考慮していないため，ということを考慮することによって説明できよう．

表 2 に，枝刈り率 50% 時の JNAS と CSJ に対する **C2** の結果を示す．「出力のみ」は **C1** と同じ条件であるため，JNAS と CSJ の結果は，図 5 と図 6 の 枝刈り率 50% のエントロピーベースの結果と同じである．「出力のみ」は入力部の枝刈りを行わないため，原理的な上限値である．入力部の 3 つの枝刈り条件の中では，提案手法である層間ペアリングは「出力のみ」と同等の性能となり，最もよい性能を示した．これは，対象としている層の入力部が前の層の出力部と直接接続されているのに対し，同じ層の出力部とはボトルネック層を通じて間接的に接続されていることを考慮すると，直感的に理解できよう．

図 7 と図 8 に，**C3** の JNAS と CSJ に対する結果を示す．横軸と縦軸は枝刈り率と WER を表す．JNAS では，いずれも同等の性能となったが，CSJ では，入力部や出力部のノードが枝刈りされていた場合でも，対応するバイパス接続は枝刈りせず残したほうが良好な性能が得られた．これらの結果を考慮すると，バイパス接続は枝刈りしない方がよいと考えられる．これは，バイパス接続がネットワーク内の他の部分と比較して，音声認識性能に大きく貢献していることを示唆している．

図 9 と図 10 に，**C4** の JNAS と CSJ に対する結果を示す．横軸と縦軸は，それぞれ枝刈り率と WER を表す．枝刈り率が 40%未満の場合，JNAS と CSJ 共に，枝刈りポリシーの違いで性能の差は見られなかった．しかし，枝刈り率が 50% を超えると，提案する層ベース枝刈りポリシーが優位となった．これを分析するために，図 11 に，ネットワークベース枝刈りポリシーを適用した際に各層で残ったノード数を示す．上位層になるほど枝刈り率が高くなっていること，そのために上位層のノード数が少なくなっていることがわかる．つまり，枝刈り率が高いと，上位層のノード数が少なくなりすぎ，性能が劣化している可能性があることを示している．一方，層ベース枝刈りポリシーを適用した場合は，残ったノード数はどの層でも同数に保たれるため，枝刈り率が高くても性能が安定していると考えられる．

表 3 に，**C5** の結果を示す．テストデータには，18,376 秒の音声データからなる CSJ テストデータセット (eval1-

3) を使用した．音声認識の処理速度は，音響モデルの尤度計算部だけではなく，音声認識システム全体の処理時間として，Intel(R) Xeon(R) E5-2697A v4 (2.6 GHz) の 1 コアを用いて計測した．枝刈りを行わない場合，処理時間は 3,036 秒で，平均 WER は 8.49% であった．提案する枝刈り手法では，枝刈り率 50% で，それぞれ 2,388 秒と 8.57% であった．つまり，音声認識性能を維持しながら，31% の速度向上が達成することができた．

C1–**C5** の結果から，エントロピーベースのノード活性度推定，層間ペアリング，バイパス部の枝刈りをしない，層ベース枝刈りポリシーの組み合わせである提案法が最もよい性能となることを示すことができた．**C1**，**C3**，および **C4** の結果から 2 つのサイズが異なるコーパスである JNAS および CSJ ともに，枝刈り率は 50% が適切であることが得られ，これによって，**C5** で 31% の速度向上を達成した．WER と音声認識の処理速度のバランスは，アプリケーションによって変わる可能性がある．提案法は，JNAS のようにモデル圧縮と過学習しやすい小規模なコーパスだけでなく，CSJ のように規模の大きいコーパスでも効果的であった．これは，ネットワークトポロジを考慮して枝刈りを行うことが重要であることを示している．また，過剰パラメータ化 [1, 2, 3] の知見から得られる予測と異なり，層ベース枝刈りポリシーが良い結果を示した．これは，証明には，さらなる研究が必要なものの，3.2 節で述べた背景となるアイデアが正しいことを示唆している．

5　おわりに

本稿では，複雑で不均一なニューラルネットワークのモデル圧縮手法を扱った．このために，ノードエントロピーベースのノード活性度推定法，層間ペアリング，バイパス接続の枝刈りをしない，層ベース枝刈りポリシーという 4 つの手法と規範からなるノード枝刈り方法を提案した．提案法を Kaldi の音響モデルに適用し，その有効性を明らかにした． より大きなコーパス，多言語での有効性検証，ニューラルアーキテクチャ探索を使用した自動パラメータ推定，アテンションモデルなどより複雑なネットワークへの適用が今後の課題である．

謝辞

HRI-JP の 尾西ダニーロ，瀧ケ平 雅行，住田 直亮，中塚 雅樹の各氏に感謝する．

参考文献

[1] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” *Proc. of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[2] C. Zhang, S. Bengio, and Y. Singer, “Are all layers created equal?” *CoRR*, vol. abs/1902.01996, 2019. [Online]. Available: http://arxiv.org/abs/1902.01996

[3] C. Zhang, S. Bengio, M. Hardt, M. C. Mozer, and Y. Singer, “Identity crisis: Memorization and generalization under extreme overparameterization,” in *Proc. of 8th International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=B1l6y0VFPr

[4] Y. Le Cun, J. S. Denker, and S. A. Solla, “Optimal brain damage,” in *Proc. of the 2nd International Conference on Neural Information Processing Systems (NIPS)*. Cambridge, MA, USA: MIT Press, 1989, p. 598–605.

[5] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, “A survey of model compression and acceleration for deep neural networks,” *CoRR*, vol. abs/1710.09282, 2017. [Online]. Available: http://arxiv.org/abs/1710.09282

[6] J. Xue, J. Li, and Y. Gong, “Restructuring of deep neural network acoustic models with singular value decomposition,” in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, 2013, pp. 2365–2369.

[7] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, “Low-rank matrix factorization for deep neural network training with high-dimensional output targets,” in *Proc. of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6655–6659.

[8] Y. Gong, L. Liu, M. Yang, and L. D. Bourdev, “Compressing deep convolutional networks using vector quantization,” *CoRR*, vol. abs/1412.6115, 2014. [Online]. Available: http://arxiv.org/abs/1412.6115

[9] V. Sindhwani, T. Sainath, and S. Kumar, “Structured transforms for small-footprint deep learning,” in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 3088–3096.

[10] M. Pöllot, R. Zhang, and A. Kaup, “An efficient alternative to network pruning through ensemble learning,” in *Proc. of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4022–4026.

[11] J. Ba and R. Caruana, “Do deep nets really need to be deep?” in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, 2014, pp. 2654–2662. [Online]. Available: http://papers.nips.cc/paper/5484-do-deep-nets-really-need-to-be-deep

[12] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *NIPS Deep Learning and Representation Learning Workshop*, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[13] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, 2014.

[14] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016. [Online]. Available: http://dx.doi.org/10.18653/v1/D16-1139

[15] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 3903–3911.

[16] T. He, Y. Fan, Y. Qian, T. Tan, and K. Yu, "Reshaping deep neural network for fast decoding by node-pruning," in *Proc. of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 245–249.

[17] T. Ochiai, S. Matsuda, H. Watanabe, and S. Katagiri, "Automatic node selection for deep neural networks using group lasso regularization," in *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5485–5489.

[18] R. Takeda, K. Nakadai, and K. Komatani, "Acoustic model training based on node-wise weight boundary model for fast and small-footprint deep neural networks," *Computer Speech & Language*, vol. 46, pp. 461 – 480, 2017.

[19] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 1135–1143.

[20] W. Chen, J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, "Compressing convolutional neural networks," *CoRR*, vol. abs/1506.04449, 2015. [Online]. Available: http://arxiv.org/abs/1506.04449

[21] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," in *Proc. of 4th International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1510.00149

[22] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," in *Proc. of 5th International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=rJqFGTslg

[23] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *Proc. of 5th International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=SJGCiw5gl

[24] A. See, M.-T. Luong, and C. D. Manning, "Compression of neural machine translation models via pruning," in *Proc. of The 20th SIGNLL Conference on Computational Natural Language Learning*. ACL, 2016. [Online]. Available: http://dx.doi.org/10.18653/v1/K16-1029

[25] R. Takeda, K. Nakadai, and K. Komatani, "Node pruning based on entropy of weights and node activity for small-footprint acoustic model based on deep neural networks," in *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, 2017, pp. 1636–1640.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of the 27th International Conference on Machine Learning (ICML'10)*, 2010, p. 807–814.

[27] "Kaldi CSJ recipe (confirmed on Aug. 14, 2020)," https://github.com/kaldi-asr/kaldi/tree/master/egs/csj/s5/local/chain/tuning/run_tdnn_1a.sh.

[28] "Kaldi liblispeech recipe (confirmed on Aug. 14, 2020)," https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech/s5/local/chain/tuning.

[29] "Kaldi nnet3-chain model (confirmed on Aug. 14, 2020)," https://kaldi-asr.org/doc/chain.html.

[30] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features," in *Proc. of 148th Acoustical Society of America Meetings*, no. 1aSC7, 2004.

[31] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "Srilm at sixteen: Update and outlook," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE SPS, 2011.

[32] "Pocolm web (confirmed on Aug. 14, 2020)," https://github.com/danpovey/pocolm.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[34] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, "Amc: Automl for model compression and acceleration on mobile devices," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018, pp. 784–800.

[35] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan*, vol. 20, no. 3, pp. 199–206, 1999.

[36] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. of the Second International Conference on Language Resources and Evaluation (LREC'00)*. ELRA, 2000. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2000/pdf/262.pdf

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-057-11 (11/20)

住居内環境でのLiDAR・マイクアレイ統合による移動音源の追跡

Tracking a Moving Sound Source in Indoor-Environment with LiDAR and Microphone-Array

伊福和己　　公文誠 *
Kazuki Ifuku　　Makoto Kumon

熊本大学
Kumamoto University

Abstract: 本論文ではLiDARによるポイントクラウドとマイクロホンアレイからの音到来方向情報を統合し，環境中の移動音源位置推定を考える．住居等の場合, 環境の多くを占める建物部分が平面でモデル化できることを利用して，形状が未知な環境中の物体を取り出し, 音源を認識する方法を提案する．分離された音源候補については，確率的移動モデルと観測の対応づけを利用して，移動音源の追跡も実現する．実環境で行った提案法の評価実験の結果も合わせて示す.

1　はじめに

音環境認識において，音源の位置を正確に推定することは重要で，例えばロボットが対話相手を検出する場合や，災害現場で要救助者を発見しようとする時，また侵入者検知など，この技術が必要となる状況は多々存在する．このような例では，対象以外にも音源が存在することが一般的で，さらに音源自体の発する音が間欠的なことから，音情報のみで対象を継続的・安定して検知，追跡することは容易ではない．特に住環境のような複雑な環境では，壁面での反射や回折など環境との相互作用によって正確な音到来方向の推定 [1, 2] が難しく，マイクロホン（以下マイク）やマイクアレイに加えてカメラ等の異なるセンサを相補的に用いることが効果的である．

カメラ動画像とマイクアレイ情報では音到来方向が得られるため，これらの情報が空間の同一方向を指すことを利用して統合することが考えられる [3]．この場合，対象の奥行き情報は得られないため，例えば複数の音源が交差するような場合，その前後を判断するには複数のカメラや深度情報を用いる [4]，複数のマイクアレイを用いる [5]，カメラとマイクアレイを異なる位置に配置する [6] など，適当な方法で音源位置を推定する必要がある．

対象の奥行情報を得るために良く用いられるセンサにLaser Imaging Detection and Ranging (以下 LiDAR) があり，多数の計測点からなるポイントクラウドと呼ばれる点群 (以下点群) で周辺環境を表現する．特に移動するLiDARの観測情報とオドメトリ情報にSimultaneous Localization and Mapping (以下 SLAM)[7, 8] 手法を援用した環境地図作成は良く研究されており，自動運転技術として実用化の取り組みがなされる段階にある．音環境地図の作成にも展開があり，田邉ら [9] がLiDARで得られた環境情報である点群と，高精度なマイクアレイによる音到来方向から音源位置を推測する方法を提案している．また，Even ら [10] は LiDAR に基づくSLAMで得られた環境地図に対して，遅延和ビームフォーマで推定した音到来方向情報を統合して音源情報を表す方法を示した．これらはいずれも静的音源を対象としたもので，点群（あるいはこれを変換したボクセル）と音源の対応を絶対位置を用いて対応づけしている．一方，SLAMで得られる環境地図は静的なものに限られるため，移動物体の検出は別に対応する必要がある．自動運転や屋内作業のようにコンテキストが明確な場合，検出される対象について車，歩行者や自転車などを仮定することは現実的で，点群の中からこれらの物体をセグメンテーションする手法が提案されている (例えば [11, 12])．あるいは移動物体が音を発するという仮定に立脚して，音源付近を取り除いて静的環境の三次元再構成を行った後，移動物体を改めて環境情報に加えるアプローチも提案されている [13]．しかし，日常空間での音源を考える場合，対象は多岐に渡るため，これらを事前に学習しておくことは難しく，また移動物体が常に音を発しているとは限らないので，静的な背景情報から対象物体の候補を分離するのは簡単ではない．特に LiDAR の点群は画像情報に

*連絡先： 熊本大学
〒 860-8555 熊本市中央区黒髪２－３９－１
E-mail: *kumon@gpo.kumamoto-u.ac.jp*

比べて疎な情報で，対象への遠近によっても疎密が変わることなどから，対象の詳細な形状が得にくいことも考慮する必要がある．

本研究でも環境中に移動音源が存在する場合に LiDAR とマイクアレイで認識することを考えるが，ここでは対象音源ではなく周囲環境を住環境のような一定の性質を仮定し，この特性を用いて点群中から対象を分離することに取り組む．

本論文の構成は次の通りである．次節でマイクアレイによる音源認識についてを概略し，第 3 節で想定する環境についての条件と，LiDAR での点群情報から対象の候補を抽出する方法，さらに音源の追跡手法について説明する．その後，提案法の妥当性について実験での検証を行い（第 4 節），最後に第 5 節でまとめる．

2 マイクアレイと LiDAR による音環境地図

マイクアレイでの音源方向推定情報は，特に室内のような反射など不確かさのある環境では不正確となるため，複数の観測を統合して推定することが重要である．ここでは，先行研究（Even ら [10]）を元にマイクアレイと LiDAR を用いて環境中の音源位置を推定し環境地図を構成する方法を概説する．

2.1 音源位置推定

Multiple signal classification(MUSIC)[14] 法などマイクアレイ処理によって音源方向を推定する手法が提案されているが，これらはマイクアレイから見た相対的な方向情報が得られるものの，その奥行き情報を得られない．

本研究では LiDAR を用いて環境地図の形状情報が得られるため，環境地図に音源方向を投射するアプローチを採るものとする．また音源方向の推定には，耐雑音性が高く，複数音源からの到来方向にも対応可能とされる MUSIC 法を用いることとする．具体的には MUSIC 法によって得られた MUSIC スペクトルが十分に大きな値を示した時，これに対応する方向を音到来方向とする方法を考える．

環境地図情報を $\mathcal{M}$ と表し，推定された音源方向 θ が得られたとき，マイクアレイから音源方向に向けた線分が環境地図との交点 x を求めることが出来る．

$$x = x(\theta, \mathcal{M}) \tag{1}$$

ただし，θ には一定の不確かさがあるため，推定された音源位置 x にも不確かさがあるため，この不確かさを考慮する必要があり，x が音源であるかどうかは確率的に表現されることになる．

2.2 環境地図中の音源確率

時刻 t において推定された 音源方向 $\theta(t)$ について MUSIC スペクトルが $P_M(\theta(t))$ とする．この観測に対して, (1) で与えられる x に音源があるという信念を表す尤度 $L(x(\theta(t)), \mathcal{M})$ を以下で与える．なお，以降，誤解のない範囲で $\mathcal{M}$ は省略する．

$$L(x(\theta(t)), P_M(\theta(t))) = p_{min} + \frac{p_{max} - p_{min}}{1 + e^{-\frac{P_M(\theta(t)) - T_a}{\alpha}}} \tag{2}$$

ここで p_{min}, p_{max} は尤度の範囲を定めるパラメータであり, α, T_a は尤度分布を定める定数である．

今, $\mathcal{M}$ の点 A が音源であるという事象を E_A, A が (1) で求まる $x(\theta(t))$ と一致する，あるいは近傍にあるという事象を $S_{A,\theta(t)}$ と表す．$S_{A,\theta(t)}$ の下で A が音源である確率 $P(E_A|S_{A,\theta(t)})$ は Bayes の定理と，$P(\overline{E_A}) = 1 - P(E_A)$ の関係より，

$$P(E_A|S_{A,\theta(t)}) = \left[\frac{1 + (1 - P(S_{A,\theta(t)}|E_A))(1 - P(E_A))}{P(S_{A,\theta(t)}|E_A)P(E_A)}\right]^{-1} \tag{3}$$

となる．対数オッズ関数

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{4}$$

を用いれば (3) は以下のように変形できる．

$$\mathrm{logit}(P(E_A|S_{A,\theta(t)})) = \mathrm{logit}(P(S_{A,\theta(t)}|E_A)) + \mathrm{logit}(P(E_A)) \tag{5}$$

また (5) を繰り返し用いることで，このような観測を N 回得た時 ($S_{A,\theta_1:\theta_N}$ などと書く），

$$\mathrm{logit}(P(E_A|S_{A,\theta_1:\theta_N})) = \mathrm{logit}(P(E_A|S_{A,\theta_1:\theta_{N-1}})) + \mathrm{logit}(P(S_{A,\theta_N}|E_A)) \tag{6}$$

のように漸進的に計算することが出来る．実際には観測を得た際の確率 $P(S_{A,\theta})$ を (2) の尤度に置き換えて，

$$\mathrm{logit}(P(E_A|S_{A,\theta_1:\theta_N})) = \mathrm{logit}(P(E_A|S_{A,\theta_1:\theta_{N-1}})) + \mathrm{logit}(L(x(\theta_N), P_M(\theta_N))) \tag{7}$$

とする．これを推定された音源方向情報全てについて計算することで音環境地図を構成する．なお，初期確率は適宜与えるが，$P(E_x) = \frac{1}{2}$ $(\forall x \in \mathcal{M})$ とすることが多い．

3 住居内環境での音源検出

3.1 住居内環境の音源

本研究で考える住居内環境についてまず説明する．

住居は人の生活に供する空間であるが，改めてその構造を考えると，外界から守るために壁面・屋根等の境界を与える建物部分があり，その内部に机や戸棚等の什器などの物体と人の移動・滞在できる空間がある．Manhattan 仮説 [15] が指摘するように人工物の環境構造には平面や平行，直交などの特性があるため，本研究でも住居の特に建物部分については平面によって構成されると考える．実際，住居内で観測をする場合，床，天井，壁面によって全方向を覆われる状況になるため，LiDAR の観測する点群の多くの部分は平面でモデル化できる特徴がある．また，音源となりうる家具等は壁際に存在することが多く，例えば壁掛け時計などを考えれば，壁と物体を点群中から適切に区別することは工夫が必要である．一方，住居内の音源は数個程度が孤立していて自由空間部分を移動するものが含まれる可能性があり，建物そのものが音を発することは少ない．移動音源が壁面近くに滞在する可能性は注意が必要である．

以上より，本研究では LiDAR で得た対象の環境の点群には，いくつかの平面を構成する多数の点と，それ以外の要素を成す独立したクラスタが存在すると考え，これらの独立したクラスタが物体を表しているというモデルに立脚して各物体を追跡する手法を提案する (図 1)．また，この物体には静的物体と移動物体双方を含むことに注意されたい．

3.2 静的平面の検出

前述の環境に対する仮定から，壁などの建物については平面構造で面積が大きいため観測点数が多く, 観測中における点群位置の変化は微小という性質がある．このような環境を構成する点群は，大きなクラスタを構成するため，比較的特定が容易であることから環境を構成する点群の内, 平面状に分布する点群を検出し, 近似平面に分類された点群を建物環境として扱う．

近似平面の検出には Random Sample Consensus[16] (RANSAC) を用いて，クラスタサイズの逆数と平面モデルとの二乗距離和から成る評価関数に基づいて法線と原点からの距離に対応する平面パラメータを推定する．さらに，平面として検出された点群を取り除いた後，残る点群に対して繰り返し同手法を適用して環境の平面点群を複数検出する．この手続きは平面として検出されたクラスタサイズが検出を実行した入力点数の一定割合未満となるまで続ける．以下 N_p 個の平面が検出されたと考え，これらの平面パラメータを $\phi_p = \{\phi_{p,1}, \cdots, \phi_{p,N_p}\}$ などと書く．

また，対象とする平面は静的と考えているため，パラメータ ϕ_p は観測時刻によらず一定となる筈である．時刻列 $t_1, \cdots, t_T$ におけるスキャンデータを前述の方法

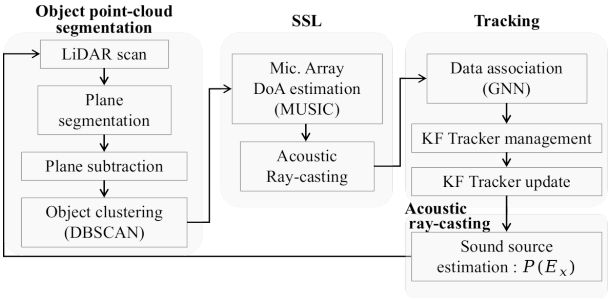


図 1: Flow of the proposed method

で処理して得たパラメータの集合 $\{\phi_p(t_1), \cdots, \phi_p(t_T)\}$ をパラメータ空間で Density Based Spatial Clustering of Applications with Noise(DBSCAN)[17] によって クラスタリングし，一定以上のサイズを持つクラスタの重心を静的平面として推定し，これらに所属する点群によって静的環境情報 $\mathcal{M}_S$ を定義する．この平面群に対応する点群を各時刻での LiDAR からの点群から取り除き，環境中の静的物体と移動物体のクラスタを示す点群を得る．

3.3 環境内物体の検出・追跡

前小節の方法で得た環境中物体の点群を物体毎に認識するため，点群をクラスタに分類しこれらのクラスタの動きを推定する．なお，以下では簡単のため 2 次元平面内の記述とする．

3.3.1 個別物体の識別

環境中の多くを占める建物由来の点群を前小節の方法で推定し取り除くことで比較的少数から成る物体に対する点群（以下物体点群と呼ぶ）が得られる．この物体点群に含まれる各物体毎の点群クラスタは互いに非連結となっていると期待されることから，物体点群に対して DBSCAN[17] を適用して，各物体毎の点群クラスタ (以下物体クラスタ）を得る．このように 時刻 t において k_t 個に分類された物体クラスタを $\{C_{t,1}, \cdots, C_{t,k_t}\}$ と表すこととする．

3.3.2 個別物体の追跡

住居環境での移動は，その空間的制約から速さや動きは制限されていることから，移動物体には適当な運動モデルを想定することは現実的である．本研究ではシンプルな例として等速直線運動を導入する．

時刻 t で l 個の追跡対象が得られており，その m 番目の物体点群の中心位置を $(x_{t,m}, y_{t,m})$ とする．状態

を $\boldsymbol{x}_{t,m} = (x_{t,m}, \dot{x}_{t,m}, y_{t,m}, \dot{y}_{t,m})^T$ とし，等速直線運動の仮定の下で状態方程式は以下のように書ける．

$$\begin{aligned} \boldsymbol{x}_{t,m} &= \boldsymbol{F}\boldsymbol{x}_{t-1,m} + \boldsymbol{G}\boldsymbol{\Delta}\boldsymbol{x}_{t-1,m} \quad (8) \\ \boldsymbol{F} &= \begin{bmatrix} 1 & \tau & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tau \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{G} = \begin{bmatrix} \frac{\tau^2}{2} & 0 \\ \tau & 0 \\ 0 & \frac{\tau^2}{2} \\ 0 & \tau \end{bmatrix} \end{aligned}$$

ここで τ はステップ間のサンプリング周期で，$\boldsymbol{\Delta}\boldsymbol{x}_{t,m}$ は 平均 $\boldsymbol{\mu}_{x,t-1,m}$, 共分散 $\boldsymbol{Q}_{t-1,m}$ の正規加速度外乱とする．

一方，クラスタの重心が物体位置として観測される．時刻 t の物体点群のうち n 番目のクラスタ $C_{t,n}$ が物体 m のものとすれば，$C_{t,n}$ の重心座標 $\boldsymbol{z}_{t,n}$ によって観測方程式は以下の式で表される．

$$\begin{aligned} \boldsymbol{z}_{t,n} &= \boldsymbol{H}\boldsymbol{x}_{t,m} + \boldsymbol{\Delta}\boldsymbol{z}_{t,n} \quad (9) \\ \boldsymbol{H} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{aligned}$$

ここで $\boldsymbol{\Delta}\boldsymbol{z}_t$ は平均 $\boldsymbol{\mu}_{z,t,n}$, 共分散 $\boldsymbol{R}_{t,n}$ の正規外乱とする．

上記モデルから，以下に示すカルマンフィルタを用いて動きの予測を行う．

1. 予測ステップ

$$\begin{aligned} \bar{\boldsymbol{x}}_{t/t-1,m} &= \boldsymbol{F}\boldsymbol{x}_{t-1,m} + \boldsymbol{\mu}_{x,t-1,m} \quad (10) \\ \bar{\boldsymbol{P}}_{t/t-1,m} &= \boldsymbol{F}\boldsymbol{P}_{t-1,m}\boldsymbol{F}^T + \boldsymbol{G}\boldsymbol{Q}_{t-1,m}\boldsymbol{G}^T \quad (11) \end{aligned}$$

2. 更新ステップ

$$\begin{aligned} \boldsymbol{y}_{t,mn} &= \boldsymbol{z}_{t,n} - \left(\boldsymbol{H}\bar{\boldsymbol{x}}_{t/t-1,m} + \boldsymbol{\mu}_{z,t,n}\right) \quad (12) \\ \boldsymbol{S}_{t/t-1,m} &= \boldsymbol{H}\bar{\boldsymbol{P}}_{t/t-1,m}\boldsymbol{H}^T + \boldsymbol{R}_{t,n} \quad (13) \\ \boldsymbol{K}_{t,m} &= \bar{\boldsymbol{P}}_{t/t-1,m}\boldsymbol{H}^T\boldsymbol{S}^{-1}_{t/t-1,m} \quad (14) \\ \boldsymbol{x}_{t/t,m} &= \bar{\boldsymbol{x}}_{t-1,m} + \boldsymbol{K}_t\boldsymbol{y}_{t,m,mn} \quad (15) \\ \boldsymbol{P}_{t,m} &= \bar{\boldsymbol{P}}_{t/t-1,m} - \boldsymbol{K}_{t,m}\boldsymbol{H}\bar{\boldsymbol{P}}_{t/t-1,m} \quad (16) \end{aligned}$$

なお，(10) で得られる状態 $\bar{\boldsymbol{x}}_{t/t-1,m}$ に対して，次節で求める対応関係が得られる場合のみ更新ステップ (12)〜(16) を行なうものとする．対応する観測が得られない場合は

$$\begin{aligned} \boldsymbol{x}_{t/t,m} &= \bar{\boldsymbol{x}}_{t-1,m} \\ \boldsymbol{P}_{t,m} &= \bar{\boldsymbol{P}}_{t/t-1,m} \end{aligned}$$

とする．

3.3.3 データアソシエーション

実際の状況では，有効領域内に観測が複数の観測を得る場合や，複数の追跡位置から重複して観測される場合があり，追跡位置と観測値との対応を求めるデータアソシエーションが必要となる．本研究ではこの対応付けに演算の簡便な Global Nearest Neighbor (GNN)[18] を用いる．

観測された物体クラスタ n と推定している物体 m との対応を求めるため，物体クラスタ n の観測位置と物体 m の推定位置の間の距離にマハラノビス距離

$$d^2_{t,nm} = \boldsymbol{y}^T_{t,nm}\boldsymbol{S}^{-1}_{t/t-1,m}\boldsymbol{y}_{t,nm} \quad (17)$$

を考える．GNN 法では, まず有効領域の閾値 g と，領域外を表す十分に大きな正定数 E を定めて次のコスト関数 λ_{nm} を考える．

$$\lambda_{nm} = \begin{cases} E & d^2_{nm} > g \\ d^2_{nm} & d^2_{nm} \leq g \end{cases} \quad (18)$$

(18) を全ての観測・対象の組合せについてまとめたコスト行列 $\boldsymbol{\Lambda}$ を以下のように定める．

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2k} \\ \vdots & \vdots & & \vdots \\ \lambda_{l1} & \lambda_{l2} & \cdots & \lambda_{lk} \end{bmatrix} \quad (19)$$

物体 m に対応する観測値の番号を $a(m)$ ($m \neq m'$ に対して $a(m) \neq a(m')$) とした時，$\Sigma^l_{m=1}\lambda_{m,a(m)}$ を最小とする $a(m)$ が求める対応関係となる．この最小化には Munkres 法 [19] を用いることで効率的に計算できる．

3.3.4 追跡管理

物体の追跡プロセスの経過とともに，新しい追跡対象の出現や，それまでの追跡対象が消失することがあり，それに応じて対応するカルマンフィルタを生成・消去する管理機構が必要である [20].

(a) 新規追跡物体の生成

得られた観測が現在追跡中のいずれのカルマンフィルタの推定値からもマハラノビスの意味で十分に離れている場合，この観測を追跡する新たなカルマンフィルタが開始される．ただし，観測値が突発的な外乱によるものである可能性を想定し，追跡開始から一定回数 (N_1 と記す) 以内に次の観測が一度でも得られない場合は追跡を中止する．

(b) 追跡の終了

オクルージョンなどで追跡している物体に対する観測が一時的に得られないことがある．観測が得られなくなった場合，予測ステップ (10) (11) を用いて物体の追跡を続けるが，長時間にわたって観測が得られない場合，推定値は信用できないものとなる．そこで十分

な回数（ここでは N_2 回とする）続けて観測の得られない状況が続いた時に当該の物体についての追跡を終了する．

以上で求められた時刻 t での各物体の点群によって物体環境情報 $\mathcal{M}_{O,t}$ を与える．これから，環境情報は $\mathcal{M} = \{\mathcal{M}_S, \{\mathcal{M}_{O,t_1}, \cdots, \mathcal{M}_{O,t_T}\}\}$ と得られる．また $\mathcal{M}_{O,t}$ は速度の推定も含んでいるため閾値処理等で移動物体と静止物体を区別する．

3.4 音源情報との統合

Ray-cast 法は，指定した場所から透明な光線を放ち，光線と交わる物体から情報を取得する方法であり，ここでは簡単のため，球との交点を ray-cast 法で求め，音源位置を推定する．

時刻 t における音源方向 θ_t が得られたとする．物体クラスタ $C_{t,j} \in \mathcal{M}_{O,t}$ の点群 $\left\{\boldsymbol{c}^1_{t,j}, \cdots, \boldsymbol{c}^i_{t,j}, \cdots\right\}$ の全ての点それぞれについて，その点を中心とする半径 r の球を考え，マイクアレイから音源方向 θ_t に向けた半直線との交点の有無を求める．交点が存在すれば，物体クラスタ $C_{t,j}$ から音が到来したと判断する (図 2)．複数のクラスタと交点を持つ場合，これらいずれか，あるいはそのうちのいくつか，または全てから音が到来している可能性があるため，全てを音源候補と考えて，前述の音源確率の計算を行う．

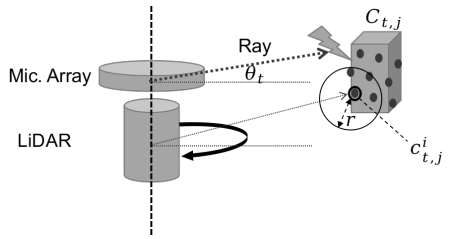


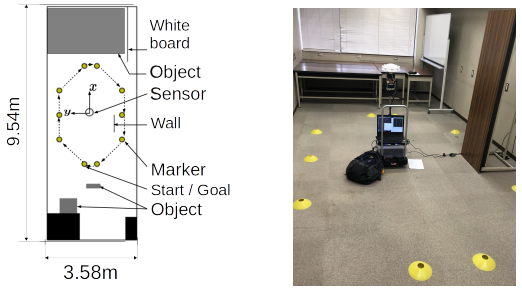
図 2: Cluster based Acoustic Ray-casting

4 検証

提案法の有効性を確認するため，居室内 (図 3) での検証実験を実施した．提案法の実装において，音源定位にはロボット聴覚用オープンソースソフトウェア HARK[21] を用い，ロボットミドルウェア ROS[22] にて LiDAR のスキャンデータと同期して収録した．また，クラスタリング等の点群処理には Open3D[23] を利用した．

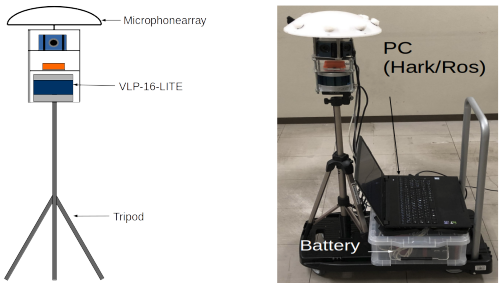
本研究で用いたマイクロホンアレイと測定装置を図 4 に示す．実験装置を構成する LiDAR は垂直視野 30 度，水平視野 360 度の範囲の点群が取得可能な，Velodyne 社製の VLP-16-LITE を使用し，円上に 8 つのマイクが配置されたマイクアレイ (System Infrontier 社) と音響処理ユニット RASP-ZX (System Infrontier 社) を用いて音信号を収録した．マイクロホンアレイと LiDAR は三脚上で軸を同一となるよう固定した．

住環境において想定した静止物体としては室内の壁面，床，天井の他，机，ホワイトボード，洗面台，PC モニタ等を考え，追跡物体のオクルードを検証するために装置中心から x 方向に-0.35m,y 方向に-1.02m 離れた地点に幅 0.71m の壁を設定した．また，環境中を移動する音源として，人が継続的に発声を行いながら装置の周りを楕円状に移動した．この移動はスタート地点からマーカーに沿って時計回りに人間の歩行速度で周回するものとした．得られたスキャン点群ならびに MUSIC スペクトルの時間発展を図 5 と 6 に示す．



(a) 実験環境模式図　**(b)** 実験環境概観

図 3: 実験環境



(a) 装置模式図　**(b)** 装置概観

図 4: 実験装置

今回のシステムでは，LiDAR のスキャン周期 $\tau = 0.1$[sec] を基準に同期して処理しており，提案法のカルマンフィルタに用いたパラメータ値は $N_1 = 5$, $N_2 = 80$, $\boldsymbol{Q} = 0.1\boldsymbol{I}$, $\boldsymbol{R} = 0.01\boldsymbol{I}$ とした．音源確率の推定では $p_{min} = 0.1$, $p_{max} = 0.9$, また予備実験から $\alpha = 40$, $T_a = 20$ とした．

まず環境から建物部分とそれ以外の住居内物体の区別の様子を図 7 に例示する．クラスタ毎に色分けしており，壁面，床，天井などの建物に対し，仮説の通り室内の人，机等が非連結のクラスタとして分けられていることが分かる．

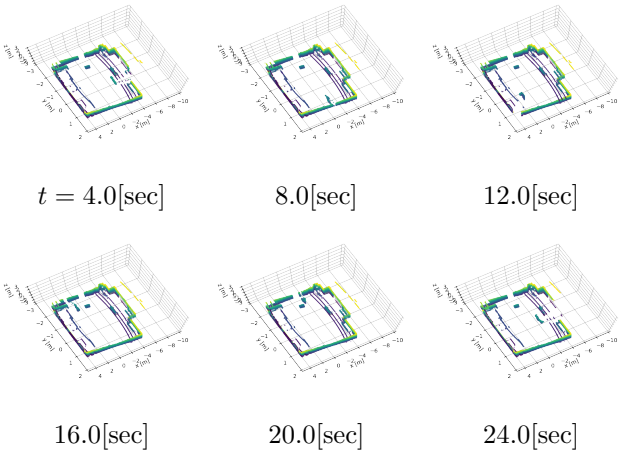


図 5: スキャンされた点群

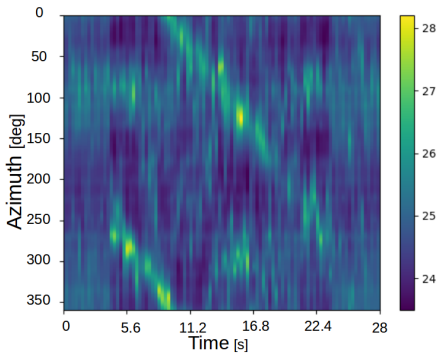


図 6: MUSIC スペクトログラム

次に各物体クラスタの追跡結果を図 8 に示す．図中黒い点は平面検出によって建物として認識された点群を成し，赤い点が物体とされた点群を示している．物体中の黒点がカルマンフィルタの平均，その周辺の等確率楕円で推定された不確かさを示しており，0.1[m/sec] 以上の速さが推定された物体については，矢印でその速度ベクトルを表している．室内中央を楕円経路に沿った移動音源を良く検出・追跡している．20[sec] 付近から図右下付近にも移動音源を推定しているが，これはオクルージョンに伴って，静止物体がいくつかの小領域に分かれたため，いくつかの小領域の重心がオクルードしている移動音源の動きに合わせて誤って推定されたものである．この推定は共分散の大きな不確かな情報で，適当な閾値処理などで区別することが可能と考えられる．

提案法はクラスタリングとカルマンフィルタの追跡によって，移動音源の点群についても一貫した音源確率の推定を行うものである．これを確認するため，各物体で推定された音源確率を図 9 に示す．提案法 (図 9(a)) では移動音源に対して高い確率を与える一方，室内に

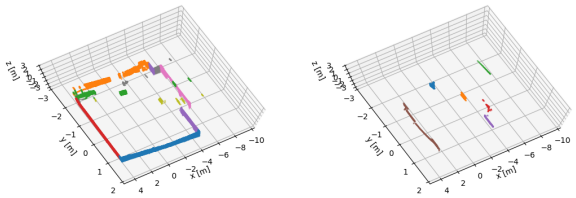


図 7: クラスタ分類結果 ($t = 20.0$[sec])

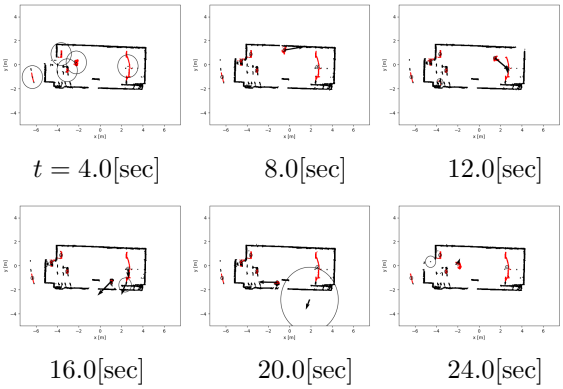
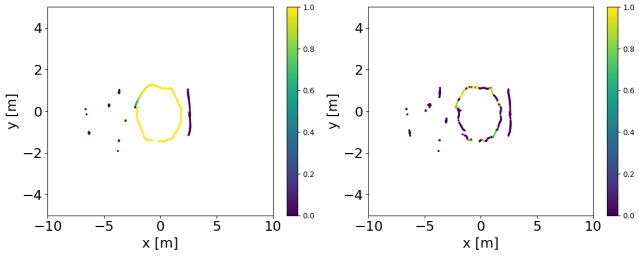


図 8: 移動音源・物体追跡

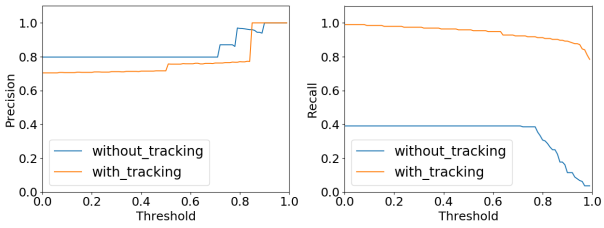
あった音を発さない物体については低い確率となった．比較のため，クラスタ追跡を行わなかった場合の結果を図 9(b) に示す．MUSIC スペクトルに強弱があったため，閾値処理によって音源の検出出来なかった区間では音源推定確率が低くなっており，また非音源との区別も難しい．このことを定量的に評価するため，音源確率を元に音源の存在判定した場合の 適合率 V_P と再現率 V_R を図 10 に示す．なお，音源確率の分布範囲が両者で異なるため，正規化して閾値処理するものとし，試行中の検出すべき延べ音源数を Num_S とし，正しく検出した音源の総数を Num_T，音源として検出した総数を Num_D とした時 適合率は $V_P = \frac{\mathrm{Num}_T}{\mathrm{Num}_D}$，再現率は $V_R = \frac{\mathrm{Num}_T}{\mathrm{Num}_S}$ で与えられる．図より適切な閾値において



(a) 提案法　　**(b)** 比較：追跡無

図 9: 推定された音源確率

追跡を行わない場合に比べ提案法の方が性能の良いこと, 特に再現率において優位であることが示された.



(a) 適合率 V_P (b) 再現率 V_R

図 10: 性能評価

5 おわりに

本研究では，住居内環境で LiDAR を用いて点群を得る時，建物部分を適当な仮定の下で取り除いて室内物体と区別し，物体と音到来方向情報を統合して音源を認識する手法を提案した．複数の物体が室内にある場合でも，対象を区別できることを実験で示した．また, 対象が移動物体であっても音源である確率を推定できることも確認し，提案法の有効性が示された．オクルージョンが発生する場合，物体クラスタの認識に影響することがあり, 今後改善が必要である.

謝辞

本研究は JSPS 科研費 19H00750 の助成をうけた.

参考文献

[1] G. Narang et al., “Auditory-aware Navigation for Mobile Robots based on Reflection-robust Sound Source Localization and Visual SLAM,” Proc. of the 2014 IEEE Int. Conf. on Sys., Man, and Cyb. (SMC2014), 2014, pp.4079-4084.

[2] K. Takami et al., “Estimation of a nonvisible field-of-view mobile target incorporating optical and acoustic sensors,” Autonom. Robot., 40(2), 2016, pp. 343-359.

[3] Y. Kokusho and M. Kumon, “Sound Source Tracking by Incorporating Target Motion Estimated by Visual Trackers,” Proc. of Int. Sym. on Sys. Integ. (SII2020), 2020, pp. 652-657.

[4] 鈴木啓他, “環境音情報と画像情報を用いた物体検出による音ラベル付きセグメントの生成” 日本ロボット学会学術講演会予稿集, 2020, 1D3-02.

[5] K. Sekiguchi et al., “Online Simultaneous Localization and Mapping of Multiple Sound Sources and Asynchronous Microphone Arrays,” Proc. of Int. Conf. on Intel. Robot. Sys. (IROS2016), 2016, pp. 1973-1979.

[6] 公文誠, 鷲崎海他, “繰り返しベイズ推定を用いた視聴覚統合による話者位置推定,” 計測自動制御学会システムインテグレーション部門講演会, 2018, 3B3-13.

[7] J.J. Leonard and H.F. Durrant-whyte, “Simultaneous Map Building and Localization for an Autonomous Mobile Robot,” Intelligent Robots and Systems (Int. Workshop at IROS91), 1991, pp. 1442–1447.

[8] 友納正裕, SLAM 入門, オーム社, 2018.

[9] 田邉亮他, “マイクロホンアレイと 3 次元 LIDAR を用いた確率的音源地図作成,”, ロボティクスメカトロニクス学術講演会, 2016, 1P1-09b4.

[10] J. Even et al., “Probabilistic 3D Mapping of Sound-Emitting Structures Based on Acoustic Ray Casting,” IEEE Tr. on Robot., 33(2), 2017, pp. 333-345.

[11] RB. Rusu et al., “Towards 3D Point Cloud based Object Maps for Household Environments,” Proc. IEEE Int. Conf. on Robot. and Autom., 2007, pp. 927-941.

[12] M. Lehtomaki et al., “Object Classification and Recognition From Mobile Laser Scanning Point Clouds in a Road Environment,” IEEE Tr. on Geo.Sci. and Rem. Sens., 54(2), 2016, pp. 1226-1239.

[13] T. Konno et al., “Audio-Visual 3D Reconstruction Framework for Dynamic Scenes,” Proc. of IEEE/SICE Int. Symp. on Sys. Integ. (SII 2020), 2020, pp. 802-807.

[14] R.Schmidt, “Multiple emitter location and signal parameter estimation” IEEE Tr. on Ant. and Prop., 34(3), 1986, pp. 276-280.

[15] J. M. Coughlan and A. L. Yuille. “The Manhattan World Assumption: Regularities in Scene Statistics which Enable Bayesian Inference,” Proc. of Int. Conf. on Neural Info. Process. Sys., 2000, USA, 809–815.

[16] M. A. Fischler and R. C. Bolles,“Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,” Comm. ACM. 24(6), 1981, 381–395.

[17] M. Ester et al.,“A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” Proc. Int. Conf. on Know. Discov. and Data Min., 1996, pp. 226-231

[18] P. Konstantinova et al.,“ A Study of a Target Tracking Algorithm using Global Nearest Neighbor Approach,” Proc. CompSysTech, 2003, pp. 290–295.

[19] J. Munkres, “Algorithms for the Assignment and Transportation Problems,” J. Soc. Indust. Appl. Math., 5(1), 1957, pp. 32–38.

[20] M. Wakabayashi et al., “Multiple Sound Source Position Estimation by Drone Audition Based on Data Association Between Sound Source Localization and Identification,” IEEE Robot. and Autom. Lett., 5(2), 2020, pp. 782-789.

[21] K. Nakadai et al., “Development, Deployment and Applications of Robot Audition Open Source Software HARK,” J. of Robot. and Mech., 29(1), 2017, pp. 16-25.

[22] M. Quigley et al., “ROS: an open-source Robot Operating System,”, ICRA workshop on Open Source Software 3 (3.2): 5, 2009.

[23] Q.Y. Zhou et al., “Open3D: A Modern Library for 3D Data Processing,” arXiv:1801.09847, 2018.

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-057-12 (11/20)

ロボット聴覚オープンソースソフトウェアHARK用ミドルウェア HARK middleware の紹介

HARK middleware: Middleware for open-sourced robot audition software HARK

木下智義[1*]　中臺一博[2,3]
Tomoyoshi Kinoshita[1]　Kazuhiro Nakadai[2,3]

[1] 株式会社ネットコンパス
[1] NetCOMPASS Ltd.
[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.
[3] 東京工業大学
[3] Tokyo Institute of Technology

Abstract: 本稿では，ロボット聴覚オープンソースソフトウェア HARK のミドルウェアとして開発した HARK middleware について紹介する．HARK では，従来，モジュール統合のオーバヘッドが小さいという実時間信号処理の要件を備えた flowdesigner/batchflow をミドルウェアとして採用してきたが，複数のデバイスの利用が難しい，分散処理ができないという制約があった．そこで，flowdesigner/batchflow の利点を踏襲しつつ，こうした問題を解決する HARK middleware を 2018 年にリリースした．しかし，HARK に関する文献の多くは，ロボット聴覚の主要課題である音源定位・分離・認識に関する信号処理アルゴリズムやその実装にかかるものであり，もう一つの特長である HARK middleware に関する文献は存在していない．今回，HARK middleware の pybind11 化，分散処理サポートを行ったので，これを契機に，HARK middleware の概要を紹介する．

1　はじめに

著者らは，これまでに，ロボット聴覚オープンソースソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) を開発し，提供してきた [1, 2]．HARK を用いることで，ロボット聴覚システムをはじめとした音響信号処理を簡便に構築することが可能となり，応用システムも紹介されている [3]．

その一方で，従来の HARK では，C++ベースで記述されているため開発の難易度が高く，データフローがいわゆる pull 型となっているために同時に複数デバイスの I/O 処理を扱うことが難しいなどの制約があった．また，処理が同一プロセス上で完結することが想定されており，別ホストと連携した処理を実現するには，ユーザが連携処理を記述する必要があった．

本稿では，これらの問題を解消しつつ，従来の HARK 機能を実現してきたソフトウェア資産を引き続き利用可能なミドルウェア HARK Middleware（以下 harkmw と呼ぶ）を紹介する．

*連絡先：〒 103-0024 東京都中央区日本橋小舟町 1 番 3 号 6 階
E-mail: kino@netcompass.co.jp

2　関連するミドルウェア

ロボットシステムの構築にこれまで多くのミドルウェアが発表されてきたが，その主要なものとして，ROS および OpenRTM-aist が挙げられる．

ROS[4] は，player/stage [5] を源流に持つ 2007 年にリリースされたミドルウェアで，その草創期は，Willow Garage 社によって開発が推し進められて，ロボット分野のデファクトスタンダードとしての地位を確立した．roscore と呼ばれるブローカを導入し，各ユーザが ROS ノードの作法に従って機能を開発することにより，比較的容易に ROS ノード間結合を実現することができる．しかし，ROS ノード間の連携には常にプロセス間通信を用いる必要があり，またそのためのシリアライズ処理においてもオーバーヘッドが存在する．これに対して harkmw では，ノード間の連携は単一プロセスで動作する限り関数コールのオーバーヘッドしかなく，

データの授受もポインタの受け渡しに相当する形で行われるため，低コストである．

一方，OpenRTM-aist[6] は，国際標準化団体 OMG (Object Manegement Group) [7] で標準化された RT コンポーネントインターフェース仕様に準拠した RT ミドルウエアの参照実装として，2008 年にリリースされ，その後も産総研を中心に研究開発が継続されている．OpenRTM-aist も ROS 同様，ブローカーベースでモジュール間通信を行うアーキテクチャを採用しており，ブローカには CORBA を採用している．CORBA も OMG 標準として汎用的に使われている技術であるが，信号処理ではフレームレベルの数十 ms 単位でのデータを連続的に処理を行っていく点を考慮すると，ROS で挙げた点と同様の問題がある．これに対して harkmw では，先述の通りノード間の連携のオーバーヘッドは小さなものとなっている．

3 harkmw の機能

HARK は，音響信号処理を主として様々な処理を，その構成の変更が容易な形で実現するフレームワークである．HARK は，個々の機能を「ノード」として管理し，ノードを必要に応じて組み合わせることで，任意の処理を実装することが可能である．

各ノードは，1 つ以上の出力と，任意数の入力を「端子」として持っており，ノードの端子を別のノードの端子に接続することで，データの流れを表現することになる．

旧来（バージョン 2.5 以前）は，フレームワークとして flowdesigner/batchflow(以降，flowdesigner と記述する) [8] をベースとしていた．バージョン 3.0 からは，flowdesigner に代わり，新規開発した harkmw を HARK のミドルウェアとして採用している．

harkmw は後述するように C++ で記述された既存の低レベル処理と連携して動作し，またプロセス間通信を用いた高レベルでの連携も同時に行う必要があることから，それらが比較的容易に実現できる Python により記述した．なお，Python による処理は必要最小限に留められており，処理速度の面でも従来のものと遜色なく動作している．

本節では，これらの harkmw の特徴のうち，特に harkmw におけるデータフローと，その実現方法について概説する．

3.1 HARK における処理の構成

HARK では，ノードと呼ばれる処理単位を組み合わせてネットワークを形成することで，処理を定義し実行することができる．

ノードには,「マイクから音声信号を取得する」といった入力機能を備えたもの,「信号に対して FFT を実行して出力する」といった加工等の機能を備えたもの,「WAV ファイルとしてファイルに出力する」といった出力機能を備えたもの，等が用意されている．

また，ノードを組み合わせたものをサブネットとして定義し，あたかもそれが複雑な機能を有したノード (DynamicNode) であるかのように，ネットワーク上に配置することも可能である．

DynamicNode には，subnet と iterator の 2 種類がある．subnet は，ノードを組み合わせてグループ化したものに相当し，ネットワーク内においてより複雑な機能を提供する．

iterator は，一定の条件を満たすまでループする機能を持ったサブネットであり，ループの最後の出力が，サブネット全体の出力となる．信号処理を実装するケースでは，MAIN ネットワーク（ミドルウェアが直接データを要求するネットワーク）に，iterator を 1 つ配置して一連の処理を実行することが多い．

次項では，これらを実現するために従来の HARK が採用していたデータフローと，harkmw におけるデータフローについて述べる．

3.2 従来の HARK におけるデータフロー

本項では，harkmw におけるデータフローの理解のために，従来の HARK におけるデータフローについて述べる．

3.2.1 pull 型アーキテクチャ

flowdesigner におけるデータフローは，いわゆる pull 型であった．すなわち，ネットワークの出力に相当するノードに対して，データを出力するよう要求し，ノードは必要に応じてその前段に相当するノードにデータを要求する．具体的には，各ノードを実装するクラスには `getOutput()` というメソッドが提供されており，下流のノード（あるいは flowdesigner）がこのメソッドを呼び出していた．

この方法では，各ノードは，要求に応じて上流からデータを取得して下流に返すという処理を行えばよく，実装はシンプルにすることができる．一方で，各ノードは自身が処理を開始するタイミングを知ることができず，特に外部との I/O に関わるノードでは問題となることがあった．

3.2.2 `count` 値に基づく時間管理

flowdesigner には `count` という状態値，および `lookAhead`，`lookBack` というパラメータがノードに

備わっていた．

count は，時間方向のフレーム番号を管理する値であり，処理開始時に 0 となり，以降順次インクリメントされる値である．すなわち，前述の getOutput() の引数は count であり，getOutput(count) は，時刻 count のデータを取得するためのメソッド呼び出しということになる．

複数のフレームをまとめて処理を行うブロック処理を行う場合には，時刻 count のデータを生成する場合に必要となる上流データは，時刻 count のものに限定することはできず，処理対象となっている複数フレームにアクセスできるよう時間的に幅を持ったスコープを設定する必要がある．この幅はノードにおける処理に依存しているため，パラメータとして持つこととなる．flowdesigner では，この値を lookAhead, lookBack という値によって管理している．すなわち，lookAhead が 0 でないノードでは，lookAhead フレーム分先のデータを用いて処理をするということであり，lookBack が 0 でないノードでは，lookBack フレーム分過去のデータを用いて処理するということである．

なお，flowdesigner では，lookAhead, lookBack の値は定数であることを想定しており，前後可変長時刻，特に「処理開始からすべて」「データの末端まですべて」という範囲に依存した処理は想定していない．

3.3 harkmw におけるデータフロー

これに対し，harkmw では push 型のデータフローを採用している．この場合，ネットワークの出力から「引き出す」データの取得ではなく，入力側から「流し込む」形で処理が進むこととなる．

harkmw の開発をスタートした時点で，HARK の提供するノードは多くの種類が既に実装され，それらは flowdesigner 用の pull 型を前提としたインタフェース設計となっていた．そこで harkmw では，push 型のデータフローに移行することとしながらも，各ノードの実装を変更するコストを抑制するため，ノードの実装は現状を維持したままとすることが求められた．

また，flowdesigner のオーバヘッドが小さいという利点を踏襲するために，ノード間のデータの送受は原則として従来同様 C++ ポインタの受け渡しとすることとした．データフローの変更により，各ノードの getOutput(count) は同一の count について複数回呼び出されることとなったが，既存のノード実装は内部にバッファ機構を持つものがほとんどであり，繰り返しの処理に対してはキャッシュされた結果を返すことができるため，速度面で不利となることはない．

3.3.1 push 型データフローの実現方法

harkmw では，全体として push 型のデータフローを用い，個々のノードにおいては旧来の pull 型を想定した呼び出し形式を維持することとなった．

これらを両立させる目的で，harkmw においては以下のようなアルゴリズムで各ノードの処理を実行することとした．

1. ネットワーク内で，他のノードに依存しない処理を行うノードを探す．
2. 当該ノードについて，ノード自身の処理を実行した後，以下 3. の処理を実行する．
3. ノードの下流にあるノードについて，以下を実行する．
   (a) ノードの各入力に接続されている上流ノードのそれぞれにつき，ノードの処理で用いるデータが準備できているかを確認する．
   (b) 準備ができていれば当該ノードの処理を実行し，当該ノードの下流にあるノードについて 3. の処理を実行する．
   (c) 準備ができていなければ当該ノードの処理の実行は保留して 3. を終了する．

例えば，図 1 のようなネットワークにおいて，従来の HARK におけるシーケンスは図 2 のようになっていた．harkmw においては，図 3 のようになる．

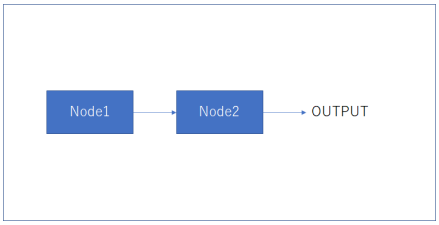


図 1: HARK ネットワークの例（1）

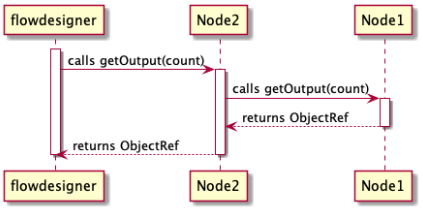


図 2: 従来の HARK におけるシーケンス

ところで，ノードの処理の実行タイミングについて，従来はノード間の getOutput() の呼び出しに依存していたのに対し，harkmw では harmkw が個々のノー

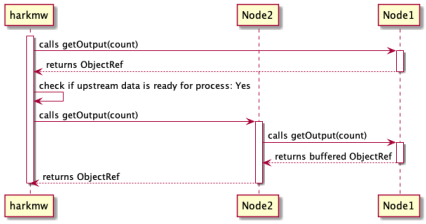


図 3: harkmw におけるシーケンス

ドの getOutput() を呼び出しているため，そのタイミングを制御する余地がある．今後のバージョンでは，I/O 処理を非同期的に実行するなど，より柔軟なタイミング制御機能を追加することも検討している．

3.3.2 count, lookAhead, lookBack の扱い

先述の通り，HARK の各ノードは lookAhead, lookBack というパラメータを持つ．これらの値は，当該ノードがその処理をするにあたって参照するフレームの範囲を定義するものであり，3.3.1 項に示したアルゴリズムにおいて「データが準備できている」の判定に用いられる．

図 4 に示したネットワークでの処理について考える．このネットワークでは，0 でない lookBack を持つノード（Average）が使われている．Average は，直近の指定したフレーム数の値を平均して出力する機能を持つ．また，Temperature は，時刻 count における周辺温度を返すノードである．(Average, Temperature は実際には実装されていないノードであるが，説明のため導入した)．

この場合，Average の上流にある Temperature では，Average からの getOutput(count) の呼び出しでの count 値が単調に増加せず，lookBack の範囲で前後することとなる．

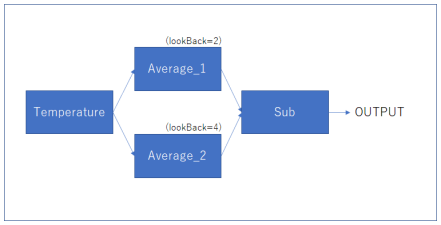


図 4: HARK ネットワークの例（2）

なお，実際のシーケンスは，図 5 のようになる．pull 型のデータフローでは，出力段のノードから引出されるように上流のデータが取得されていることがわかる．

一方，push 型データフローを導入した harkmw では図 6 のようになる．push 型のデータフローでは，上流のノードの処理を最初に実行し，その結果，下流のノードの実行の準備ができた場合に初めて下流の処理が実行されることがわかる．

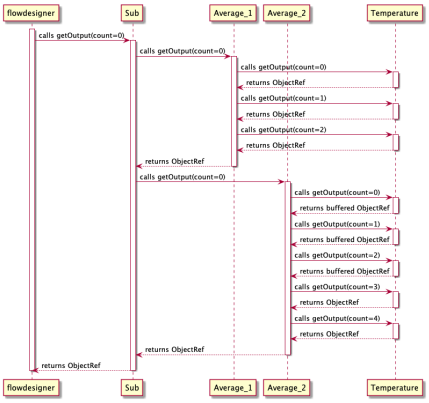


図 5: flowdesigner におけるシーケンスの例（2）

4 harkmw における新機能

harkmw においては，先述のデータフローの変更の他，いくつかの機能が新規に実装されている．本節ではそれらについて説明する．

4.1 複数プロセス実行

従来の HARK では，個々のノードがプロセス間通信の機能を実装するなどの方法を用いない限り，全ての処理は単一プロセスで実行されていた[1]．harkmw では，ネットワークを分割して複数プロセスで実行する機能が追加された．

複数プロセスにて harkmw を実行する場合，処理のプロセスへの割り当ては，ノード単位で行われる．具体的には，flowdesigner においても用いていたネットワーク定義ファイル（.n ファイル）に，どのプロセスで当該ノードを動作させるかを指定する要素を記述することで行う．

先述の通り，harkmw は，ネットワーク内の最も上流にあたるノードを探して処理を開始する．複数プロセスに分割して harkmw を実行する場合，そのようなノードが自プロセスが実行するノードの中には見つからない場合がある．そのような場合には，当該プロセスは，他のプロセスの処理結果に応じて自プロセスにおける処理を開始することとなる．

また，ノードの処理を実行した結果，下流のノードが他のプロセスが実行するものであった場合は，上流

[1]実際に，HARKDataStreamSender, SpeechRecognitionClient, HARK3.2 で新しく導入された HARK-CN などは各ノードで独自にプロセス間通信を実装している

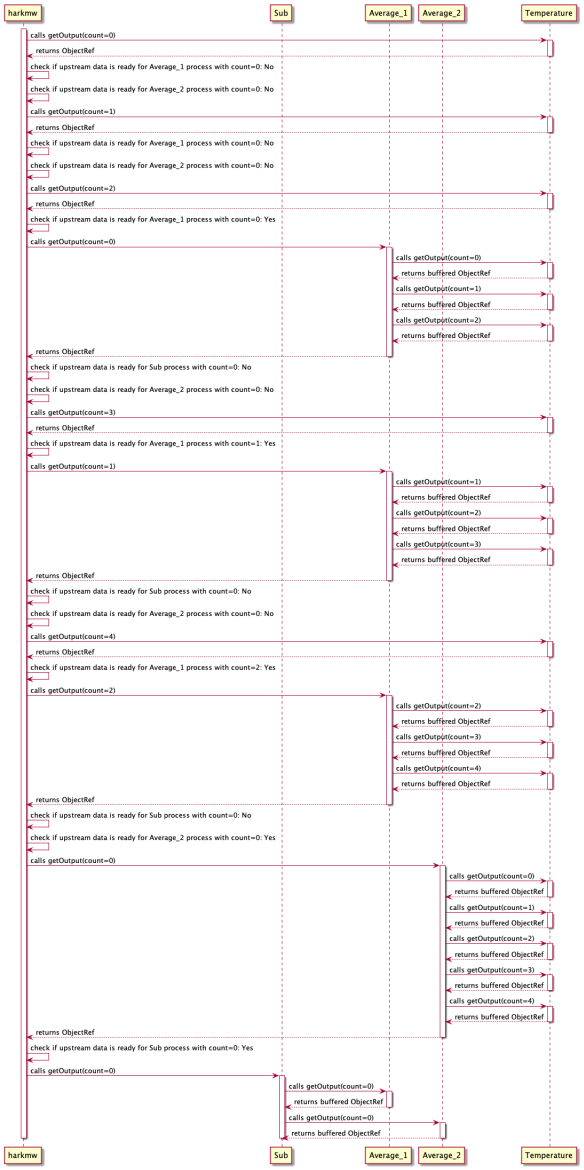


図 6: harkmw におけるシーケンスの例（2）

のノードの結果が得られたことを当該他のプロセスに通知して，他のプロセスにおける処理を開始する必要がある．これらのプロセス間連携処理に，harkmw では MQTT(Message Queuing Telemetry Transport)[9] を用いている．

具体的に，上流のノード A と，下流のノード B が定義され，以下の 2 通りのプロセス割り当ての設定の場合を比較してみる．

1. A，B ともに同一プロセスで実行されるよう設定されている場合
2. A はプロセス P において，B はプロセス Q において実行されるよう設定されている場合

1. の場合，A の処理が完了すると，B の実行可能条件を検証し，実行可能であれば B を実行することは先述の通りである．この時，B の処理の実行時には A の出力データが必要になるが，その受け渡しは，従来の flowdesigner の機構を流用する形で，ObjectRef 型の値を受け渡す実装となっている．これは実質的には C++ のポインタを渡すことに相当し，不要なデータの複製を発生させない効率的なデータの送受と，ノード間の疎結合を両立させている．

一方，2. の場合には，プロセス間でデータの送受を行う必要がある上，各ノードはその上流あるいは下流に存在するノードが同一プロセス上に存在するか，別プロセス上に存在するかを知ることができない．そこで，harkmw では，ノード間にデータの仲立ちを行うノード Proxy を挟むことでこの課題を解決している．すなわち，Proxy の両端に存在するノードが同一プロセス上で処理される場合，Proxy は何もせずに単に ObjectRef 型の値を引き渡すのみである．両端に存在するノードが別プロセス上で処理されている場合，Proxy は制御を harkmw に移し，他プロセスとのデータの受け渡しを実行する．

他プロセスへのデータのを行う際には，各ノードにおいて C++ で生成された ObjectRef 型を，Python 型に変換した上で，pickle したバイナリ列を MQTT の payload に載せる形で送信している．他プロセスからのデータの受信は，逆に，MQTT payload から得た pickle バイナリ列を Python 型に戻した上で，さらに ObjectRef へと変換して用いている．

4.2 harkmw デーモン機能

複数プロセスを用いた分散処理を行う場合，原則としてそれぞれのプロセスをユーザが個別に起動する必要がある．実際の運用を考慮すると，この制約は煩雑になることがあるため，その対処として，harkmw デーモン機能を実装した．

この機能を用いると，分散プロセスを稼働させるホストにおいてあらかじめデーモンを立ち上げておくことで，実際の処理時には，1 箇所で harkmw を起動することで分散プロセス全体を動作させることが可能となる．

5 harkmw のソフトウェア構成

harkmw は python モジュールとして実装されている．本節では，各モジュールおよびクラスについて，その機能をこれまでに述べた処理と対応させる形で説明する．

- module **harkmw.main**

プログラムの引数の解釈と，後述する Process の構築および実行を行う．実質的には，コマンドライン起動を想定した Process のドライバ，と言える．

- class `harkmw.process.Process` プロセス共通で用いる機能を提供する．具体的には，プロセスの初期化と終了の処理，MQTT ブローカとの接続およびデータの送受，等である．
- module `harkmw.defs` HARK のネットワーク定義ファイルを読み込んでメモリ上に展開管理する機能を提供するモジュールである．ネットワーク定義ファイルの読み込みと，実行時にその定義情報を参照する際に用いされる．
- module `harkmw.logic` harkmw の実行時の各種制御を担当するモジュール．
- class `harkmw.logic.builder.Builder` ネットワーク定義ファイルを元に構築された `harkmw.defs` の各クラスの情報を元に，ネットワークを構成する `Node`，`Network` 等を生成する．
- class `harkmw.logic.node.Node` `NativeNode`，`MQTTNode`，および，`Network` の基底クラスであり，共通処理を実装する．
- class `harkmw.logic.mqtt_node.MQTTNode` MQTT を介して動作するノードを管理するクラス．
- class `harkmw.logic.network.Network` ネットワーク，あるいは DynamicNode の処理を行うクラス．いわゆる subnet として動作する場合には単一のフレームについて，iterator として動作する場合には条件が成就するまでループしながら各フレームの処理を行う．
- class `harkmw.logic.proxy.Proxy` ノードとノードの間に存在する `Proxy` であり，Python 型のクラスとして各種処理を実行する他，C++ レベルの `Proxy` との連携も担当する．
- class `harkmw.types.dynamic.Dynamic` HARK の各 toolbox をロードするためのクラス．
- class `harkmw.types.native_node.NativeNode` HARK の各ノードを，C++ で実装されたクラス `harkmwnative.Node` と連携しながら Python から操作するためのクラス．
- module `harkmwnative` src ディレクトリ以下にある C++ ソースファイルによって提供されているモジュールである．C++ および pybind11 を用いて，HARK のノード等機能と python との橋渡しを行う．
- class `harkmwnative.Node` HARK ノードを wrap して Python から操作するためのクラス．
- class `harkmwtative.Proxy` HARK ノード間を接続する proxy ノードを wrap して Python から操作するためのクラス．

6 むすび

本稿では，HARK Middleware (harkmw) を紹介した．harkmw を用いることで，従来の HARK ソフトウェア資産を活用しながら，より柔軟な構成でロボットシステムを構築することが可能となる．

参考文献

[1] HARK Official Site https://www.hark.jp/

[2] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system HARK. *Advanced Robotics*, Vol. 24, pp. 739–761, 2010.

[3] 「ドローンが耳を澄まして要救助者の位置を検出〜災害発生時の迅速な救助につながる技術を開発〜」https://www.jst.go.jp/pr/announce/20171207-2/index.html

[4] ROS, http://wiki.ros.org/ja

[5] player/stage, http://playerstage.sourceforge.net/

[6] OpenRTM-aist, https://www.openrtm.org/openrtm/

[7] The Object Management Group, http://www.omg.org

[8] C. Côté, D. Létourneau, F. Michaud, J.-M. Valin, Y. Brosseau, C. Räievsky, M. Lemay, and V. Tran. Reusability tools for programming mobile robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, pp. 1820–1825. IEEE, 2004.

[9] MQTT: The Standard for IoT Messaging, https://mqtt.org/

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-057-13 (11/20)

言語獲得能力を備えた音声対話エージェントの検討

On Language Acquisition With Spoken Dialogue Agent

篠崎隆宏 * 高 聖洲 張 明鑫 侯 汶昕 田中 智宏

東京工業大学

Abstract: We propose an end-to-end neural network-based spoken language acquiring agent that combines unsupervised learning and reinforcement learning. The agent first learns sound words from unlabeled speech waveform and makes a sound dictionary. Then, the agent uses the sound dictionary as an action space during spoken dialogues. With the proposed method, reinforcement learning compensates the insufficient accuracy of unsupervised learning, while unsupervised learning significantly contributes to improve the efficiency of reinforcement learning. Simulation experiments demonstrate that the agent efficiently learns to speak appropriate word utterances based on the outside environment and its internal desire.

1 はじめに

生活環境において人と共存し人の生活をサポートするロボットが実現すれば，高齢化の進む将来社会において有用と期待される．ロボットが個別の環境下で柔軟な活動を行うためには，日々新し知識を継続的に取得するとともにそれを言語表現する高い学習能力が求められる．人にとって音声言語は一次的であり，文字言語を持たない言語はあっても音声言語を持たない言語は知られていない．また，道具を使用せずに身体のみを用いてパラ言語情報を伴った豊かな意思伝達を即座に行うことができるという，文字言葉にはない特徴がある．人の音声言語学習能力は強力で，特定の言語の知識を何も持たない状態で生まれた後社会生活を行う中で母国語を獲得することができる．しかし現在一般的な教師あり学習を用いた対話システムでは，音声対話を行う中で学習を行い言語知識を拡張するということができない．人と真に柔軟な音声対話を行えるロボットを実現するためには，人との日常会話の中で閉じた学習ループを形成する学習アルゴリズムの実現が不可欠である．

人がどのように音声言語を獲得しているのか未だ完全な理論や工学モデルは実現していないが，幾つかの試みは行われている．初期の研究として，スキナーは行動心理学の立場から人は単語に意味を結びつける強化理論に基づいて言語を学習しているとの仮説を提唱している [1]．近年では，ロボットに音声言語を自動獲得させることで構成的に音声言語獲得を理解すると共に工学的に応用しようとする研究が行われている．しかし，語彙獲得と意味理解および状況に応じた適切な発話発声の全てを同時に実現しているシステムは存在していない [2]．本研究では教師なし学習と強化学習により音声言語獲得におけるこれら全ての側面を同時に実現するニューラルネットワークシステムを提案し，評価実験を行う [3, 4] [1]．また，今後の課題について検討する．

*連絡先：東京工業大学工学院情報通信系
〒 226-8502 神奈川県横浜市緑区長津田町 4259-G2-2
www.ts.ip.titech.ac.jp

2 関連研究

Roy 等によるシステム [5] は，連続的に発話された音声から単語境界の推定を含めて語彙を学習することができる．また共起関係をもとに単語と画像オブジェクトの対応を学習することもできる．しかし事前に教師あり学習した音素認識器の利用が前提となっており，完全なゼロ知識からの学習にはなっていない．また単語の学習を目的としており，行動の学習は対象とされていない．事前学習された音素認識器としてであるが，ニューラルネットが HMM やヒストグラムとともにシステムの一部として用いられている．岩橋は，隠れマルコフモデル (HMM) と統計文脈自由文法 (SCFG) により構成されたシステムを提案している [6]．このシステムでは一単語ごと発話した単語の発声から語彙を獲得し，相互情報量に基づき共起関係を定式化することで単語と視覚を結びつける学習を行っている．ロボットへの要求発話の認識や行動も学習に基づいているが，認識結果を行動に結びつける方策はヒューリスティックなプログラムに依存している．杉浦等のシステムでは，単語学習結果をもとに確率推論結果の信頼度を学

[1] 提案法を実装したプログラムを `https://github.com/tttslab/spolacq.git` で公開している．

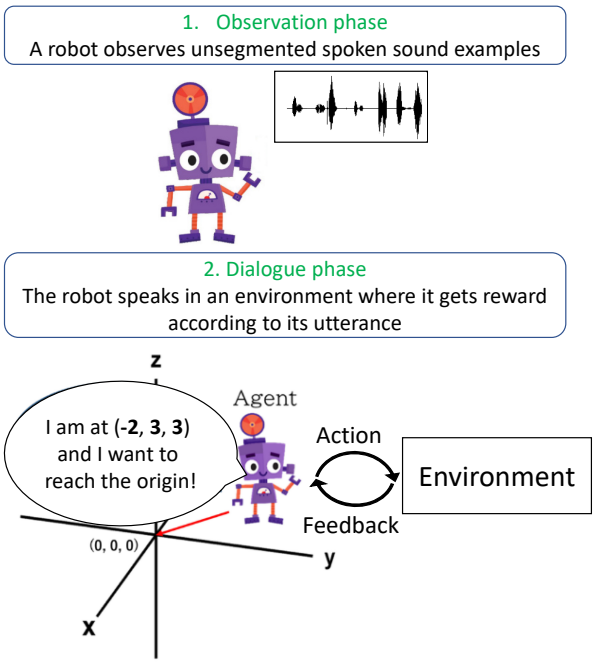


図 1: Spoken language acquisition system based on unsupervised word learning and reinforcement based dialogue leanring.

習して対話制御に用いている [7]．しかし，信頼度を用いた対話制御および確認発話は固定されている．谷口等は階層的なノンパラメトリックベイズ法と特徴量抽出にニューラルネットを組み合わせたシステムを提案している [8]．しかし学習済みの音素認識器が仮定されている点でゼロ知識からの学習ではなく，また発話方策の学習は含まれていない．Harwarth 等はニューラルネットを用いてラベル付けのされていない音声と画像から音声単語と対象物の関係を直接学習させる手法を提案している．しかし音声と画像のグランディングを対象としており，発話方策の学習は含まれていない．[9]．羽鳥等は，ロボットが強化学習による学習に基づき音声で指示された行動を実行するシステムを提案している [10]．このシステムでは行動が強化学習により学習されるが，音声入力は学習済みの音声認識器を用いてテキストに変換されており，音声処理は学習に含まれていない．また，音声の発話は対象とされていない．

3 提案方法

音声言語獲得機能を備えた音声対話エージェントは，原理的には感覚入力と音声合成器を備えたニューラルネットエージェントに強化学習を適用することで実現できるはずである．そこでは，音声発話が強化学習における行動に該当する．しかし音声発話は可変長の高次元連続データであるため，ランダムに初期化された方策関数が状況に応じた意味のある発話を生成する確率は限りなく 0 に近い．発話行動が成功しなければ強化学習による学習は進まない．そのため強化学習のみを用いて音声言語獲得をさせようとすると，学習の収束に非現実的な時間が必要になってしまう問題がある．同様の困難性は人間の子供が音声を学習する際にも存在するはずであるが，人間の場合は周囲が話す音声を観察して真似ることで対話における試行を効率化していると考えられる．そこで，周囲の音声やその他画像入力等を観察し教師なし学習を行う観察学習と，教師なし学習により得られた音声単語集合を行動空間として強化学習を行う対話学習を組み合わせて行う学習アルゴリズムを提案する．

エージェントの基本設計として，エージェントに内在する欲求を満たすように外部世界からの入力に応じて音声発話を行うシステムを想定する．エージェントの内部欲求をどのように設計するかは重要な点であるが，本研究では対象外である．本研究では，任意に設定された欲求をもとに音声言語獲得を行う一般性のある仕組みについて取り組む．

以下では，はじめに観察学習として教師なし単語学習のみを用いる基本システムについて説明し，ついで音声画像接地をもとにロボットの注意を視界中のオブジェクトに集中させることで単語学習効率を向上させた拡張システムについて説明する．

3.1 教師なし単語学習に基づく音声言語獲得システム

教師なし単語学習法として，ES-KMeans 法 [11] などが提案されている．現状単体で十分な精度の学習を行える手法は存在しないが，強化学習と組み合わせて使用する場合は行動空間を離散化し探索を大幅に削減する効果が期待できる．そこで，図 1 に示す教師なし単語学習と強化学習を組み合わせた自動音声言語獲得手法を提案する．単語の教師なし学習では間違った単語セグメントを排除するよりも必要語彙のすべての単語の正しいセグメントが含まれることを優先し，生成される音声辞書のサイズは十分に大きくとる．

対話学習では，エージェントが 3 次元空間内のランダムな初期位置に置かれた状況を想定する．エージェントは原点に到達したいという内部欲求を持っている．エージェントは自分で直接移動することはできないが，方向を示す音声コマンドを発声するとエージェントが乗っている乗り物がそちらの方向に一ステップ進むことを想定する．強化学習には，教師なし単語学習で得た音声辞書を行動空間とする Q 学習を用いる [3]．

3.2 音声画像接地に基づく注意機構を導入したシステム

提案するエージェントの構成を図 2 に示す．提案法は，観察フェーズにおいて音声辞書の作成とともに音声と画像の関係を教師なし学習しておき，対話フェーズにおいて強化学習を行う際にロボットの視界にある画像と関連の高い音声単語の確率が高くなるように行動価値関数を補正することを原理とする．確率の補正は，行動価値関数を実装するニューラルネットに入力画像との類似度を入力することにより行う．さらに，教師なし学習により得た初期の音声辞書にある誤りを音声対話学習時に上書きし精度を高めるための機構として Action filter を提案し導入した [4].

想定する音声対話タスクを図 3 に示す．観察学習フェーズでは，ロボットは食べ物の写真を見ながら音声を聞く．音声は 1 から数単語からなる食べ物の名前についての発話である．発話ラベルや発話に含まれる単語数は，ロボットには一切与えられない．対話フェーズではロボットは 2 種類の食べ物の写真を提示される．ロボットは自身の内部状態に依存して，食べ物に対する嗜好を持っている．提示されている食べ物のうち自分の欲する方の食べ物の名前を正しく発声できれば，その食べ物が与えられ報酬を得る．反対の食べ物や提示されていない食べ物の名前を発声した場合や，発声が食べ物の名前として認識されない場合，報酬は得られない．

4 3D 空間での原点回帰対話実験

4.1 実験条件

教師なし単語学習には，Google Speech Commands Dataset の音声を用いた．データセットの発話のうち方向を表す 6 種類の単語 (up, down, left, right, forward, backward) および空間移動とは関係のない 1 種類の単語 (marvin) のそれぞれについて 200 サンプル，合計 1400 サンプルを連結して一つの音声ファイルとした．教師なし単語学習法は ES-KMeans 法とともに，比較のために単にランダムに音声をセグメント化するランダム法も用いた．教師なし単語学習により得た，壊れたセグメントを含む音声単語辞書のサイズはおよそ 2000 である．対話タスクにおいてエージェントが発話した音声は，一般タスクの音声認識器である Google Speech-to-Text API [2] を用いて認識した．教師なし学習した単語辞書中で音声認識器により方向を表す単語に認識されたセグメントの割合は，ES-KMeans 法を用いた場合が 22.4%，ランダム法を用いた場合が 11.8%であった．対話フェーズにおいて，ランダムな位置に初期配置されたエージェントが原点にたどり着くまでが 1 エピソードである．強化学習は初期値を変えながら 100 回繰り返し，エピソード数に対する報酬の平均と分散を求めた．

[2] https://cloud.google.com/speech-to-text/docs/reference/rest/

4.2 実験結果

教師なし単語学習を用いた音声言語学習エージェントの学習の様子を図 4 に示す．ES-KMeans 法を用いた方が学習効率が高いが，ランダム法を用いた場合でも学習が進めばほぼ同じ対話精度が得られている．

5 注意機構を導入した食べ物要求対話実験

5.1 実験条件

画像データとして，20 種類の食べ物（ apple, banana, carrot, cherry, cucumber, egg, eggplant,green pepper, hyacinth bean, kiwi fruit, lemon, onion, orange,potato, sliced bread, small cabbage, strawberry, sweet potato,tomato, white radish) の写真をそれぞれ 120 枚撮影した．各食べ物画像に対して，4 種類のテンプレート（e.g., “Apple,” “An apple,” “A red apple,” and “It’s an apple.”）をもとに音声合成器 [3]を用いて説明音声を作成した．合成音声には 20dB のガウス雑音を重畳させた．ロボットの食べ物に対する嗜好としては，ロボットの内部状態として RGB カラーをランダムにセットし，その色に近い食べ物を欲するものとした．

5.2 実験結果

音声画像接地による意識集中機構を導入した音声言語学習エージェントの学習の様子を図 6 に示す．語彙数が増えているため意識集中機構を導入しない場合は学習がほとんど進行しない一方で，意識集中機構を導入することで学習が大幅に効率的に進むことがわかる．さらに Action filter 機構を導入することで初期音声単語辞書にある誤りを修正し，高精度な音声対話が学習されることが分かる．また意識集中機構や Action Filter 機構を導入しない場合は，教師なし音声辞書学習に ES-KMeans を用いる方法はランダム法と比べて学習速度の向上に貢献していることが分かる．しかし，意識集中機構や Action Filter 機構を導入するとそれらの効果が大きく，音声辞書の作成方法の違いの差は殆どなくなる結果となった．

[3] https://pypi.org/project/gTTS/

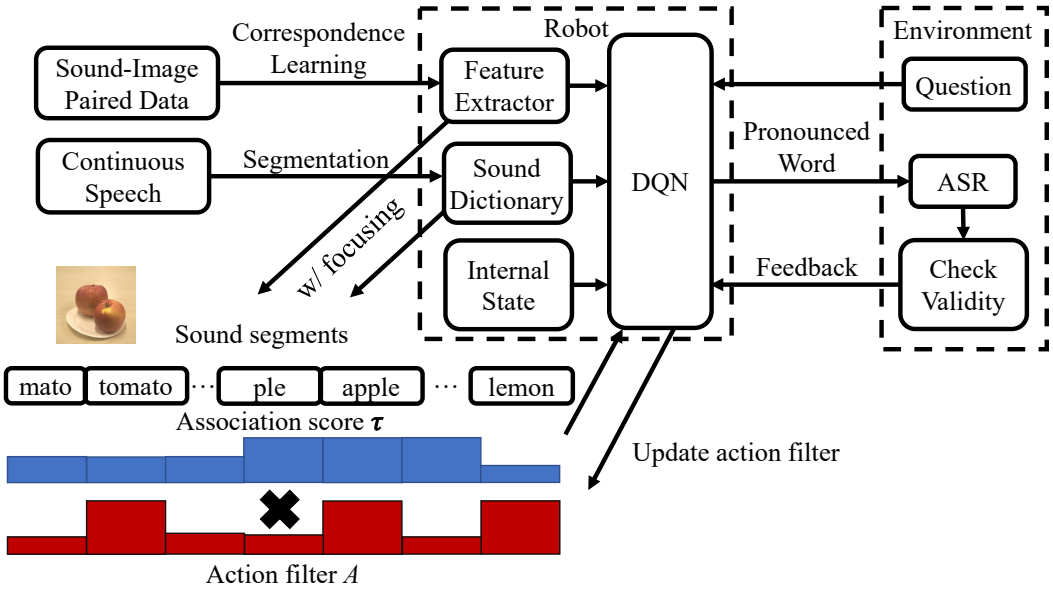


図 2: Proposed spoken language acquisition agent with sound-image grounding based focusing mechanism and the learning environment.

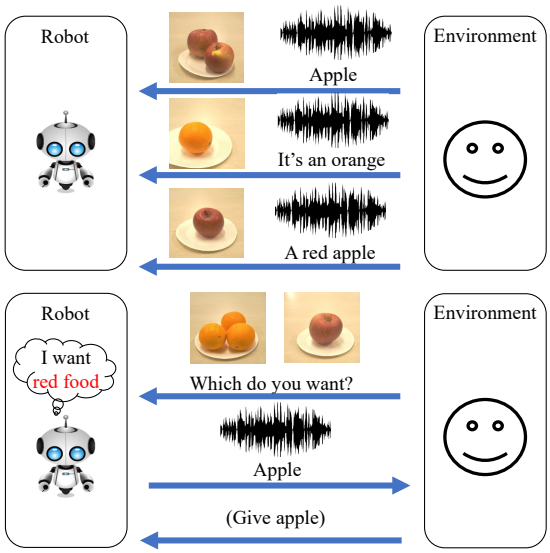


図 3: Spoken language acquisition task with image presentation.

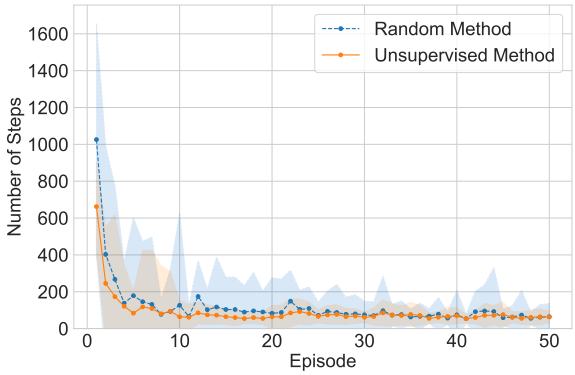


図 4: Learning process of the agents by the spoken dialogue with the environment that recognizes the speech commands.

6 エージェントの発話機構について (ディスカッション)

提案した２つの音声言語獲得エージェントの発話はどちらも，教師なし学習で得た音声辞書から選択した音声セグメントの再生に限られている．そのため，声質は観察学習の時に観察した音声の話者のままである．また感情その他を表現する抑揚の制御も音声辞書の要素の選択を通してしか行えず，自由度が低い．より柔軟な音声発話を行えるようにするためには，音声生成部を自由度の高い音声合成器に置き換える必要がある．一方強化学習を効率的に進める観点からは，エージェントのランダムな試行が音声発話として意味のあるものになる確率がある程度の大きさを持つ必要がある．そのためには，コンパクトな連続空間から音声発話を合成する機構が必要である．このために，今後の課題として音素ラベルによる条件付けを必要とせず音声を生成できる WaveGAN [12] を応用することなどが考えられる．また，人の場合は声帯や声道などから構成され

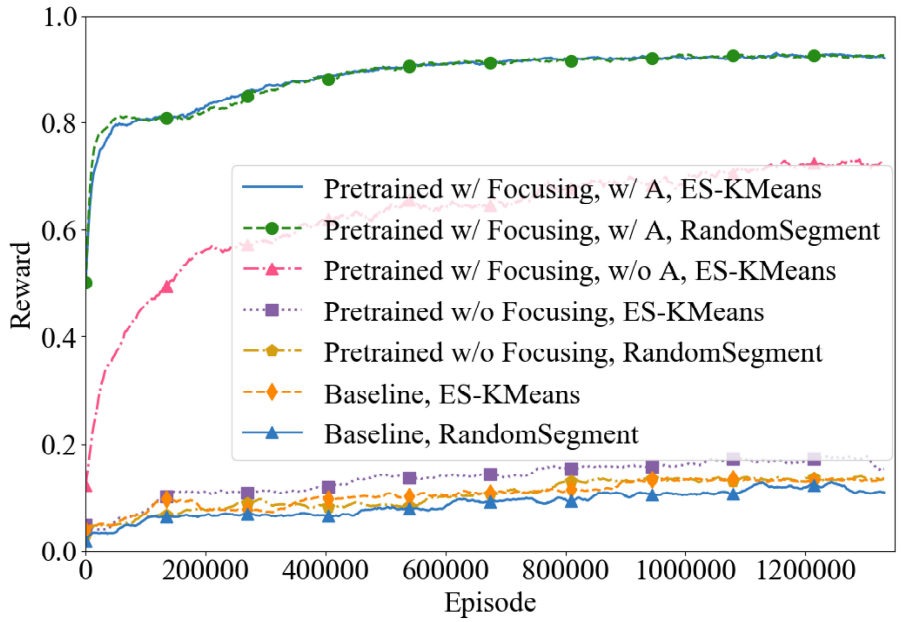


図 5: Learning performance of the language acquisition agent in the dialogue phase.

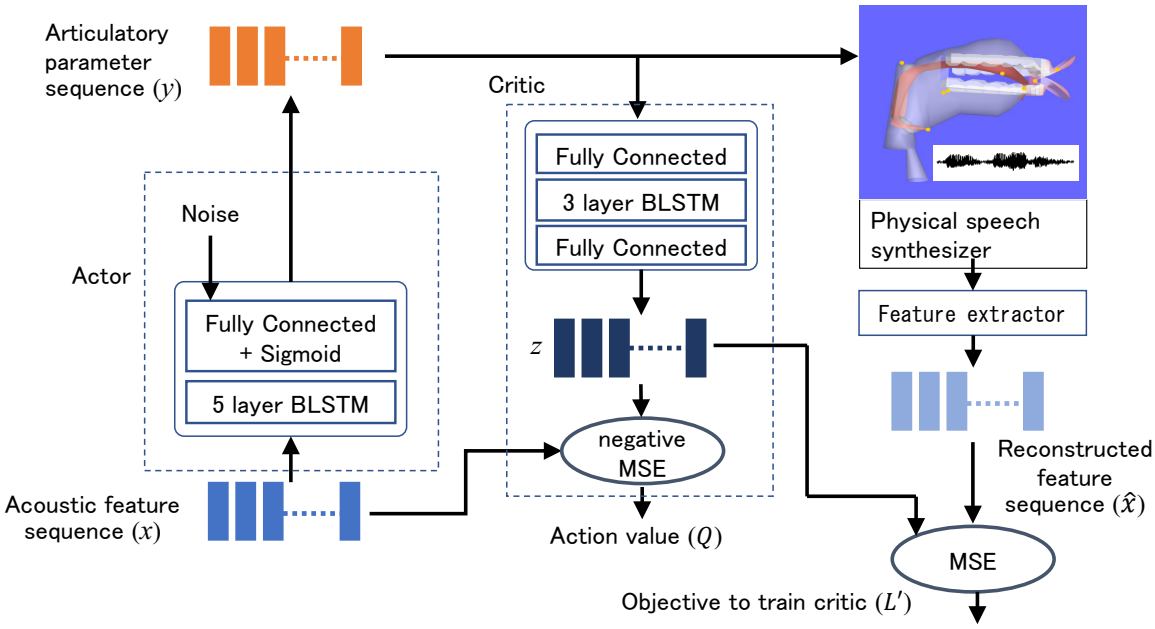


図 6: Hybrid autoencoder based self-learning of physical speech synthesis system using deterministic policy gradient.

る音声生成機構の限定された自由度が効率的な音声発話学習に貢献している可能性も考えられる．また発声器官の物理的な形状が，音声の話者性を生じさせる重要な因子となっている．音声言語獲得エージェントにおいても，人間の音声生成機構をそのまま物理的に模擬する物理音声合成器 [13] を事前知識の一形態として用いることも考えられる．制御器となるニューラルネットと物理音声合成器を接続してハイブリッドなオートエンコーダを構成し決定的方策勾配法を用いることで自己教師あり学習を実現することができ [14]，これを応用することが考えられる．

7　まとめ

教師なし単語学習と強化学習を用いた，音声言語獲得エージェントの仕組みを提案した．教師なし単語学習を用いて強化学習の探索空間を離散化することで，現実的な試行回数で学習が進むことを示した．辞書の精度は高いほうが望ましいが，音声辞書が必要単語の正しいセグメントを含んでいれば精度が低くても学習が進むことを示した．また，視覚に基づき音声辞書中の特定の単語に注意を向ける仕組みを提案した．ロボットが自分の望む食べ物を音声発話により得るタスクを設定したシミュレーション実験を行い，提案法の有効性を示した．今後の課題として，音声発話の柔軟性をより高めることが挙げられる．

謝辞

本研究は東レ科学振興会の助成を受けたものです．

参考文献

[1] B. Skinner, “Verbal behavior,” in *Verbal behavior*. New York: Appleton-Century-Crofts, 1957.

[2] T. Tangiuchi, D. Mochihashi, T. Nagai, S. Uchida, N. Inoue, I. Kobayashi, T. Nakamura, Y. Hagiwara, N. Iwahashi, and T. Inamura, “Survey on frontiers of language and robotics,” *Advanced Robotics*, vol. 33, no. 15-16, pp. 700–730, 2019. [Online]. Available: https://doi.org/10.1080/01691864.2019.1632223

[3] S. Gao, W. Hou, T. Tanaka, and T. Shinozaki, “Spoken language acquisition based on reinforcement learning and word unit segmentation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6149–6153.

[4] M. Zhang, T. Tanaka, W. Hou, S. Gao, and T. Shinozaki, “Sound-image grounding based focusing mechanism for efficient automatic spoken language acquisition,” in *Proc. Interspeech*, 2020, pp. 1436–1440.

[5] D. ROY, “Learning words from sights and sounds : A computational model,” *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

[6] N. IWAHASHI, “Language acquisition through a human-robot interface,” *Proc. Int. Conf. Spoken Language Processing*, 2000.

[7] 杉. 孔明, 岩. 直人, 柏. 秀紀, and 中. 哲, “言語獲得ロボットによる発話理解確率の推定に基づく物体操作対話,” **日本ロボット学会誌**, vol. 28, no. 8, pp. 978–988, 2010.

[8] T. Taniguchi, T. Nakamura, M. Suzuki, R. Kuniyasu, K. Hayashi, A. Taniguchi, T. Horii, and T. Nagai, “Neuro-serket: Development of integrative cognitive system through the composition of deep probabilistic generative models,” *New Generation Computing*, vol. 38, no. 1, Jan 2020.

[9] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, *Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input: 15th European Conference, Munich, Germany, September 8 – 14, 2018, Proceedings, Part VI*, 09 2018, pp. 659–677.

[10] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, “Interactively picking real-world objects with unconstrained spoken language instructions,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3774–3781.

[11] H. Kamper, K. Livescu, and S. Goldwater, “An embedded segmental k-means model for unsupervised segmentation and clustering of speech,” in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 719–726.

[12] C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” in *Proc. ICLR*, 2019.

[13] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” *PLOS ONE*, vol. 8, no. 4, pp. 1–17, 04 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0060603

[14] H. Shibata, M. Zhang, and T. Shinozaki, “Unsupervised acoustic-to-articulatory inversion neural network learning based on deterministic policy gradient,” in *Proc. IEEE Spoken Language Technology*, 2020, accepted.

口蓋形状から呼吸系・心臓系疾患を予測する手法の検討

Examination of methods for predicting respiratory and cardiac diseases from the shape of the palate

馬場嘉朗[1]　馬場達朗[2]　酒井経雄[3]

Yoshirou Baba[1], Tatsurou Baba[2], and Tuneo Sakai[3]

[1] 九州工業大学生命体工学専攻博士課程学生(D3)
[1]Kyushu Institute of Technology
[2] 馬場技術士事務所代表
[2] Baba Engineer Office
[3] 酒井デンタルクリニック代表
[3] Sakai dental clinic

Abstract: To investigate the relationship between palatal shape and sleep apnea syndrome (OSAS) and heart diseases such as premature ventricular contraction (PVC) and atrial fibrillation (AF). The maxillary tooth profile, OSAS, PVC, and AF indexes were collected and modeled. It was found that the OSAS index of respiratory diseases can be predicted with high accuracy from the palate shape parameters. On the other hand, sufficient prediction accuracy was not obtained for cardiac diseases. We examined whether it is possible to predict the disease itself from palate photographs of critically ill patients. It was found that the Residual Neural Network model can discriminate image data for which discrimination accuracy could not be obtained with the cNN model.

1　はじめに

口蓋形状や噛み合わせは、呼吸器系疾患や心臓系疾患と関係がある。口蓋形状から求められる口蓋断面積と閉塞性睡眠時無呼吸症候群（OSAS: Obstructive Sleep Apnea Syndrome）と深い関係があることや噛み合わせ不良によるストレスが心筋梗塞の引き金になることは知られている。本報告では、これらの関係を解釈可能な回帰モデルを使って解析し簡単に個々人の重症度を予測する手法を示した。心臓系疾患レベルは回帰不能なため機械学習（ブラックボックス）モデルで疾患予測を試みた。

第２章では、図１に示すように、口蓋形状と睡眠時無呼吸症候群（OSAS）重症度を表す STOPBAN 指標、心室期外収縮（PVC:Premature Ventricular Contraction)重症度を表す Lown 指標、心房細動（AF:Atrial Fibrillation)重症度を表す CHADS 指標のモデル化を試みた。入力の口蓋形状のパラメータは上顎歯形から抽出し 121 人分のデータを収集した。

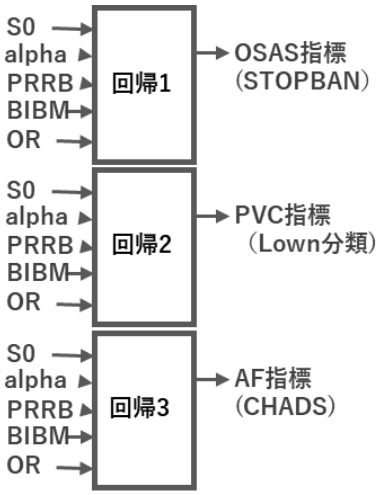


図１　口蓋形状パラメータと疾患指標

連絡先:1 Tel:045-513-9390 mail: sastbaba@yokohama.email.ne.jp
2 Tel:0287-29-2360 mail: yff57718@nifty.com
3 Tel:0287-23-6485

出力の呼吸器系 OSAS 指標と併せて心臓系 PVC 指標と AF 指標を収集した。

第３章では、図２に示すように、上顎歯形画像から直接疾患の予測が可能かを検討した。
正常、OSAS、PVC、AF、OSAS＋AF 重症者の上顎歯形写真合計 22 枚をベースにデータ拡張を行い機械学習で判別器を生成した。

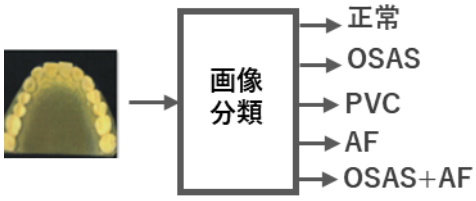


図２　上顎歯形画像による疾患予測

第 2 章、第 3 章両アプローチの結果、歯形を提供頂いた患者さんの OSAS レベル予測は、気道断面積と気道角度が支配的要因であることが判り、口蓋形状の回帰モデルで精度よく予測できることが判った。一方で、PVC や AF は口蓋形状からは予測できないことが判った。呼吸器系疾患と心臓系疾患を予測する別手法として、上顎歯形画像を使って直接画像分類することを検討した。Sony 製の NNC（Neural Network Console）を使用し CNN(Convolution Neural Network)と ResNet（Residual Neural Network）で分類器の試験を行った。結果、CNN で分類できない画像でも ResNet では分類できることが判った。

2　口蓋形状による呼吸器系（OSAS）及び心臓系(PVC、AF)重症度の予測

2.1　口蓋形状パラメータ

図３に、口蓋形状から抽出した 5 個のパラメータを示した。口道の長さと断面積に関係するものである。alpha は歯顎部放物平面(カンペル平面に平行)と x 軸（OJ 平面, J は劣弧中心）のなす角度で正常者は約 30° と言われている。OR は切歯乳頭上線（前歯の付け根）から左右第二大臼歯遠心中心までの距離である。S0 は口蓋断面積で、口蓋形状を放物線近似して求める。第二大臼歯(遠心位)を基準に臼歯間の距離の 1/2 を BM (BI*)、その高さ RB(PR*)とする。*S0 を適切に求めるため放物線近似の補正が必要であり、第二大臼歯が欠損している場合は第一大臼歯を基準にする場合もある。[1][2]

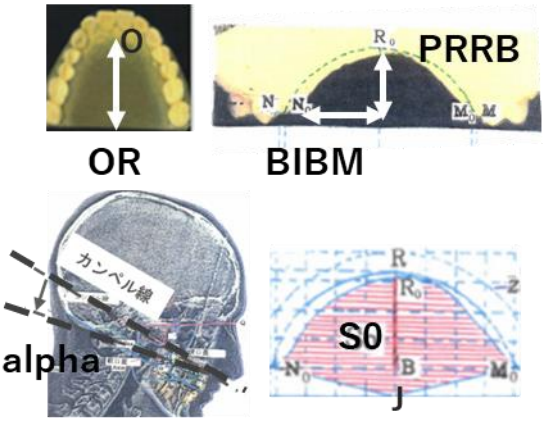


図３　口蓋形状パラメータ

2.2　重症度指標

閉塞性睡眠時無呼吸症候群（OSAS）の重症度指標としては患者の問診票をベースに作られる STOPBAN 指標（8 段階）と Mallampati 指標（5 段階）が一般的である。今回は、連続値から離散値への回帰を行うことから指標レベルが多い STOPBAN 指標を採用した。STOPBAN は問診内容の頭文字をとったものでありその内容を表１に示した。

表１　OSAS 指標

S	Snoring	いびき
T	Tired	疲労感
O	Observed apnea	無呼吸
P	Presure of Blood	高血圧
B	BMI	肥満度
A	Age	年齢
N	Neck circumfence	首回り

心臓系疾患として、不整脈のひとつである心室期外収縮（PVC）と血栓形成を引き起こし脳を含めた全身に塞栓症を起こす心房細動（AF）の指標を表２と表３に示した。PVC は一般的に不整脈発生頻度による Lown 分類が用いられる。今回のデータ収集に使った患者にはレベル 5 の人は存在しなかった。今回は正常(0)～レベル 4 までの 5 段階を採用した。

AF 指標は、問診票をベースとした指標で症状の頭文字を取って CHADS で 6 段階を採用した。[3]

表２　PVC 指標

Lown grade	PVC出現状況	ラベル
0	期外収縮なし（正常）	0
1	散発　30回未満/時間	1
2	多発　30回以上/時間	2
3	多形成	3
4a	2連発	4
4b	3連発	
5	RonT（T波にQRS波乗重）	5

表３　AF 指標

C	Congesive Heart Failure	心不全
H	Hyper Tension	高血圧
A	Age	75歳以上
D	Diabetes Mallitus	糖尿病
S	Stroke/TIA	脳卒中

2.3　回帰モデルと回帰精度

表４に今回使用したデータテーブルを示した。左側の青色着色部が５個の口蓋形状パラメータであり、右側のピンク色着色部が OSAS 指標と PVC 指標、AF 指標である。データは 121 組あり、指標ごとに最適な回帰モデルの適用を試みた。

線形回帰、Lasso 回帰（基本的に線形に同じ）、サポートベクトル回帰（SVR）を使用し、121 データすべてを使って回帰モデルを生成し、実データと回帰予測データの誤差評価を行った。回帰精度評価には R^2（R-square 値）を使用した。

表４　口蓋形状パラメータと指標データ

OR	PRRB	BIBM	S0	alpha	OSAS	PVC	AF
4	1.3	2	4.6	10	6	0	3
4	1.4	2	5.2	6	1	0	3
4	1.7	2.2	5.5	4	5	0	3
4	1.6	2	5.6	6	7	0	3
4	1	1.8	4.2	17	4	0	1
4	1.5	2	4.5	6	3	0	3
4	1.3	2	4.4	9	2	0	3
4	1.5	2	4.6	7	5	0	3
4	1.5	2	5.4	7	4	0	3
4	1.7	2	5.9	6	4	0	0

各種回帰による 3 指標の回帰精度を図４に示した。呼吸器系の OSAS 指標に使用した STOPBAN は全ての回帰で、R^2 が 0.85 以上と高い精度が得られた。

一方で、心臓系の PVC 指標の R^2 は 0.6 程度、AF 指標に至っては 0.25 程度と口蓋形状パラメータから正しく予測することは不可能であることが判った。

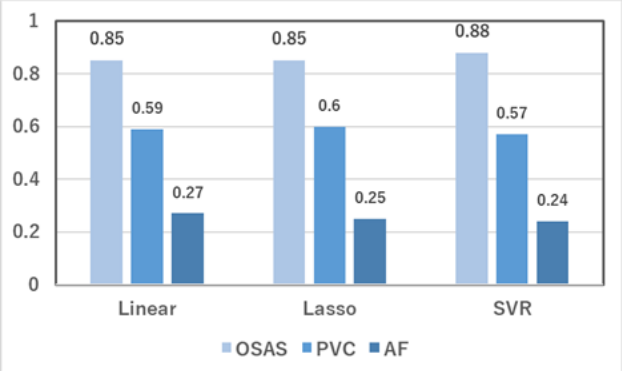


図４　疾患指標の各種回帰による予測精度（R^2）

2.4　OSAS 指標の高精度モデル化とユニバーサル化

呼吸器系疾患の OSAS 指標は、口蓋形状パラメータで高い精度の予測が可能であることを示したが、更なる予測精度向上を目指し、連続値-離散値に対応する Random Forest 回帰を追加し回帰精度を比較した。結果を図５に示す。Random Forest 回帰により R^2 値が 0.96 となり、SVR（R^2 値 0.88）以上の高い精度が得られた。重症度レベルの予測結果と実レベルの正解率は 107/121（89%）であった。

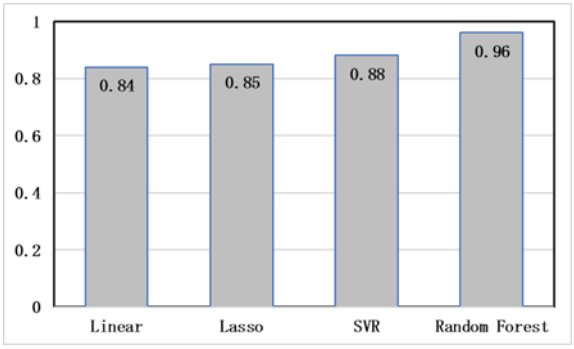


図５　OSAS 指標の回帰モデル精度比較（R^2 値）

Random Forest 回帰でのパラメータ感度解析結果を図６に示した。パラメータの感度は、S0＞alpha>PRRB の順番となった。気道面積(S0)と上顎の角度(alpha)が重要ファクターであることが分かった。

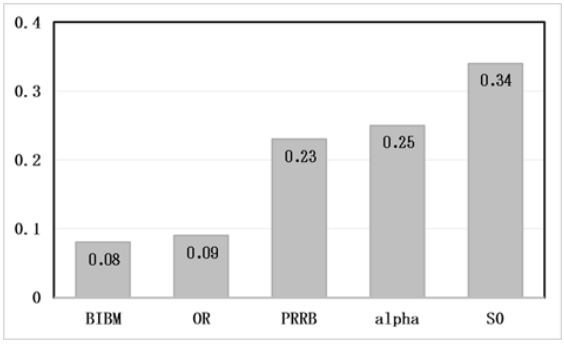


図６　呼吸疾患のパラメータ感度評価

今回の回帰モデルに使用したデータ数は 121 と少ないが、Random Forest 回帰を採用することで、R^2 値 0.96 の高精度なモデルを作ることができた。 このモデルを患者個々人の診断や将来リスク予測に役立てるためには、簡単な操作で予測結果が得られる手法が必要である。モデルのユニバーサル化と動的平行線図を使ったインターラクティブなモデル表現を行った。

人間の体重や身長などは正規分布に従っている。今回の口蓋形状も正規分布に従うと考えて 121 個のデータの共分散行列と平均値行列から 5 個の説明変数をすべて正規分布と仮定して正規乱数を 10000 組発生させユニバーサルモデルを生成した。モデル表現には、図 7 に示したインターラクティブな平行線図を使用した。(Python ライブラリの plotly を使用) 入力と出力を同一の平行線図に表示しており、灰色の帯状のバンドは、ユニバーサル化で生成した 10000 本の線の塊である。

操作の一例として、OSAS 指標が、大きい（レベルが 5,6）人は、S0 が 5cm2 以下で、alpha が 10° 以下といったように定量的な結果が得られる。

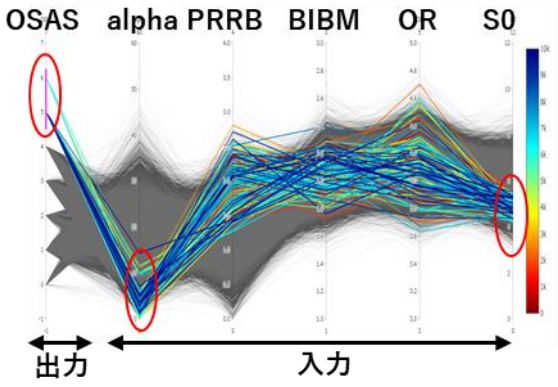


図 7 ユニバーサルモデルの動的平行線図表示

本手法を使うことで、上顎歯形から患者個々人の形状パラメータを抽出することにより OSAS 指標を簡単に予測することができる。 [4][5][6]

3　口蓋（上顎歯形）画像を使った疾患(OSAS, PVC, AF)予測手法の検討

3.1　口蓋画像と疾患ラベリング

第 2 章で、呼吸器系疾患 OSAS 指標は、口蓋形状パラメータで精度よく予測できることを示したが、心臓疾患系の PVC や AF 指標は予測できないことが判った。 噛み合わせ不良によるストレスが血中コレステロールを増加させ心筋梗塞や動脈硬化の引き金になることは知られている。今回、口蓋パラメータに頼らずに口蓋全体の画像を利用して、OSAS、PVC,AF 疾患を予測可能か検討した。口蓋画像には、正常者 4 枚、OSAS 患者 5 枚、PVC 患者 6 枚、AF 患者 3 枚に加え OSAS と AF 両方の疾患を持つ患者 4 枚の上顎歯形画像を使用した。図 8 に実際に使用した画像を示す。 () 内の 0～4 は、機械学習時の分類ラベルを示す。

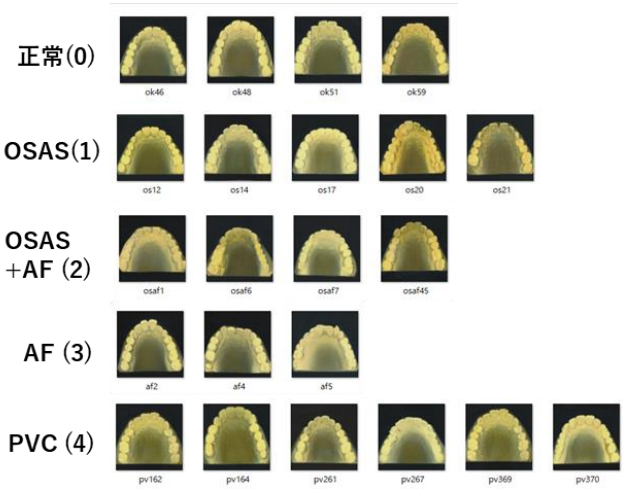


図 8　口蓋画像（上顎歯形）と疾患ラベリング

機械学習に使用する画像は、keras ライブラリのデータ拡張ツール ImageDataGenerator を使用した。X,Y シフトと回転を加え、各ラベルごとに約 100 枚の画像を生成した。機械学習に 450 枚を使用し、検証用にランダムに選んだ 50 枚を使用した。画像は、サイズが 128x128 のモノクロ画像に変換した。図 9 に正常画像 4 枚から 100 枚に拡張した画像例を示した。

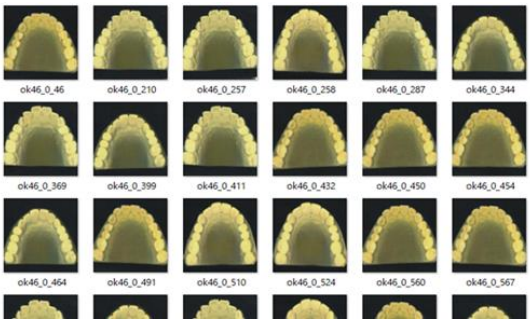


図９ ImageDataGenerator によるデータ拡張例

3.2 画像分類機械学習

画像の機械学習には、Sony が無償提供している NNC（Neural Network Console）を Windows-PC に導入し使用した。有益な例題がいくつか含まれており入力層の画像サイズを変更するだけで、簡単に学習とテストデータの推論結果を出すことが可能である。まず、手書き文字（数字）を 10 分類する例題に従い分類を試みた。
図１０に自動でレイヤーを生成する機能を ON にした手書き文字を分類した実績のあるネットワーク構造を示す。(ネットワークモデル：deep_MLP)

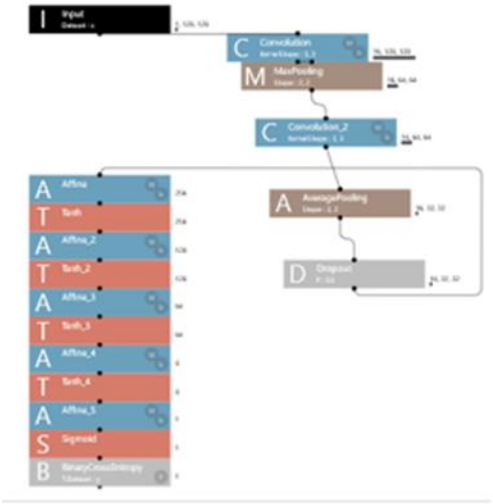


図１０ 手書文字を分類したネットワークモデル

Learning Curve は、EPOCH 数が 80 で収束したことを確認して、EPOCH 数 100 回目の結果を検証用に採用した。
検証結果の一部を図１１に示す。

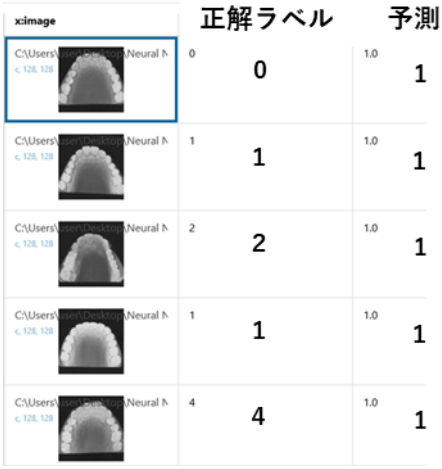


図１１ deep_MPL モデルの検証結果

正解ラベルと予測値の一致率は極めて低く、すべての検証画像をラベル 1（OSAS）と答え、正解確率は 0.24 であった。NNC の他の例題に ResNet(Residual Network)という勾配消失問題に対応したネットワークモデルが含まれており検討した。
ResNet の構造は複雑でその一部を図１２に示した。

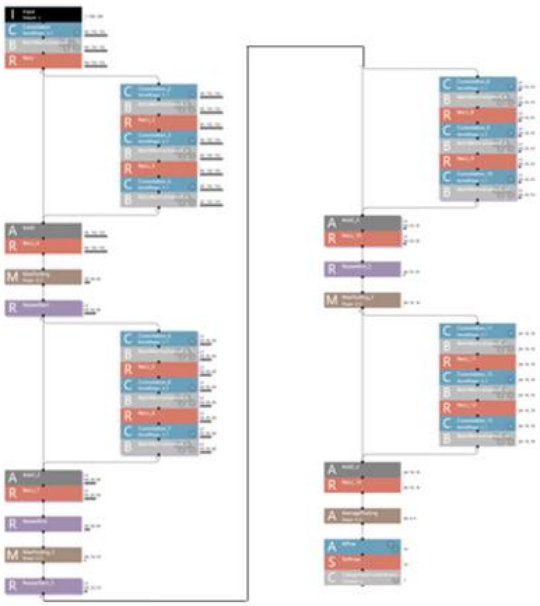


図１２ Residual Neural Network（ResNet）

EPOCH 数 5 付近で、Learning Curve は収束したことを確認して、EPOCH 数 10 回目の結果を検証用に採用した。
検証用 50 画像の分類結果を表 5 に示した。表中 y-label が答、y_0 が分類（0）である確率、y_1 が分類(1)である確率（以降同様）を表す。確率の高い箇所

に緑色でマークを施した。正解率は 50/50 で 100%であった。

表５　ResNet による疾患分類予測結果

x:image	y:label	y'__0	y'__1	y'__2	y'__3	y'__4
C:¥Users¥	2	6.94E-06	0.0076004	0.9848545	0.0015078	0.0060304
C:¥Users¥	2	5.16E-05	0.0115259	0.9870507	0.0005992	0.0007727
C:¥Users¥	3	4.02E-06	0.006379	0.0025584	0.9889092	0.0021494
C:¥Users¥	3	3.00E-06	0.0032929	0.0026086	0.9916661	0.0024294
C:¥Users¥	4	0.005795	0.0001915	0.0011195	3.79E-05	0.9928563
C:¥Users¥	3	7.82E-06	0.0002005	0.0001475	0.9994468	0.0001973
C:¥Users¥	1	0.0055172	0.981308	0.0096977	0.0001249	0.0033522
C:¥Users¥	3	0.0005057	0.0003137	0.0002227	0.9985905	0.0003674
C:¥Users¥	1	0.0063522	0.9922535	0.0005273	0.0003024	0.0005645
C:¥Users¥	0	0.9964904	0.0010515	0.0001038	1.22E-05	0.0023422
C:¥Users¥	3	0.0016792	0.0005083	0.0002535	0.9973441	0.0002149
C:¥Users¥	0	0.9955443	0.0043582	1.21E-05	1.31E-05	7.22E-05
C:¥Users¥	4	0.0073786	0.0009244	0.0012284	7.48E-05	0.9903938
C:¥Users¥	0	0.9916517	0.0015148	0.0001818	3.28E-05	0.0066189
C:¥Users¥	0	0.996531	0.0014316	0.0001375	1.34E-05	0.0018864
C:¥Users¥	3	1.71E-05	0.0089378	0.0037569	0.9886192	0.0036689
C:¥Users¥	2	2.84E-05	0.0076713	0.9918527	0.0002591	0.0001885
C:¥Users¥	3	0.0001852	8.17E-05	9.36E-05	0.9994742	0.0001653
C:¥Users¥	3	0.001497	0.0004075	0.0003089	0.9975764	0.0002102
C:¥Users¥	2	4.63E-06	0.0039285	0.9918335	0.0010007	0.0032327
C:¥Users¥	2	3.45E-05	0.0128989	0.9860363	0.0004192	0.0006112
C:¥Users¥	0	0.9970796	0.0025358	2.13E-05	9.17E-05	0.0002716
C:¥Users¥	1	5.20E-05	0.9760723	0.007654	0.000877	0.0153446
C:¥Users¥	4	0.001179	0.0480386	0.0109022	0.0048803	0.9349999
C:¥Users¥	1	0.0006241	0.9990692	0.0001464	4.02E-05	0.0001201
C:¥Users¥	3	0.0007766	0.0001644	0.0002463	0.9985002	0.0003125
C:¥Users¥	1	0.0064349	0.9836735	0.0026171	0.0035994	0.0036752
C:¥Users¥	4	0.0014638	0.013389	0.006424	0.0085835	0.9701397
C:¥Users¥	3	0.0023961	0.0004929	0.0003026	0.9964915	0.0003169
C:¥Users¥	3	0.0002988	0.0002142	0.0002102	0.9990566	0.0002202
C:¥Users¥	0	0.9981002	0.001592	1.11E-05	5.56E-05	0.000241
C:¥Users¥	0	0.998118	0.0004131	8.86E-06	0.0001399	0.0013201
C:¥Users¥	1	0.0028643	0.9961401	0.0005015	0.0001445	0.0003497
C:¥Users¥	3	0.0003563	9.69E-05	0.0001612	0.9989585	0.0004271
C:¥Users¥	3	1.07E-05	0.0005313	0.0004316	0.9987798	0.0002465
C:¥Users¥	1	0.0022357	0.9806019	0.0161421	0.0002654	0.0007548
C:¥Users¥	4	0.0002425	0.0216079	0.0088094	0.0063203	0.9630199
C:¥Users¥	0	0.9470006	0.0158844	0.0074093	0.0207419	0.0089639
C:¥Users¥	1	0.0017026	0.9965179	0.0013549	0.0003086	0.000116
C:¥Users¥	1	0.0042288	0.9791175	0.0129891	5.38E-05	0.0036108
C:¥Users¥	3	1.09E-05	0.0061652	0.0036008	0.9877613	0.0024618
C:¥Users¥	0	0.9956488	0.0041538	1.56E-05	4.05E-05	0.0001412
C:¥Users¥	2	0.0004256	0.0016154	0.9913625	0.0030006	0.0035959
C:¥Users¥	4	0.000454	0.0062149	0.004514	0.0175403	0.9712768
C:¥Users¥	4	8.33E-05	0.0065928	0.0040237	0.0014321	0.9878681
C:¥Users¥	3	2.58E-05	0.0010482	0.0002997	0.998345	0.0002813
C:¥Users¥	4	0.0002373	0.0129289	0.0082196	0.0023322	0.9762819
C:¥Users¥	3	4.48E-05	0.0007384	0.000309	0.9984901	0.0004177
C:¥Users¥	2	1.93E-06	0.0013702	0.9878238	0.00202	0.0087841
C:¥Users¥	2	0.0005501	0.0061741	0.9920075	0.000108	0.0011604

ResNet の分類能力は従来の CNN に比べ非常に高いものと思われる。今回の結果では混同行列の対角成分以外はほぼゼロである。対角成分を抜き出し、自己予測確率を求めた結果を図１３に示した。

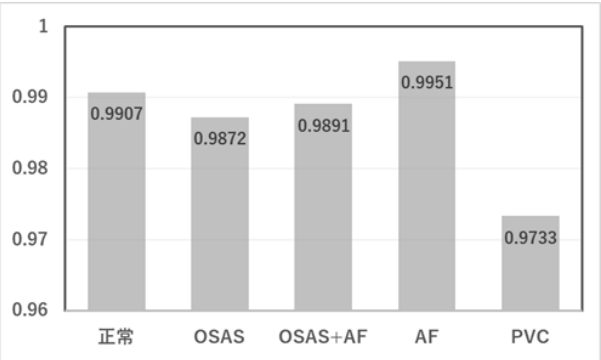


図１３ ResNet による分類の自己予測確率

検証画像 50 枚の分類確率は 97%以上であった。学習用元画像数が 22 枚で、500 枚に画像拡張を行い 450 枚を機械学習に使用したため、高精度に分類できたのかは不明である。　今後、新たな患者の歯形写真を多量に収集して分類検証してみる必要がある。

4　結果

睡眠時無呼吸症候群のリスクを示す OSAS 指標を口蓋形状の 5 パラメータから予測する高精度モデルを生成した。モデルによると口蓋面積(S0)と上顎の角度（alpha）が主要パラメータでることが判った。生成した予測モデルを可視化するため各形状パラメータが正規分布することを仮定して説明変数を正規乱数で発生させユニバーサル化を行った。ユニバーサルモデルを可視化するため説明変数と目的変数を平行線で結び、変数間の関係を線の色で表示する動的平行線図プロットを生成した。この動的平行線図を使用することで、睡眠中無呼吸リスクが高い（レベル 5,6）人の特徴は、S0<5cm2 で alpha<10° であることが予測された。

一方で心室期外収縮（PVC）指標や心房細動（AF）指標は、口蓋形状パラメータではモデル化できなかった。口蓋画像を使って、OSAS、PVC、AF 症状を予測する手法を検討した。結果、ResNet (Residual Neural Network)による機械学習で分類が可能であることが検証できた。

5　考察

OSAS 指標には BMI（肥満度）や Neck circumference（首回）などの後天的ポイントも含まれている。
一方で、今回のモデルでは口蓋形状という先天的な要素を多く含むパラメータを使って高精度な予測モデルを生成することができた。逆に言えば、睡眠時無呼吸症候群（OSAS）は先天的性であり将来リスクを予測することが可能と思われる。
個々人の疾患リスクレベルをより簡便に予測するためには、口蓋形状パラメータの抽出（3 次元計測）という手間のかかる方法より、歯形写真（2 次元データ）を ResNet 等の高精度な画像分類器を使って推論予測する方法が望ましい。今後、画像からより精度の高い分類ができるようにデータを多く収集する予定である。

参考文献

[1] 池原　晃生
咀嚼筋に対する交感神経刺激の影響
阪大歯学誌（1993)、pp232-260

[2]姜　英雄・豊田　博紀・斉藤　充・佐藤　元
生命歯科医学のカッテイングエッジ、三叉神経中脳核ニューロンにおけるインパルストラッフィング
大阪大学出版会 2（2007)、pp157-166

[3]小林　建三郎・池田　隆徳
虚血性心疾患に伴う心室不整脈に対するリスク層別化と薬物療法
JPN.J.ELECTROCARDIOLOGY Vol.34 No.2(2014)
PP98-107

[4]本田　啓介・中野　純司
三次元平行座標プロット
統計数理（2007）　第 55 巻 第 1 号、pp69-83

[5] Takayuki Ito, Ashmil Kumar, Karsten Klein, J. Kim
“High-dimensional data visualization by interactive construction of low-dimensional parallel coordinate plots”
Journal of　Visual Languages & Computing, Volume 43, 2017, pp1-13

[6]馬場　嘉朗・馬場　達朗・酒井　経雄
口蓋形状から呼吸器疾患レベルを予測するモデルの一般化*
*計測機器自動制御学会に投稿中

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-057-15 (11/20)

Improving Conditional-GAN using Unrolled-GAN for the Generation of Co-speech Upper Body Gesture

Bowen Wu[1,2*] Chaoran Liu[3] Carlos T. Ishi[2,3] Hiroshi Ishiguro[1,3]

[1] Osaka University
[2] RIKEN
[3] ATR

Abstract: Co-speech gesture is a crucial non-verbal modality for humans to express ideas. Social agents also need such capability to be more human-like and comprehensive. This work aims to model the distribution of gesture conditioned on human speech features for the generation, instead of finding an injective mapping function from speech to gesture. We propose a novel conditional GAN-based generative model to not only realize the conversion from speech to gesture but also to approximate the distribution of gesture conditioned on speech through parameterization. Objective evaluation show that the proposed model outperforms the existing deterministic model in terms of distribution, indicating that generative models can approximate the real patterns of co-speech gestures more than the existing deterministic model. Our result suggests that it is critical to consider the nature of randomness when modeling co-speech gestures.

1 Introduction

Human-like robots and virtual agents have human appearance. Therefore, they are expected to use both verbal and non-verbal behaviors to express themselves like humans during the interaction with humans. One crucial non-verbal behavior that can be observed is the hand gestures[1][2]. These spontaneous hand movements accompany speech to complement or even to supplement the information relayed in a speaker's speech[3]. Modeling the relationship between gestures and speech will provide a useful tool for expressing ideas comprehensively for human-like agents and promoting humans' perception.

At the early stage, robot gestures were only designed for a few pre-defined scenarios[4]. For the automatic generation, the first trial was the so-called ruled-based method. A set of human gesture patterns is recorded as sequences of joint data, and their occurrence was statistically studied in the relationship with the lexicon. These results were then summarized as a bunch of rules to decide which gesture to select from the recorded database[5]. An advanced rule-based method was proposed to separately model different parts of the human body to generate different combinations as a whole[6]. The shape of the gesture was constrained on those appearing in the collected data in these studies.

Beyond writing rules, data-driven statistical models were also adopted. The relation between iconic gestures and lexicon was automatically learned from the corpus using a Bayesian Decision Network[7]. Dynamic Bayesian Network was also utilized to model several meaningful behaviors (e.g., nod) while considering the synchronization with speech[8]. The relationship between the prosodic feature of speech and rhythmic gesture was modeled using modified hierarchical factored conditional restricted Boltzmann machines(HFCRBMs)[9]. Taking both prosody and text as input, a probabilistic model that maps the concept extracted from text using WordNet to gesture clusters, integrated with the superimposed beat gesture, was proposed[10]. However, the methods proposed in these studies require an elaborate feature engineering on human data.

Since human data analysis is tedious and time-consuming, machine learning and deep learning approaches have been utilized to automatically map speech to gesture. Hidden Markov Model was used to generate pointing gesture from audio features[11]. The effectiveness of recurrent models such as gated recurrent unit(GRU) and long-short term

*連絡先： 大阪大学大学院基礎工学研究科
〒 560-0043 大阪府豊中市待兼山町 1 丁目 3
E-mail:wu.bowen@irl.sys.es.osaka-u.ac.jp

memory(LSTM) on mapping Mel-Frequency Cepstrum Coefficients(MFCC) features of speech to gestures has been analyzed[12][13]. In [13], a bi-directional LSTM network learned the mapping from MFCC features to 3D joint coordinates of the skeleton from the dataset collected using MOCAP toolkit. The text was also used as input to generate meaningful gestures by sequence to sequence neural networks[14]. In [15], the text was encoded using bidirectional encoder representations from transformers(BERT) to be concatenated with audio features to generate gesture sequences. Due to the high dimension characteristic of human motion, Denoising autoencoder(DAE) was employed to reduce the dimension of motion to help the neural network to generalize[16]. [17] made use of labeled gesture phase information to constrain the dynamic of generated gestures. The individual style was concerned with separately training different neural networks with L1 distance and discriminative loss on a particular person's data[18]. A style transfer model aiming at generating gestures with personal style for others' voice was also proposed[19].

The generation methods mentioned above are based on a strong assumption: the mapping from speech to gesture is injective, i.e., only one gesture can be generated by these models for one speech segment. On the contrary to this assumption, there are alternatives for almost any gestures. There are numerous examples of this phenomenon, such as using left or right or both hands, hand at different height and radius, and so forth. Additionally, a human may perform new gestures that have never been performed before for a particular speech. We treat this randomness as an essential nature of co-speech gestures. As a result, we aim to design a generative model to realize the randomness of co-speech gestures.

Relatively few studies have walked into this field. [20] uses MoGlow to generate gestures while controlling the height, radius, or speed by inputting a controlling variable, realizing the gestures' variation. However, they rely on a manipulated signal, whereas we leave this to be entirely random by sampling a random variable from a prior distribution.

Inspired by the success of generative adversarial nets(GAN) on generation tasks, we proposed a GAN-based generative model to realize the conversion from speech to gesture while preserving the randomness. To optimize the model, we designed a discriminator to give dynamic feedback on the generator results. Mode collapse, a common failure in GAN training, is minimized by using the algorithm of unrolled generative adversarial nets(Unrolled-GAN). We experimented with our model on a Japanese speech/gesture dataset. The objective evaluation showed that the proposed model can better approximate real gesture distribution. User studies also confirmed the proposed model's effectiveness and showed no significant difference between the generated results of the proposed model and the ground truth(original human motions).

The contributions of this work are three-folded: (1) We proposed a novel deep-learning-based generative model for co-speech gesture generation. (2) We proposed a strategy for changing gesture patterns by manipulating the randomly sampled vector, and improved the performance. (3) We confirmed that the proposed model outperformed the existing deterministic model through experiments.

2 Method

2.1 Problem Formulation

The notations used in the rest of this article are denoted as follows: for a speech segment with length T, the features extracted from the audio signal is $\mathbf{s} = [s_t]_{t=1:T}$. The sequence of the absolute position of each joint in the 3-dimensional space is $\mathbf{j} = [j_t]_{t=1:T}$, where $j_t = [x_t^i, y_t^i, z_t^i]_{i=1:K}$, K is the total number of joints. The problem of generating gesture from speech then can be defined as to parameterize a model G by a parameter set θ such that $\mathbf{j}^{(m)} = G_\theta(\mathbf{s}^{(m)})$. Furthermore, we aim to model the conditional distribution $X_\mathbf{j}$ conditioned on the distribution $X_\mathbf{s}$. To achieve this, a random variable z sampled from a normal distribution $N(0, 1)$ will be taken as input by the model. Thus, the problem is to find a paramter set θ such that $p(\mathbf{j}|\mathbf{s}) = G_\theta(z|\mathbf{s}), \mathbf{j} \sim X_G, \mathbf{s} \sim X_\mathbf{s}, z \sim N(0, 1)$. The error between the parameterized distribution and real distribution is defined as $d(p(\mathbf{j}|\mathbf{s})_{\mathbf{j}\sim X_G}, p(\mathbf{j}|\mathbf{s})_{\mathbf{j}\sim X_j})$ to optimize G_θ. A model D paramterized by ϕ will be optmized to be the measurement of this error.

2.2 Feature Extraction

Motion features. The motion data in the corpus is composed by joint rotation and offset of each joint. We used the protocol provided in [16] to convert the joint's rotation values to absolute position values(APV) in three-dimensional(3D) space to meet

our problem established in section 2.2. As the movements are concentrated on the upper body part, we only used the upper body's APV as the training label.

Speech features. The speech features used in this work are the prosodic features. Prosodic features include fundamental frequency(f0), intensity, and their first derivative and second derivative, as they reflect the rhythm of speech. Although MFCC features are frequently used in automatic speech recognition(ASR), they are not preferred here because the extracted features will be used as conditions in model D. Low dimensional features are expected to yield better results than a high dimensional one since high-dimension conditions will drastically reduce the number of samples included in that condition. An opensource audio signal processing package Parselmouth was used to extract intensity and fundamental frequency from the speech signal. First, using a window size of 10 milliseconds and hop length of 5 milliseconds, 200 frames of every second feature are extracted. Then, every ten frames' feature are averaged to be 20 frames per second(fps) to match the fps of motion data.

2.3 Methodology

Fully-connected layer(FC). FC is essential in the deep learning area. It consists of a weight matrix $W \in \mathbb{R}^{m \times n}$,and a bias vector $b \in \mathbb{R}^n$, where m is the input's dimension to the FC, n is the output's dimension of the FC. The computation inside FC is defined as equation (1).

$$A = x \cdot W + b \tag{1}$$

where $x \in \mathbb{R}^m$, $A \in \mathbb{R}^n$.

Bidirectional Long-Short Term Memory. Bi-LSTM is a general solution for sequence data modeling. It utilizes the past and future information to compute the output for the current timing. Considering that past and future information in the speech has influences on the current motion, we used bi-LSTM to capture the information flowing through time.

Generative Adversarial Nets(GAN). The essence of GAN is a min-max game between a generator and a discriminator. While the discriminator is being optimized to recognize whether its inputs are sampled from real data or generated fake data by the generator, the generator is trying to deceive the discriminator by learning to generate data that resembles real data. This adversarial system will reach the nash equilibrium in the end after the generator learned to generate real data. Intuitively, this is equivalent to that the generator approximates the real data distribution. [21] confirmed this hypothesis by proving that the generator is trying to minimize the Jensen–Shannon divergence between the generated distribution and the real data distribution when the discriminator is optimal.

Conditional-GAN(CGAN). CGAN can generate an entity in a specific category[22]. It adds the same conditional labels for both generator and discriminator. Mathematically, the distribution to which the GAN's generator is trying to approximate is replaced by the conditional distribution conditioned on a specific category. [23] used CGAN to model head motion with speech as conditional input.

Unrolled-GAN. Mode collapse is a common failure in GAN training, i.e., the generator outputs identical results for any noise vector from the prior. By unrolling the discriminator, unrolled-GAN allows the generator to "look into the future" to prevent the discriminator from overfitting on a specific training sample, reducing the mode collapse[24].

Proposed method. Our proposed model utilizes the architecture of CGAN, where the speech features are used as a condition. An overview is shown in Figure 1 and 2. During the generating phase, a randomly sampled vector(noise vector) z from the Gaussian prior is repeated to the time step length of speech features. Then, z and the speech feature s are concatenated and fed into a two-layer bi-LSTM. A sequence-wise fully-connected layer then takes the output of previous layers and outputs a sequence of vectors indicating each joint's absolute positions in the 3D space. The reason for repeating a fixed-length random vector instead of sampling a sequence length wise random vector is that we want to maintain the output motion's consistency along the entire sequence. To optimize the generator, we optimize a discriminator simultaneously to compute the error between the generated distribution and the real distribution conditioned on speech features. A vector of motion sequence and the corresponding speech features are concatenated and fed into a two-layer bi-LSTM layer. The output is squashed between 0 and 1 through a sigmoid function, indicating whether the input motion is real and corresponding with the speech features. Instead of outputting only one scalar for the whole sequence by the discriminator, we prefer to output one scalar for each time step. The reason for doing so is that though

LSTM is claimed to be capable of capturing long term dependencies, in practice, the effectiveness decreases when the sequence grows relatively long. The equation for optimizing generator and discriminator is defined as equation (2).

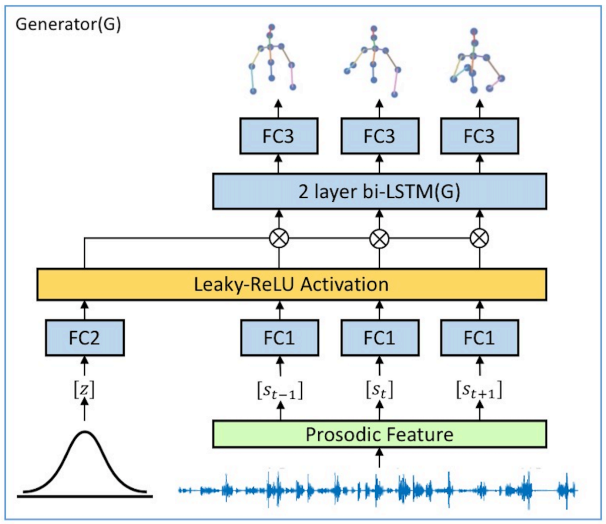


Fig. 1: The generator of proposed model. The output of FC2 is manipulated to the same time steps with **s**, and then be concatenated.

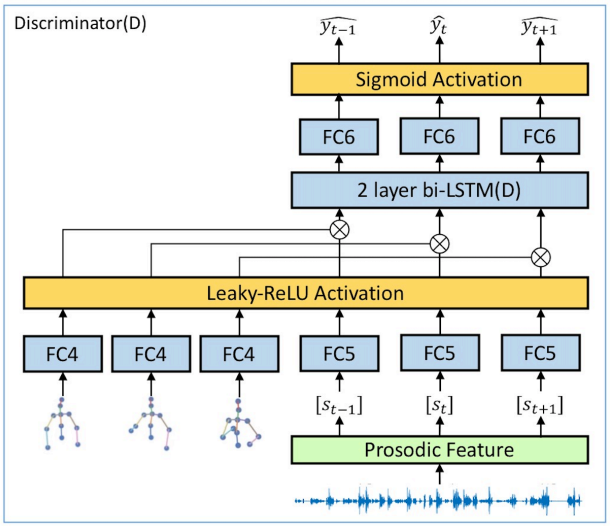


Fig. 2: The discriminator of proposed model. The outputs of FC4 is of the same length as **s**. The concatenation follows the order of the sequence.

$$\max_D \min_G \frac{1}{m}\sum_{i=1}^{m} log(D(j^{(i)}, s^{(i)})) - log(D(G(z, s^{(i)}), s^{(i)})) \quad (2)$$

where m is the number of samples.

Algorithm 1 Algorithm for training.

Require: α, the learning rate. k_{unroll}, the unrolling steps. m, the batch size. *iteration*, the number of training iterations.

Require: ϕ_0, initial discriminator parameters. θ_0, initial generator parameters.

for 0 to *iteration* **do**
  Sample $\{\mathbf{j}^{(i)}, \mathbf{s}^{(i)}\}_{i=1}^{m} \sim (X_\mathbf{j}, X_\mathbf{s})$ a batch from the real data
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim N(0,1)$ a batch from the prior
  $g_\phi \leftarrow \nabla_\phi[\frac{1}{m}\sum_{i=1}^{m} log(D_\phi(j^{(i)}|s^{(i)})) - log(D_\phi(G_\theta(z^{(i)}, s^{(i)})|s^{(i)}))]$
  $\phi \leftarrow \phi + \alpha \cdot g_\phi$
  $backup_\phi \leftarrow \phi$
  for 0 to k_{unroll} **do**
    Sample $\{\mathbf{j}^{(i)}, \mathbf{s}^{(i)}\}_{i=1}^{m} \sim (X_\mathbf{j}, X_\mathbf{s})$ a batch from the real data
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim N(0,1)$ a batch from the prior
    $g_\phi \leftarrow \nabla_\phi[\frac{1}{m}\sum_{i=1}^{m} log(D_\phi(j^{(i)}|s^{(i)})) - log(D_\phi(G_\theta(z^{(i)}, s^{(i)})|s^{(i)}))]$
    $\phi \leftarrow \phi + \alpha \cdot g_\phi$
  end for
  Sample $\{\mathbf{s}^{(i)}\}_{i=1}^{m} \sim X_\mathbf{s}$ a batch from the real data
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim N(0,1)$ a batch from the prior
  $g_\theta \leftarrow \nabla_\theta[\frac{1}{m}\sum_{i=1}^{m} log(D_\phi(G_\theta(z^{(i)}, s^{(i)})|s^{(i)}))]$
  $\theta \leftarrow \theta - \alpha \cdot g_\theta$
  $\phi \leftarrow backup_\phi$
end for

On the other hand, a common failure during GAN training is mode collapse, i.e., the generator outputs identical results for any noise vector from the prior. In practice, we found that the algorithm proposed in unrolled-GAN successfully reduced the mode collapse that appeared in our experiment setting. However, since we used the LSTM layer, the original unrolled-GAN algorithm will tremendously increase the training time. As a result, we simplified the algorithm and observed that a similar result is achieved in our experiment with shorter training time. Note that we are not claiming that the original algorithm is replaceable by this simplified version. The proposed algorithm is shown in Algorithm 1.

In our experiment, we found that each noise vector

corresponds with a particular pattern of motion, i.e., motions with the same pattern are generated when using the same noise vector throughout the sequence, a result that is not desirable. To increase variations of the generated motions, we proposed a strategy of generating variating noise vectors for a certain length of speech sequence. The algorithm is shown in Algorithm 2.

Algorithm 2 Algorithm for generating noise vectors.

Require: T, time steps of speech features. F, time steps of repeating the same noise vector.

$K \leftarrow \text{ceil}(T/F)$
$zs \leftarrow [z, ..., z]_F \sim N(0, 1)$
for 0 to K **do**
  Sample $P \sim Uniform(0, 1)$
  if $P > 0.5$ **then**
    append $zs_{:-F}$ to zs
  else
    $zs_1 \leftarrow [z_1, ..., z_1]_F \sim N(0, 1)$
    append zs_1 to zs
  end if
end for

3 Experiment and results

3.1 Corpus

We evaluated our model on the dataset proposed in [25], in which pairs of recorded audio and motion are provided. The content is an undergraduate student answering questions in Japanese like in an interview while standing and gesturing. The motion data was recorded using a motion capture studio. The motion data files contain information of offset and rotation of each joint, from which each joint's absolute position can be derived. The audio is saved as WAV files(sampling rate 22050 Hz, 16bits). There are 1049 sentences in this dataset, 298 minutes, 68.41% are metaphoric gestures, 23.73% are beat gestures, and others are iconic and deictic gestures.

3.2 Baseline

To compare the proposed model with the deterministic generation method, the model proposed in [16] is used as a baseline. We use the protocol provided by the authors and reproduced the reported result. We cut the upper body motion generated using the baseline model in order for comparison. Since the dataset is already split for the baseline model into train, development, and test set, we use the split test set for evaluation. There are 45 samples in the test set.

3.3 Training setting

Numerous works for gesture generation cut gesture sequence into several slices to approximate the effect of data augmentation. Instead, we used the entire sequence of speech and motion as samples. We saved the trained model with every ten iterations and generated some samples using speech utterances in the test set. By viewing the quality of these generated results, we finally chose the generator of the 1000 iteration.

3.4 Quantitive Evaluation

Numerical evaluation metric. It is common for a deterministic model to use L1 distance or average position error(APE) to evaluate the generated results. Since our motivation is to model the distribution of gestures, it is not appropriate to evaluate the precision of generated key-points compared with the ground truth. Instead, kernel density estimation(KDE) is a useful tool for approximating the distribution of data, as also used in [21] for image generation and in [23] for head motion generation. The output of KDE is the log-likelihood of input samples based on the fitted density function using samples. In this work, we use generated gesture sequences in the test set to fit the density function and use the ground truth as input to KDE. Therefore, the larger the output value towards 0, the similar the generator fits the real data distribution.

By using the algorithm 2, we generated one motion sequence for every speech in the test set. The generated motions are used to fit a distribution. The optimal bandwidth in the KDE model is obtained using a grid search with 3-fold cross-validation. Then, the log-likelihood of real motions in the test set is calculated using the fitted distribution. We also studied how F in the algorithm 2 affects the results. The results are shown in Table 1. The values are the average of 5 times calculation.

Table. 1: Quantitive comparison between models. Ground Truth is the log-likelihood of real motions in the test set in the KDE distribution fitted using the ground truth itself, indicating the best results that can be approached. * used repeated noise vector to generate motions. ** jointly used the proposed model and the proposed algorithm 2.

Model	Log-likehood	SE
Ground Truth	*-29.98*	*1.03*
Baseline[16]	-508.82	87.61
CGAN*	-245.67	44.72
Unrolled-CGAN*	-118.91	17.03
Proposed(F=20)**	-177.86	29.36
Proposed(F=30)**	-161.30	26.78
Proposed(F=40)**	**-107.58**	**15.21**
Proposed(F=50)**	**-107.98**	**15.77**
Proposed(F=60)**	-126.20	19.01

4 Discussion

4.1 Why Generative Models Perform Better?

Modeling the distribution. L1-distance or L2-distance are usually used as the loss function for training deterministic models. A potential risk of doing so is that if there are two similar speech utterances paired with different gestures in the training set, an average value of these gestures will be the optimal solution for these utterances, and this average value is likely not being an available gesture at all. This risk may either cause the generated motions to be not human-like or small range gestures. On the other hand, generative models do not have such a problem since generative models aim to approximate the likelihood of real data. The generated gestures are more reliable and have the potential to cover a broader range of gestures. This difference can be seen from the comparison between baseline and other generative models in Table 1.

4.2 Detailed performance analysis

Motion dynamics(i.e., velocity) are imperative to human perception. Since we are aiming at modeling the distribution of human gesture, one reason that the proposed model outperforms the baseline model is assumed to be that the velocity distribution of the motion generated by the proposed model is more similar with the ground truth than the baseline model. By plotting the histogram of average velocity of all joints, shoulder, and wrist, we confirmed this assumption by observing that the histograms of proposed model are more similar with the ground truth compared with the baseline, as shown in Fig. 3, 4, and 5.

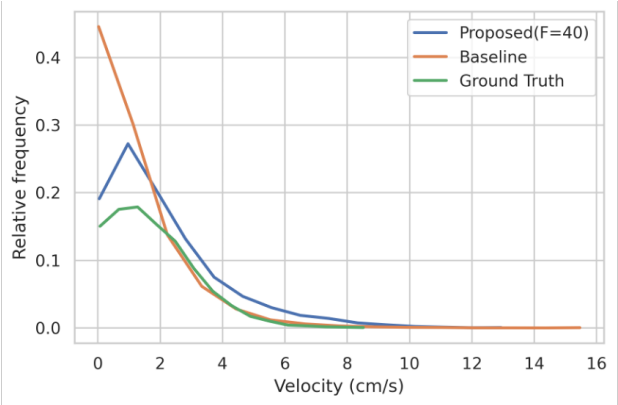


Fig. 3: Histogram of average velocity of all joints.

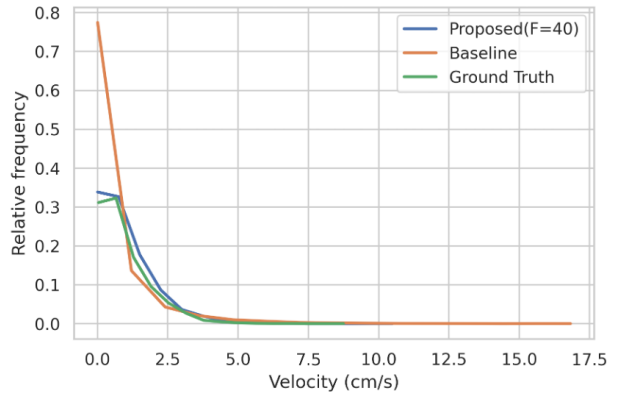


Fig. 4: Histogram of shoulder velocity of all joints.

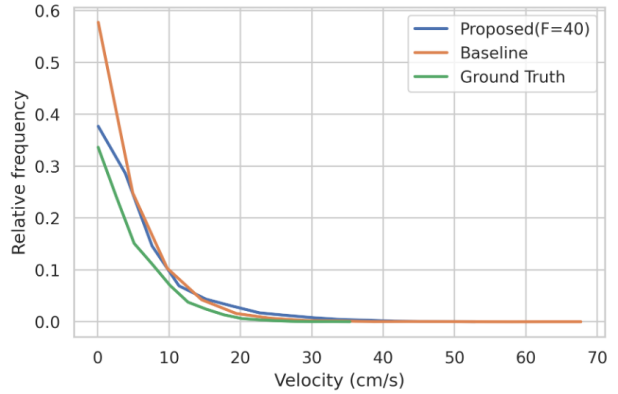


Fig. 5: Histogram of wrist velocity of all joints.

5 Conclusions

It is crucial for human-like agents to gesture well to be comprehensive and expressive. We present a model for producing co-speech gestures by modeling the conditional distribution of gesture conditioned on speech features. Improved by unrolled-GAN and our proposed algorithm, the proposed model outperforms the existing deterministic model in objective evaluation. Our work provides a powerful tool for Human-like agents to express thoughts, enhancing human perception. Moreover, probabilistic modeling's success reveals that future research in this field should focus more on gesture distribution. Human-like agent is expected to realize complicated interaction with human. However, without the ability to gesture well, they are inexpressive to be understood or empathized with by humans. Though our gesture generation model performs better in terms of gesture distribution, the lack of semantics(i.e., meaningful gesture) is still a considerable obstacle to perfectly model human gesture, which requires further research.

Acknowledgement

This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas JP20H05576.

References

[1] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[2] David McNeill. *Gesture and thought*. University of Chicago press, 2008.

[3] Jana M Iverson and Susan Goldin-Meadow. Why people gesture when they speak. *Nature*, Vol. 396, No. 6708, pp. 228–228, 1998.

[4] Justine Cassell. A framework for gesture generation and interpretation. *Computer vision in human-machine interaction*, pp. 191–215, 1998.

[5] Björn Hartmann, Maurizio Mancini, and Catherine Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *International Gesture Workshop*, pp. 188–199. Springer, 2005.

[6] David Baumert, Shunsuke Kudoh, Masaru Takizawa, et al. Design of conversational humanoid robot based on hardware independent gesture generation. *arXiv preprint arXiv:1905.08702*, 2019.

[7] Kirsten Bergmann and Stefan Kopp. Gnetic–using bayesian decision networks for iconic gesture generation. In *International Workshop on Intelligent Virtual Agents*, pp. 76–89. Springer, 2009.

[8] Najmeh Sadoughi and Carlos Busso. Speech-driven animation with meaningful behaviors. *Speech Communication*, Vol. 110, pp. 90–100, 2019.

[9] Chung-Cheng Chiu and Stacy Marsella. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*, pp. 127–140. Springer, 2011.

[10] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, Vol. 3, No. 4, pp. 3757–3764, 2018.

[11] Mehmet Emre Sargin, Oya Aran, Alexey Karpov, Ferda Ofli, Yelena Yasinnik, Stephen Wilson, Engin Erzin, Yücel Yemez, and A Murat Tekalp. Combined gesture-speech analysis and speech driven gesture synthesis. In *2006 IEEE International Conference on Multimedia and Expo*, pp. 893–896. IEEE, 2006.

[12] Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 93–98, 2018.

[13] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 79–86, 2018.

[14] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning

of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4303–4309. IEEE, 2019.

[15] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. *arXiv preprint arXiv:2001.09326*, 2020.

[16] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97–104, 2019.

[17] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*, pp. 1–10. 2019.

[18] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3497–3506, 2019.

[19] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. *arXiv preprint arXiv:2007.12553*, 2020.

[20] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, Vol. 39, pp. 487–496. Wiley Online Library, 2020.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[22] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[23] Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6169–6173. IEEE, 2018.

[24] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[25] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction*, pp. 198–202. Springer, 2017.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-057-16 (11/21)

RoboCupサッカーおける秘匿通信のための スペクトル拡散を用いた音声電子透かし法の提案

A Digital Speech Watermarking Method Using Spread Spectrum for Secret Communication in RoboCup Soccer

坪倉和哉[1]　久保谷空史[1]　舘拓磨[1]　小林邦和[1]
Kazuya TSUBOKURA[1], Takashi KUBOYA[1], Takuma TACHI[1], Kunikazu KOBAYASHI[1]

[1] 愛知県立大学 情報科学部
[1] School of Information Science and Technology, Aichi Prefectural University

Abstract: 現在 RoboCup の標準プラットフォームリーグ (SPL) では, ロボット間のコミュニケーション手段として, 主に無線通信が用いられている. しかし, 人間のサッカーにより近づけていくためには, 無線通信に代わるコミュニケーション方法の開発が必要となる. 本研究では, 無線通信に代わるものとして, 音声を用いたコミュニケーションについて検討する. 具体的には, 自然言語の合成音声に対し, スペクトル拡散法を用いて秘匿データを埋め込む手法を提案する. そして, ロボット間の距離と透かし情報の大きさを変更した場合でのビット誤り率の比較検証を行った.

1　はじめに

RoboCup では 2050 年までに, サッカーワールドカップの世界チャンピオンに勝てるロボットサッカーチームを作ることを目指している [1]. RoboCup サッカーの標準プラットフォームリーグ (Standard Platform League: SPL) では, 全チームが同一のロボットを用いてソフトウェアシステムの優劣を競うリーグである [2]. 現在は, SoftBank Robotics 社のヒューマノイドロボット NAO が用いられている.

現在 RoboCup では, ロボット間コミュニケーションに無線通信を用いることが主流である. しかし, 将来人間と対戦する上で, より人間と同じ条件にするためには, 無線通信に代わるコミュニケーション方法が必要となる. とりわけ, SPL では, 無線通信におけるパケット量を制限したり, ホイッスルの音源定位に関するテクニカルチャレンジが行われたりするなど, 無線通信ではなく音を利用したロボット間コミュニケーション技術の開発が求められている.

一般的に, 音を利用したコミュニケーションには, モールス信号のような符号的な音や自然言語の発話が考えられる. 前者は, 人間には識別が困難である. エンターテインメント性を考慮すると, 人間のサッカーに見られるような, プレイヤーと観客とのインタラクションは生まれにくく, 観客を置き去りにしてしまいかねない. 一方, 自然言語の発話によるコミュニケーションは, 観客にもロボット同士のコミュニケーションが理解でき, ロボットが何を考えて動いているかが推測できる. しかし, 単位時間あたりに伝えられる情報量が少ないということや, 相手チームにも対話内容が伝わってしまうという危険性がある.

本研究では, この問題を解決するために, 自然言語の発話の中に, 秘匿データを埋め込む手法を提案する. 具体的には, スペクトル拡散法を用いて合成音声にビット列を埋め込む. これにより, 観客は合成音声からロボット同士の対話を観察でき, ロボットは埋め込まれたビット列から命令や意思疎通を行うことができる. さらに, 相手チームからはビット列が読み取られないよう, 秘匿性を担保することも可能となった.

2　先行研究

2.1　スペクトル拡散

スペクトル拡散 (Spread Spectrum : SS) 方式とは, 信号の変調方式のことであり, 携帯電話や無線 LAN などの無線通信や, 音響電子透かしなどで用いられている [3][4][5]. SS 方式では, 通常の情報伝送に必要な帯域幅を大幅に超えたより広い帯域幅を持つ信号を用いて通信を行っており, 秘匿性や耐ノイズ性といった通信を行うにあたって非常に優れた特性を有している.

SS 方式には, 直接拡散 (Direct Sequence : DS)/SS 方式や周波数ホッピング (Frequency Hopping : FH)/SS 方式などがある. DS/SS 方式では, 信号強度を常に弱い状態で保てるため, FH/SS 方式よりも秘匿性が高いと言える. 本研究の趣旨として, 秘匿性がある方が望ま

しいと考えられることから,DS/SS 方式を用いた音声通信について考察する.

図 1 に DS/SS 方式の送信機, 受信機の構成におけるデータの流れを, 図 2 に各状態でのデータの様子を示す. 最初に, データを BPSK(Binary Phase Shift Keying) や QPSK(Quadrature Phase Shift Keying) を用いて変調する. 高速通信には向かないものの, 伝送品質が劣化しにくいことから, 本研究では BPSK 変調を用いることとする. 次に, 拡散系列を用いてデータの拡散を行い, 送信機から受信機へ拡散したデータを送信する. 受信機では, 受信したデータに対し逆拡散を行うことで, 拡散を行う前のデータ得ることができる. 最後に, 復調器でデータの復調を行うことで, データの復元ができる.

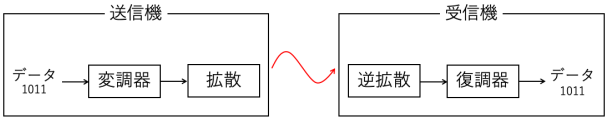


図 1: DS/SS 方式におけるデータの流れ

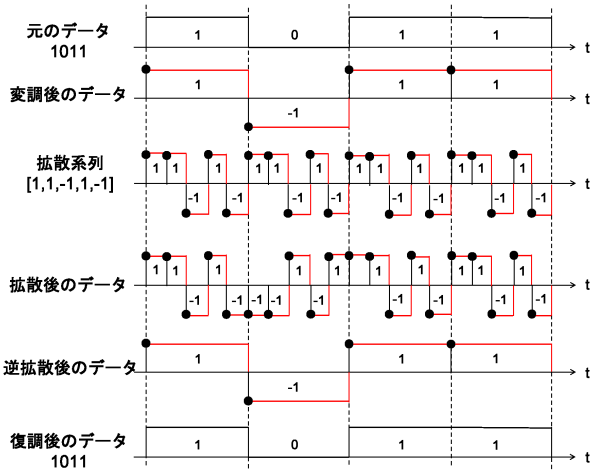


図 2: DS/SS 方式における各状態でのデータの様子

2.1.1 逆拡散

逆拡散では, 図 3 のように重み係数として拡散系列を与えた相関器を生成し, 受信信号との相関を求めている. 相関が高い程出力される値の絶対値が大きくなる. そのため, 図 4 のように相関器出力の各ピークを取り出すと拡散前の変調したデータを復元できる.

2.1.2 レイク合成

スペクトル拡散方式では, 伝送信号の波形が歪になる選択性フェージングを解決することは可能である. し

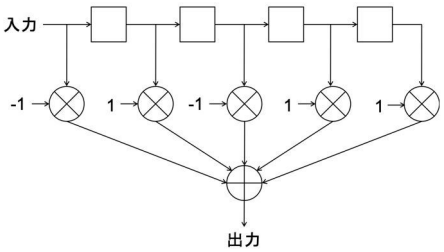


図 3: 相関器の入出力のイメージ

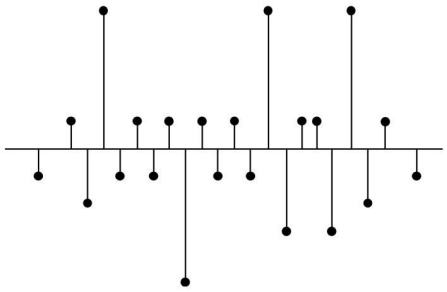


図 4: 相関器の出力例

かし, 伝送信号波形の振幅が小さくなるフラットフェージングの影響を受けてしまう. そのため, 必要となるのがレイク合成である [6]. レイク合成では, スペクトル拡散方式で分離されたマルチパス信号を SN 比が最大になるよう合成する. すると, 受信した伝送信号が大きくなり, フラットフェージングの影響を軽減できる. SN 比を最大にするには, 最大比合成を行う. 最大比合成とは, 各伝送経路に重み係数を掛け, 位相を揃えて合成する手法である. この重み係数をチャンネル係数と呼び, 1 を拡散して送信したものを受信機で相関器に通すことで得られる.

2.2 線形予測分析

線形予測分析とは, 現在の出力サンプル値 x_t がそれ以前の N 個のサンプル値の線形結合によって予測されるものとするモデルで, 声道フィルタの推定など音声分析に広く用いられている手法である [7][8].

声道フィルタ $H(z)$ を,

$$H(z) = \frac{1}{1 + \sum_{n=1}^{N} \alpha_n z^{-n}}$$

と表される全極型のフィルタで近似することを考える. α_n を線形予測係数と呼び, この値を調整することで, 声道フィルタを近似する.

3 提案手法

本研究では, 送信したいビット列をスペクトル拡散させて, 合成音声に埋め込んだ. この音声をロボットのスピーカから送信し, 別のロボットで受信して得た音声からビット列を復元した. 提案手法における音声生成からビット復元までの流れを図 5 に示す. この手法により, 人間にも理解できる音声に, 秘匿化された情報を埋め込むことができる.

3.1 音声へのビット列の埋め込み

音声生成部分では, まず, 送信したいビット列を BPSK 変調する. 例えば, 送信したいデータが $\{1, 0, 1, 0\}$ であれば, BPSK による変調後は $\{1, -1, 1, -1\}$ となる (0 が-1 に変換される). これと, 拡散系列をビット毎に掛け算し, 拡散ビット列を得る. 拡散ビット列の音圧レベルを調整するため, 合成音声に重み係数を乗じてパワーを増幅した後, 拡散ビット列と足し合わせ, 再生する音声を得る.

3.2 ビット列の復元

ビット復元部分では, まず, 録音した音声にハイパスフィルタをかける. これにより, 合成音声の影響を低減させるとともに, 低周波ノイズが除去できる. 次に, 拡散ビット列を作成するときに使用した拡散系列を用いて, ハイパスフィルタをかけた録音音声の自己相関を求める.

実環境で録音すると, 壁や障害物などに反射して遅延波が発生し, 相関器の出力しまうため, レイク合成により, 先行波と遅延波のピークの合成を行う. レイク合成後, BPSK 復調を行い, ビット列を復元する.

4 実験

本研究では, 前節の提案手法を用いて合成音にビット列を埋め込み, 実環境にて送受信を行った. 実験の評価のポイントは次の 3 点である.

- 送信側ロボットと受信側ロボット間の距離
- 合成音と埋め込みビット列の SN 比 (Signal-to-Noise Ratio : SNR)
- ビット誤り率 (Bit Error Rate : BER)

4.1 実験条件

実験は, 愛知県大学次世代ロボット研究所のアリーナ (室内) にて行った. 実験中の環境音の大きさは 40〜45[dB] 程度であり, 換気扇の運転音のみが聞こえる状態であった.

音声の送信, 受信には, 現在, SPL で使用されている SoftBank Robotics 社の NAO V6(図 6, 7) を用いた. NAO V6 のスペックを表 1 に示す [9]. 送信には左右両方のスピーカから音声を出力し, ビット列の復元には 4 つのマイクのうち 1 つのマイクからの情報のみを使用した. なお, 送信, 受信時のサンプリング周波数は, NAO で録音可能な値である, 48,000[Hz] とした.

表 1: NAO V6 のスペック

項目	詳細
OS	NAOqi 2.8
高さ	57.4cm
スピーカ	左右 2 つ
マイク	指向性マイク 4 つ

音声を送信側 NAO と受信側 NAO 間の距離は, 1.0m, 3.0m, 5.0m, 7.0m, 9.0m, 10.8m の 6 パターンとし, それぞれの BER を比較した. NAO は, 図 8 に示すように, それぞれが正面を向き合うように配置した. 送信側 NAO は直立しており, 受信側 NAO はしゃがんだ状態である. 現在, SPL のフィールドの大きさは 6.0m × 9.0m であり, 10.8m まで誤りなく情報を伝送できれば, 実用上, 十分にフィールドをカバーできたといえる.

4.2 実験用音声の生成

まず, 実験を行うために, 送受信するための音声を作成した. 合成音声は表 2 のように作成した. 声帯音源は, 基本周波数 250[Hz] の音源に共振周波数 1,000[Hz] の 2 次フィルタをかけ, ランダム雑音を足し合わせた. 声道情報には, 人間が発話した音声から線形予測分析 (次数 20) で抽出した線形予測係数を使用し, フレーム単位 (フレーム長 : 1,024 点) で声帯音源にフィルタリングした. また拡散ビット列は表 3 のように作成した.

表 2: 作成した合成音声

項目	詳細
発話内容	“ガンバレーガンバレー”
発話時間	7.0[sec] (末尾 0.2[sec] は無音区間)
サンプリング周波数	48,000[Hz]
基本周波数	250[Hz]

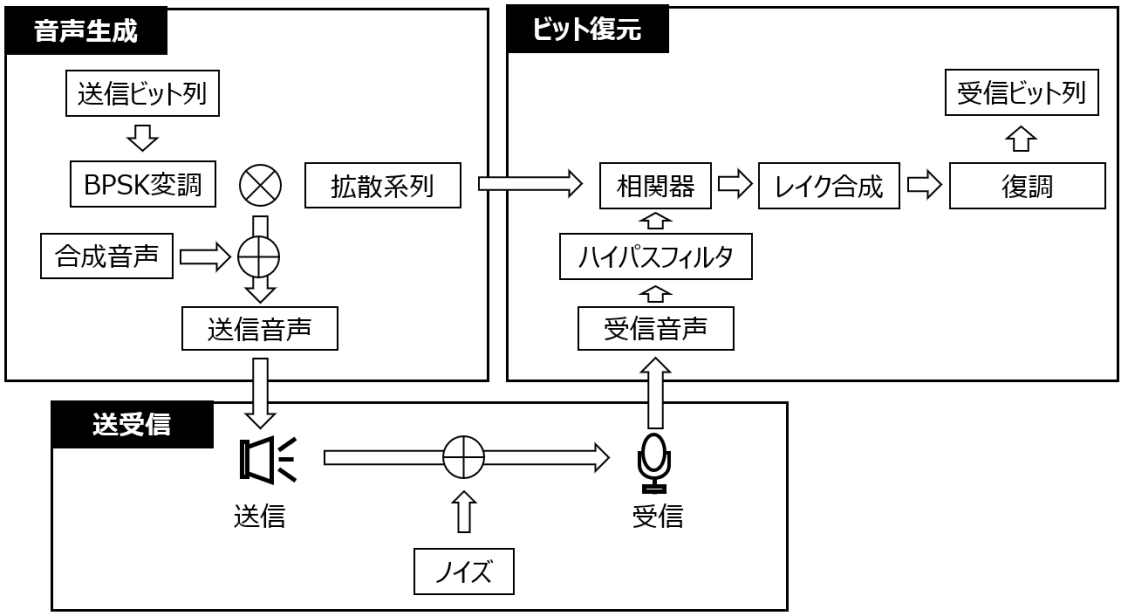


図 5: 提案手法における音声生成からビット復元までの流れ

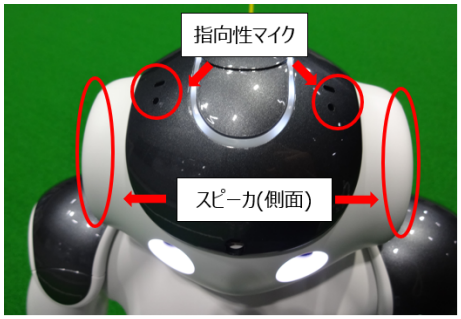


図 6: NAO の前頭部

図 7: NAO の後頭部

なお, 拡散系列には, 1 と-1 がランダムで現れる系列を使用した.

表 3: 作成した拡散ビット列

項目	詳細
送信ビット列	$\{1,1,0,1,0,0,1\}$
伝送速度	1[bps]
拡散系列長	48,000/bit
変調方式	BPSK

合成音声と拡散ビット列を足し合わせる際, そのまま足してしまうと, 拡散ビット列の音の大きさが大きすぎるため, 500, 900, 1,600 の 3 パターンで合成音声に重みづけを行った後, 拡散ビット列と足し合わせて正規化をした. 重み係数を乗じた合成音声と拡散ビット列との SNR を表 4 に示す. 同表より, 重み係数が大きくなると SNR も大きくなるので, 生成音声のノイズ成分が小さくなり, 人間が聴取したときに合成音声が聞き取りやすくなる. なお, SNR は合成音声を信号, 拡散ビット列を雑音成分として算出した.

表 4: 重み係数と SNR の対応

重み係数	SNR[dB]
500	40.8
900	45.9
1,600	50.9

それぞれの重みで合成音声と足し合わせ, 送受信したときの BER を観察した. なお, 全体で 7 ビット送信することになるが, 先頭 1 ビットはチャネル係数を求めるためのテストデータとして用いるので, ビット誤り率は 6 ビット中の誤り率となる.

図 8: 実験中の NAO の配置

4.3 ビット列の復元

音声の送受信には NAO を用いたが, 計算コストのため, ビット列の復元には計算機を使用した. ハイパスフィルタは, 受信音声に表 5 の条件で処理を行った. レイク合成のチャネル係数には, 受信音声の初めの 1 秒の自己相関を用いた.

表 5: ハイパスフィルタの条件

項目	詳細
使用ソフトウェア	MATLAB R2019b
使用関数	highpass
通過帯域周波数	4,000[Hz]

4.4 結果と考察

合成音声の重み係数ごとに, ロボット間距離を変更したときの BER の推移を図 9 に示す. ただし, BER は, 重み係数とロボット間距離の組合せ毎に 5 回送受信を行ったときの平均値である. 重み係数が 500 のときは, すべてのロボット間距離で BER が 0[%] であった.

環境音が 40〜45[dB] という実験環境では, 重み係数が 500 のとき (SNR は 40.8[dB]), SPL フィールド全体に誤りなく情報を伝送できると思われる. 重み係数が 900, 1600 のときは, 拡散ビット列の音の大きさを弱めすぎてしまい, 環境音と分離できなくなったものと考えられる.

5 おわりに

本研究では, 合成音声にスペクトル拡散させたビット列を埋め込み, 実環境で送受信した. 環境音が 40〜45[dB] であれば, SNR が 40.8[dB] の音声を用いて SPL フィールド全体に誤りなく情報を伝送できることがわかった.

今後の課題として, ノイズの少ない実験環境ではなく, より環境音が大きく, 瞬間的なノイズも混在し得る

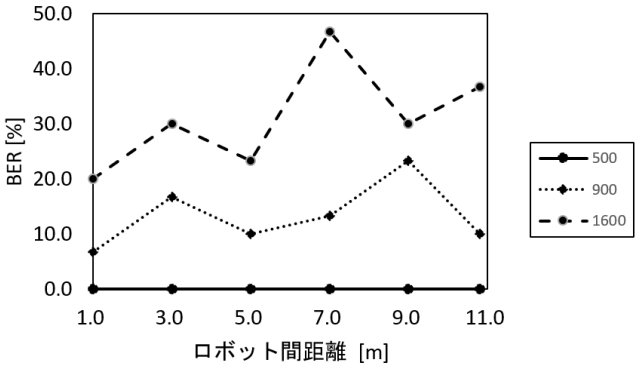


図 9: 重み係数毎のボット間距離に対応する BER の推移

状況での対処方法を検討していきたい. また, 観客にとって違和感のない SNR の限界や, 伝送速度を増加させる方法も探りたい.

謝辞

スペクトル拡散法について助言していただいた本学情報科学部准教授の神谷幸宏先生に感謝する. またロボカッププロジェクトチームのカメリアドラゴンズは, 2019 年度本学から活動費の支援を受けている.

参考文献

[1] ロボカップ日本委員会. “ロボカップとは”. http://www.robocup.or.jp/robocup/, (cited 2020-02-24).

[2] RoboCup Standard Platform League. https://spl.robocup.org/, (cited 2020-02-24).

[3] 神谷幸宏. MATLAB によるディジタル無線通信技術. コロナ社. 2008.

[4] 神谷幸宏. RF ワールド No.31 特集 スペクトル拡散の技術. CQ 出版社. 2015. pp.46-56.

[5] 竹花進悟, 近藤和弘, 中川清司. スペクトル拡散を用いた音響信号用電子透かしの基礎検討. 電気関係学会東北支部連合大会講演論文集. p.344. 2003.

[6] 神谷幸宏. RF ワールド No.31 特集 スペクトル拡散の技術. CQ 出版社. 2015. pp.92-95.

[7] 坂野秀樹. 考えて使いこなす音声のスペクトル分析 (やさしい解説). 日本音響学会誌. 2012, 68 巻, 4 号, pp.188-194.

[8] 板橋秀一 編著. 音声工学. 森北出版. 2005. 256p.

[9] SoftBank Robotics. “NAO - Documentation”. SOFTBANK ROBOTICS DOCUMENTATION. http://doc.aldebaran.com/2-8/home_NAO.html, (cited 2020-02-23).

複数人対話における役割に応じた視線の振る舞いの解析とロボットへの実装

Analysis of Role-Based Gaze Behaviors and their Implementation in a Robot in Multiparty Conversation

新谷太健 [1,2*]　石井カルロス寿憲 [2,3]　石黒浩 [1,3]
Taiken Shintani[1,2]　Carlos T. Ishi[2,3]　Hiroshi Ishiguro[1,3]

[1] 大阪大学
[1] Osaka University
[2] 理化学研究所
[2] RIKEN
[3] 国際電気通信基礎技術研究所
[3] ATR

Abstract: In a multi-person face-to-face dialogue, people naturally gaze according to their roles. The goal of this research is to develop an agent that can control eye movement according to its role in a face-to-face dialogue with multiple users. In this study, we analyze the gaze behaviors in three-party dialogue data accounting for dialogue roles, implement gaze models on a robot based on the analysis results, and conduct evaluation experiments. We show that natural behaviors are achieved by our proposed gaze control system, which accounts for dialogue roles and eyeball movement control.

1　はじめに

近年，遠隔地に存在する他者とコミュニケーションを行うとき，多人数が参加するテレビ会議のようなものだけでなく，アバターを介する対話も増え始めている．モニタ上に表示された擬人化エージェントのようなものから物理的媒体を持つロボットを使用したものまで多種多様なアバターが開発され，それらを遠隔対話のインターフェースとしたアプローチによる研究がなされ，その有用性が明らかにされている．たとえば，小川らはロボットである Telenoid を用いて携帯電話を用いた対話よりも対話エージェント Telenoid を介した対話の方が有益であることを示している [1]．また，遠隔対話のインターフェースとしてのアバターだけでなく，自律型エージェントのアプローチによる研究もなされている．

対話において，言語情報は会話を進める上で相手に発話意図や意味・内容・情報を中心となる要素である．同様に，非言語情報も言語情報を補佐する機能だけでなく，言語・発話だけでは伝えられない意図・欲求を表出したり，対話を円滑に進めたりと対話において重要となる要素である．非言語情報の重要な要素の 1 つとして，視線がある．視線の動きは対話を円滑に進める腕最も重要な要因の一つであり，視線に関する研究は様々になされている．Kendon らや Argyle らは視線には「会話の流れを制御する機能，対話相手にフィード学区を与える機能，感情情報を伝える機能，共存感覚を与える機能，対話相手に正もしくは負の評価を与える機能」があると論じている [2][3]．

対話においても，遠隔操作エージェント・自律型エージェント問わず視線は非常に重要な役割を持つ．ロボットやバーチャルエージェントの視線が対話に影響を与えることは様々な研究者によって示されている [4]．しかし，エージェントの視線が対話において重要な機能を果たすことが明らかにされているものの，視線行動の人工的生成手法に関してはまだ十分に研究されていない．特に，眼球検出機器の推定精度がまだ十分であると言えず，実際の対話時における眼球の動きに関する研究・眼球の動作生成手法に関する論文は少ない．

二者対話に関する研究は盛んに進められている．一方，現実の対話において大部分を占める複数人対話に関する研究は少ない．複数人対話では，二者対話以上

*連絡先：大阪大学大学院基礎工学研究科
〒 560-0043 大阪府豊中市待兼山町 1 丁目 3
E-mail:shintani.taiken@irl.sys.es.osaka-u.ac.jp

に複雑な情報 (今誰が発話権を所持しており，発話しているのか，話題は何かなど) を認識する必要があり，また，振る舞い方もより困難なものとなる．しかし，三者以上の複数人で行われる対話の場にエージェントが居合わせ，適切な振る舞いをする動作生成システムの存在は必要である．そのため，本研究では複数人対話，その中でも三者対話を対象として扱う．

また，タスク指向の対話のように何か目標が存在しそこに向かって進んでいくような対話に関する研究は多いが，井戸端会議のような目的がなく明確に議論の進め方のルールが存在しないにもかかわらず会話の弾むような対話に関する研究も少ない．エージェントが社会進出していくうえで，エージェントは会議のような意味のある対話から何気ない対話まで対応していく必要がある．そのため，本研究では，主に雑談 (フリートーク) に着目し，扱っていく．

Sacks らによってターンテイキングと視線に相関があることは既に示されている [5]．また，複数人対話において非常に重要となる要素として今の自分の役割を認識することがあげられる．今自分自身が話者であるのか，話者から主に話しかけられている聞き手であるのか，会話に少ししか関与しないサブリスナーであるのかという役割と視線の振る舞いに相関があると考えられる．そこで，本研究では，実際の対話データに基づいた視線の動き・眼球の動きをターンテイキングに注意しながら，また，役割ごとに分類して解析を行い，視線パターンの行動分析を行った．その後，社会的対話ロボットである CommU に提案である視線生成手法を実装し，その有用性を示す．

2　従来研究

複数人対話における視線の解析・実装に関しては，さまざまな方向から研究されている．

例えば，J.Lee らは 4 人の被験者に対して 7 分の西部劇を演じさせ，それぞれの役割 (Speaker,Adressee,side participant, bystander)，4 者のキャラ間の関係性 (好き，嫌い，どちらでもない)，実際の演じ方から視線の特徴を抽出しようとした [6]．

また，中野らは 1 人のロボットと 3 人の人による 4 者対話を行わせ，対話における支配度と対話者の役割に視線が相関するのではないかと考え，解析・実装を行っている.[7]．

前述したように，ターンテイキングと視線の動きに相関があることは明らかにされているが，例えば，石井らは 4 者対話を解析し，IPU 終了時における視線パターンからターンテイキングの推定を行おうとした [8][9]．

本研究では，対話内容に目的がなく，また，エージェントが司会進行するような対話でない，対話者全員が対話参加者となり，何気ない会話をエージェントも一人の対話者に過ぎない雑談の場において自然な振る舞いをするエージェントの解析・実装を目的としている．次に，本研究で扱う三者対話に関する研究について述べる．三者対話に関しては主に 2 種類の状況が考えられている．一つは三者で対話を行うものであり，もう一つは二者対話に陪席者を用意した対話である．前者に関する視線の研究として, 武川らは発話交替時のスピーカーではない聞き手二人に着目し，聞き手同士の視線の配分・スピーカーへの視線量と次の話者に関係があることを解析した [10]．榎本らは三者対話における参与役割の交代にかかわる非言語行動の分析を行っている [11]．三者対話の会話分析を行った結果，現話者は次話者を見ていることが多く，次話者・非話者は話し手を見ていることが多く，現話者の役割から，会話参与者のうち視線を合わせることがその参与者を次話者に選択する手段とみなせるが，次話者の立場からは現話者を見ることにより発話が増えたと解釈が可能であると主張している．また，後者の研究としては，有本らは陪席ロボットと対話ロボットの二つのロボットを用意し，この両者のロボットがアイコンタクトを行うことで対話ロボットと会話をする被験者からのロボットへの印象・社会的存在感が上がることを示した [12]．酒井らは 2 つのロボット議論を行わせ，その議論を見て参加した気になっている陪席者である被験者から対話に参加したように見える陪席者の視線モデルを作成した [13]．Mutlu らは両者の場合において，話者の視線配分がどのようになっているのかを解析し実際にロボットに実装している [14]．

本研究では，主に三者対話を扱うことを考えている．そのため，陪席者の役割となる人は存在はせず，すべての参与者に発話権は存在していると考え，陪席者とは異なるサブリスナーという考え方を用いている．役割ごとに視線モデルは異なる．しかし，対話における陪席者でなくサブリスナーとなる人の視線，特に発話中のサブリスナーの視線の解析・実装を行った研究は不十分であったと考えられるため，本研究では収録されたデータセットからサブリスナーの役割となる人の視線配分についての解析を行い実装した．また，収録したデータセットから実際の対話における視線の動きに関して顔よりも目玉の動きの方が支配的であったが，実際に三者対話における黒目の位置の配分について解析・実装を行っている研究がなかったため，瞳の位置についても収録されたデータセットから解析・実装を行った．

3　データ解析 (ラベル付け)

3.1　データセット

データセット本研究で使用する三者対話のデータセットについて述べる．1 グループ 3 人 6 グループ，男女 9 人の実験参加者が対話を行った．対話の内容はフリートークであり，各々自由に 20 分から 30 分間談笑しあった．各被験者は三角形を描くように椅子に座って話を行い，各被験者の前にカメラが設置され真正面から顔・体の動きが撮影されている．また，各被験者ごとにヘッドセットマイクが装着されており，各参加者の音声データ・議論全体の動画が収集されている．また，録画・音声データから対話の書きおこしを作成した．

本研究では，視線方向・眼球の方向・発話権の有無・ターンテイキング・役割が視線の動きと非常に強い相関があると考えた．そのため，アノテーターの方に前述した対話データのラベル付けを行ってもらい，このラベルを下に解析を行った．また，ラベル付けのルールの説明を行う．

3.2　データの分析

このセクションでは，実際にどのようなラベルを作成・付与し解析を行ったかについて述べていく．

3.2.1　視線方向

被験者は図**??**に示すように三角形を描くように斜めに座り対話を行っている．被験者がどの被験者を見ているのかラベル付けを行う．ラベルは被験者番号 2 つとそれ以外の 3 つである．この時，顔の角度だけでなく眼球の方向も含めて視線とし，この 2 つの要素から総合的に判断された方向を視線の方向としてアノテーターの主観により決定している．実際にラベル付けをしている様子を図 1 に示す．例えば，A という ID を持つ人は B→C→env→B→env と視線が遷移している．ここで，env というラベルは B も C も見ていない状態，つまり視線を逸らした状態を意味する．

3.2.2　眼球の方向

人を見る動作は，ラベルと顔・瞳の位置にそこまで大きなずれは生じない．しかし，視線を逸らすという動きにおいては，視線・特に瞳の位置は非常に重要な要素となる．だが，瞳の位置検出を自動でやることは精度が低く困難であり，眼球の方向を自動で検出する精度は現状低い．そのため，眼球の方向も上下左右斜め真ん中の 9 方向でラベル付けを行う．特に，被験者を見ていない動き，つまり視線逸らしを行うとき顔の角度だけでなく眼球の方向は非常に重要な要素となる．実際のラベルの様子を図 1 に示す．視線を逸らすラベルの時，付加情報として，瞳はどこを向いていたのかをラベル付けしている．

3.2.3　発話権の有無・ターンテイキング

図 2 に実際の対話の様子を示す．

まず，ターンを取った時，”Turn Taking”というラベル付けを行っている．これにより，実際にターンが交替したタイミングを知ることができる．”Turn Taiking”ラベルはアノテーターの方がターンが交替した・話者が発話権を取ったと判断した発話の開始時に付与している．そのため，オーバーラップや相槌・フィラーのような発音に影響されず，また，これにより IPU(Inter-Pausal. Unit) を用いなくても正しいターンテイキングのタイミングを知ることができる．

また，ターンテイキング開始時がわかっているため，ターンが切り替わった瞬間の前後 1 秒ずつ計 2 秒間 (図 2 の赤丸で示す部分) における視線ラベルの解析を行う．

3.2.4　役割

本研究では，対話における役割が重要であると考えている．本研究では以下のように役割を定義した．

- **Speaker(Sp)** 発話権を持っている人/発話権を持っており，話している人・発話をしている人
- **MainListener(ML)** 上記のスピーカーが主に話しかけている対象
- **SubListener(SL)** Sp でも ML でもない聞き手の人

本研究ではスピーカー・メインリスナー・サブリスナーをターンテイキング時に着目し以下の手法により分別して解析を行った．図 2 において，このターンテイキングで発話権を取得した人 (図 2 では C に当たる人) を話者 (Sp)，発話権を譲渡したまたは取られた人 (図 2 では A に当たる人) をメインリスナー (ML)，このターンの交替に関与しなかった人 (図 2 では B に当たる人) をサブリスナー (SL) とする．

- **Speaker(Sp)** ターンテイキング時 (発話交替) に発話権を取った人
- **MainListener(ML)** ひとつ前にターン (発話権) を持っていて，スピーカーに発話権を譲った人
- **SubListener(SL)** ターンのやり取りに関与しなかった人

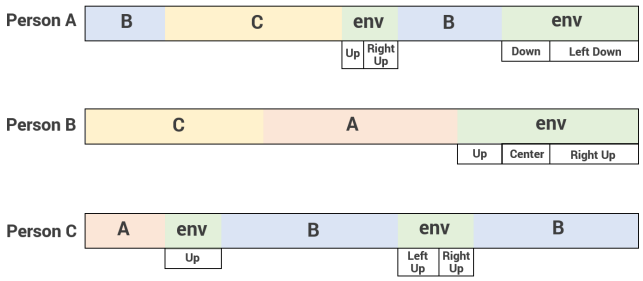


図 1: Gaze Label

3.3 解析結果

本研究ではターンテイキングに注意して，ターンテイキング時とそれ以外に分けて視線パターンの考察を行った．

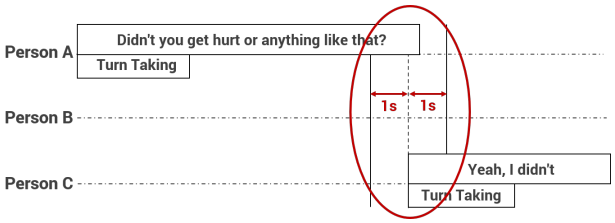


図 2: ターンテイキング

3.3.1 ターンテイキング時の解析

ターン交替時の各役割ごとの視線の動きの割合の推移について図 3 に示す．横軸は時間 (単位は秒) であり，時間 t=0.0 の時に話者がターンを取り，話し始める．また縦軸は時間 t においてどこを見ているのかの統計を取りその割合 (0.0-1.0) を表している．例えば図 3 の (a) の図の横軸 t=0.0s の縦軸の値を見ると ML,SL,env それぞれ 0.37,0.19,0.44 であるが，これは話し始めたタイミングでの話者の視線の先の総計のうち 0.37 がメインリスナーを，0.19 がサブリスナーを，そして残る 0.44 視線を逸らしていたことを表している．0.1 秒ごとに算出しており，図 3 では，話者がターンを取り話し始めたタイミングの前後 1 秒ずつ合計 2 秒の区間における視線のやり場の割合が時間経過とともに遷移する様子を表している．

スピーカー

図 3 の一番左の図 (a) にターン交替時における話者の視線の割合の推移を示す．話者はターンを取る 1 秒前はメインリスナー (ML)，つまり，前の話者を見ている割合が高い．しかし，ターンを取り，発話を開始する地点 (t=0.0s) に向かうにつれて，視線を外す割合が高くなっている．そして，話し始めて 0.1 秒経過したあたりで視線を逸らす割合はピークを迎えた後，視線はメインリスナー (ML) を向く割合と視線を逸らす割合がほぼ同等に高くなりターン交替における視線の遷移を終える．この結果より，人は最初前の話者を見て，話を始める際に視線を逸らし，以降は視線を逸らしたまま，あるいは前の話者であるメインリスナーを見る傾向にあると考えられる．

メインリスナー

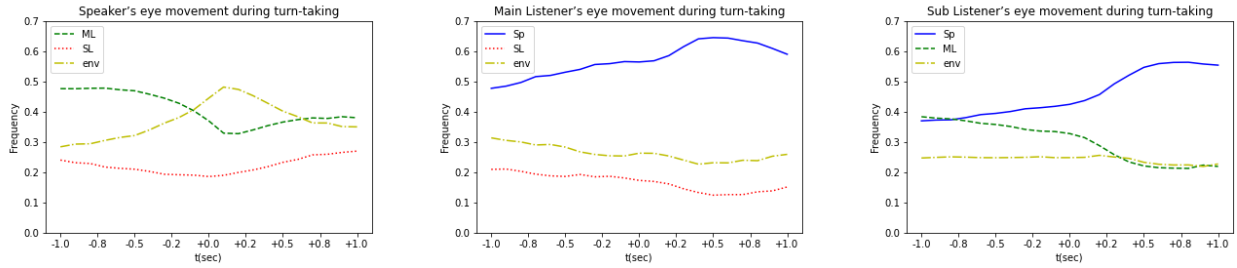
図 3 の真ん中の図 (b) に，0.0 秒に話者が話し始めるがその直前までターンを持っていたメインリスナーの視線の割合の遷移を示す．このターンテイキング時に，メインリスナーがターンを渡した，取られた，もしくは自然に交替した等，状況は様々に考えられるが全体として話始める前から次話者がわかっている或いは決めており，1 秒前からターンを取られた後 1 秒後までのターンテイキングの区間において話者の方を見続ける傾向にあると考えられる．

サブリスナー

最後に，ターンを渡しターンを取るターン交替のやり取りに関与せず，その様子を見ていたサブリスナーの視線の割合の遷移を図 3 の右の図 (c) に示す．横軸 t=-1.0s の時，即ち，話者が発言する 1 秒前までは前話者であるメインリスナー (ML)，次話者であるスピーカー (Sp) を見る割合が同党に高く，そこから t=1.0 つまり話者が発話を開始し 1 秒経過した地点に向かうにつれて話者を見る割合が高くなっている．そのため，サブリスナーは 1 秒前まではメインリスナ (前話者) の方もしくは次話者を推測しスピーカー (Sp) の方に視線を向け，そこから次の話者へと視線を動かす傾向にあると考えられる．

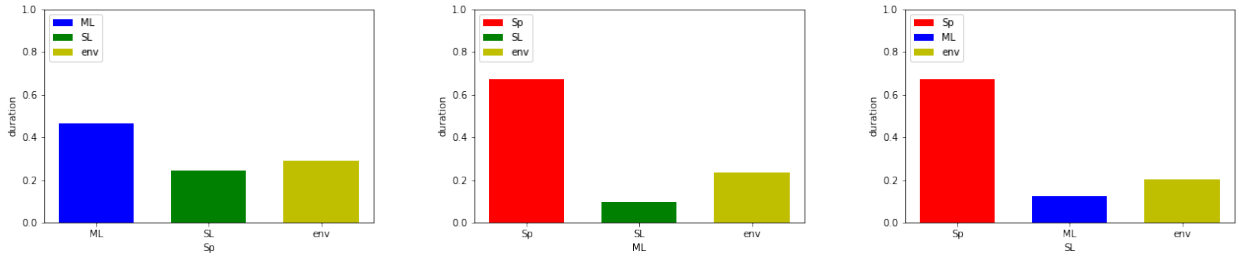
3.3.2 ターンテイキング外の解析

では，ターンテイキング時以外での視線の動きの解析を行う．前後 2 秒ずつ合計 4 秒はターンテイキングからの影響を受けているものと考えられるため，この 4 秒を省いて，また，短すぎる発話を省いて解析を行った．



(a) 話者の視線の動きの遷移の割合　(b) メインリスナーの視線の動きの遷移の割合　(c) サブリスナーの視線の動きの遷移の割合

図 3: ターン交替時の各役割の視線の動き



(a) 話者の発話中の視線の時間の割合　(b) メインリスナーの発話中の視線の時間の割合　(c) サブリスナーの発話中の視線の時間の割合

図 4: 発話中における各役割の視線を向ける時間の割合

3.4　発話中の視線の動きの解析

次にターンテイキングでない区間，つまり発話の冒頭部分，末端部分を除いた話者が発話をしている区間における視線の解析を行う．図 4 に各役割ごとの発話中における視線の時間の割合を示す．横軸は視線の方向を現しており，縦軸は時間の割合を示す．ここで、ターンテイキングを除くため，4 秒以上の発話を抽出し，前後 1 秒ずつ合計 2 秒間を取り除いている．また，冒頭部分，末端部分のラベルに関して，発話区間外と一続きでかつ，区間内では視線方向の持続時間が 1.0 秒以内，区間外で視線の持続時間が 1.0 秒以上の時，ターンテイキング要素と強く結びついていると考え取り除いている．

表 1 に図 4 で示した発話中の視線の割合の具体的な数字を示している．

図 5 に発話中における視線を逸らした回数の散布図を示す．横軸は発話中の時間 (先ほど選別したものを使用) を示しており，縦軸はヒトから視線を逸らした回数を示している．また，その時の近似直線も表している．近似直線の回帰係数 r だが図 (a),(b),(c) それぞれ 0.83,0.82,0.67 である．図 5 及び表 1 より，役割に依らず視線を逸らす感覚が変わらないことがわかる．

話者

まずは発話中の話者の視線の解析を行う．図 4 の一番左の図 (a)，表 1 より，話者は発話中に前の話者であるメインリスナー (ML) を見る割合が少し高いが，全体としてバランスよく視線の配分をしていることがわかる．

メインリスナー

次に，発話中のメインリスナーの視線の解析を行う．図 4 の真ん中の図 (b)，および表 1 より，メインリスナーは 7 割近く話者の方向を向いている．

サブリスナー

発話中のサブリスナーの視線に関してもメインリスナー同様，図 4 の一番右の図 (c)，および表 1 より，7 割近く話者の方向を向いている．

3.4.1　視線逸らしの解析

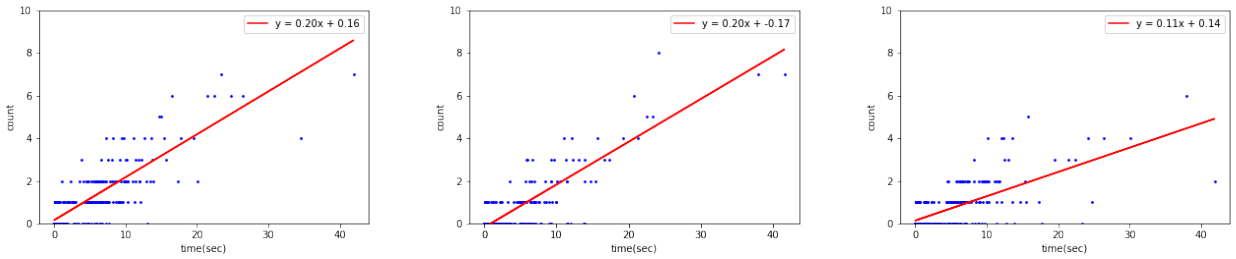
では，次に視線逸らしにおける解析を試みる．ヒトを見るときの視線の動きは非常に単純であり，違和感のないように顔を向けばよい．しかし，視線を逸らすときは単純な動きだけでは不十分である．実際に収録されたデータを見ると眼球だけ移動させて視線をそらしている動きがみられた．そのため，視線を逸らすときの動きの解析として，今眼球がどこを向いているのかに特に着目した．また，視線を逸らすときにどれだけの時間逸らしているのかも分布にして解析を行った．

視線逸らし時における瞳の配分

次に，視線を逸らす動きをした時の継続時間のヒス

表 1: 発話中の視線の割合と逸らす時間

Role	Speaker	Main Listener	Sub Listener	Look away	Interval(look away)(s)
Speaker	0.0	0.47	0.24	0.29	5.03
Main Listener	0.67	0.0	0.10	0.23	6.05
Sub Listener	0.67	0.13	0.0	0.2	6.90



(a) 話者の発話中の視線を逸らした回数　(b) メインリスナーの発話中の視線を逸らした回数　(c) サブリスナーの発話中の視線を逸らした回数

図 5: 発話中における各役割の視線を逸らした回数

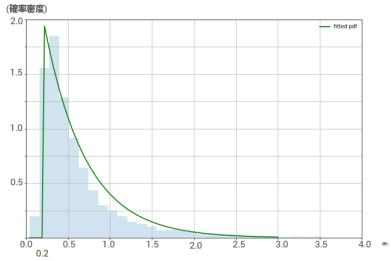


図 6: 視線を逸らすときの時間のヒストグラム

トグラムを図 6 に示す．指数分布

$$p(x;\mu,\lambda) = \lambda \mathbb{1}_{x\geq\mu} \exp\left(-\lambda(x-\mu)\right) \qquad (1)$$

で近似される．ここで，$\mu = 0.2, \lambda = 0.63$ である．なお，0.2s 以下のサンプルは実際のエージェントへの実装を考えたとき，非現実的であると考え，0.2s 以上のものに限った．平均値は 0.83s，中央値は 0.55s である．

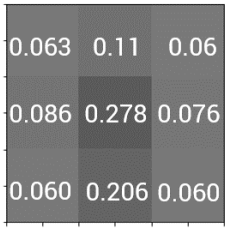


図 7: 視線を逸らすときの瞳の位置の割合

視線逸らし時における瞳を向ける時間

次に，視線を逸らす動作をしたときにおける黒目の位置についての解析結果を図 7 に示す．黒目が上下左右斜め中央の 9 か所のうち，それぞれの時間の合計がわかっているため，そこから相対度数の計算を行った．傾向としては，中央・ましたに瞳を動かして視線を逸らす傾向強く，残りの 7 か所は同党の割合でまばらに見ていることがわかる．

4　実装

解析した内容を視線制御システムとして実装を行う．瞳の動きは Pejsa らの提案した手法を参考に行う [15].

4.1　ターンテイキング時の実装

次に，ターン交替時におけるロボットへの実装方法について記述する．図 3 に示すように，各役割ごとにターン交替時における視線の動きには傾向がある．そのため，自分の今の役割に応じたターンテイキングの視線の動きを生成し実装する必要がある．

本研究では，ターンテイキングの 2 秒の区間を-1.0s -0.3s,-0.3s 0.3s,0.3s 1.0s の 3 つの区間に分割し，それぞれの区間における特徴的な所の割合から，その区間ごとにどこを見るべきかを抽出，算出し，ターンテイキングの 2 秒間で 2 回の遷移，合計 3 か所を見る実装を行う．

スピーカー

役割がスピーカーの時の視線の割合の推移が図 3 の (a) に示されている．この図より，-1.0s -0.3s の区間

において，-1.0s の時に特徴があると考え，この時の ML,SL,env を見る割合がそれぞれ，0.48,0.24,0.28 であった．この割合を確率とし，一様分布からどこを見るかを決める．次に，-0.3s 0.3s の区間において，0.1s の時に特徴がある．この時の ML,SL,env を見る割合はそれぞれ，0.33,0.19,0.48 である．最後に，0.3s 1.0s の区間における特徴は 1.0s に見られる．この時の ML,SL,env を見る割合はそれぞれ，0.38,0.26,0.36 である．

メインリスナー

次に，役割がメインリスナーの時のターンテイキング時の視線の動きの実装について考える．この時の視線の割合の推移は図 3 の (b) に示している．この図より，-1.0s -0.3s の区間において，-1.0s の時に特徴がみられ，この時の Sp,SL,env を見る割合はそれぞれ，0.48,0.21,0.31 であった．次に，-0.3s 0.3s の区間において，0.0s の時に特徴がみられ，この時の Sp,SL,env を見る割合はそれぞれ，0.56,0.17,0.27 であり，また，0.3s 1.0s の区間においては 0.5s の時に特徴がみられ，その時の SP,SL,env の割合はそれぞれ 0.65,0.12,0.23 であった．

サブリスナー

最後に，役割がサブリスナーの時のターンテイキング時の視線の動きの実装について考える．この時の視線の割合の変化の推移は図 3 の (c) に示している．-1.0s -0.3s の区間において，-1.0s の時に特徴がみられ，この時の Sp,ML,env を見る割合はそれぞれ，0.37,0.38,0.25 であった．また，-0.3s 0.3s の区間において，0.2s の時に特徴がみられ，この時の Sp,ML,env を見る割合はそれぞれ 0.46,0.28,0.26 であり，0.3s 1.0s における特徴は 0.8s の時に見られ，Sp,ML,env を見る割合は 0.57,0.21,0.22 であった．

ターンテイキング時の実装

以上の 3 つの役割における特徴を用いて各区間における確率より，一様分布から抽出した値から視線を向けるべき方向を算出し，視線の動きを作成する．

4.2 ターンテイキング時以外の実装

図 4 に示すような時間の割合を維持しつつ，また，各場所を見る時間も実際のデータに近くなるように発話中の視線の動きの実装を行う．今，誰を (或いは視線を逸らす) どれだけの時間継続して見ていたのかがわかる視線ラベルデータが大量にある．エージェントの役割ごとにラベルの数がわかり，そこから割合が決まる．次に割合を確率と考えどの役割の人を見るのかを決める．見る対象が決まったら，図 6 のように視線を向ける継続時間のヒストグラムがわかっているため，そこから見る時間を抽出する．

4.3 視線逸らし時の動き

実際に視線を逸らす動作をロボットの CommU に実装した様子を図 8 に示す．左右における，上中下それぞれの場所に瞳を動かした様子である．画像 8 から目玉の位置が印象に大きな影響を与えそうなことがわかる．

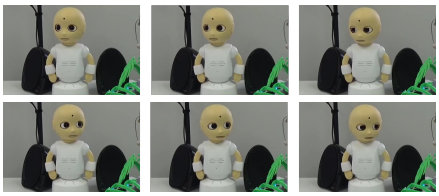


図 8: 実験の様子

5 実験

本研究では，役割ごとの視線パターン・視線を逸らすときの瞳の動きが対話における重要な要素であると考えている．そのため，以下の 4 つの条件で比較することにした．

同割合-頭部モデル

ベースラインとして，話者の時，2 人の対話者を同じ割合で自然配分するモデルを使用した．2 人の対話者とそれ以外のところを見る割合であるが，先行研究 [14] より，人・人・それ以外を見る割合をおおよそ 3:3:4 になるように設定した [14]．また，視線を向ける場所であるがこれも同論文より，人の顔ないし 2 人の対話者の間に角度 4 度を分散とした 2 次元ガウス分布を適用して得られた場所を見るようにした．聞き手の時は違和感のないように話者の方向を向くようにした．

同割合-頭部-眼球モデル

眼球の動きが大切であると考えているため，前述の同割合モデルで動かす部位を頭だけでなく眼球も追加し共に動かすモデルとした．

提案-頭部-眼球モデル

4 節に示した役割に基づいた視線制御システム・視線逸らしシステムの実装を行った提案モデルである．

提案-頭部モデル

前述の提案モデルのうち，眼球の動作をなくし，頭部動作のみを行う．

5.1 印象評価

ヒトが自然にこなしている役割の理解・視線逸らしを実装することで一番期待したい効果としてロボットが自然らしいかどうかが大切であるため，「全体的にロボットの振る舞いは自然に感じましたか」という質問項目を設けた．

男女 30 人 (平均 32.1 歳，分散 10.3) に対して 4 つの動画を視聴していただきそれぞれに対して印象評価をしてもらった．

本実験では，実際に収録した 3 人で行われた対話の中の 1 人の会話音声をロボットに搭載したビデオを視聴し，その時の視線の動き方について評価してもらった．実際には図 8 で示すような動画を見ていただき，2 名の対話者は，ロボットの斜め左側と斜め右側にいて，左側の人の声は左耳に，右側の話者の声は右耳に，ロボットの声は両耳に聞こえるようにした．そして 4 つの手法についてそれぞれ個別に印象評価をして頂いた．

実験手順は次のとおりである．まずは，順序効果を減らすために 4 つの動画をの提示順序をランダム化した．次に，各 4 つの手法の動画について個別に 7 点スケール（1 ：とても不自然、4 ：どちらともいえない、7 ：とても自然）で印象評価していただいた．このセットを 3 つの対話区間（各対話区間の長さは 1 分程度）に対し合計 12 個の動画について印象評価していただいた．

5.2 評価結果

3 つの対話区間のうち、2 つでは条件間に有意差がみられなかったため、有意差がみられた 1 対話区間の結果について報告する．

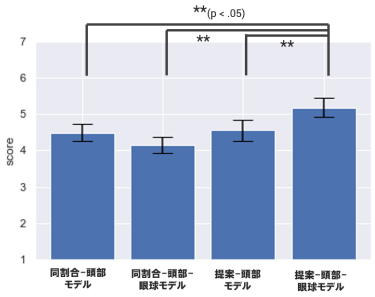


図 9:「全体的にロボットの振る舞いは自然に感じましたか」の結果

「全体的にロボットの振る舞いは自然に感じましたか」への評価結果を図 9 に示す．各モデルにおいて結果の平均値，標準誤差，ライアン法に基づいた多重比較の結果を算出した．ライアン法の結果, 同割合-頭と提案-頭部-眼球モデル間に p 値 0.020($\leq$ .05)，同割合-頭部-眼球と提案-頭部-眼球モデル間に p 値 0.001($\leq$ .05)，提案-頭部と提案-頭部-眼球モデル間に p 値 0.008($\leq$ .05) とそれぞれ有意差が算出された．この結果から，視線逸らしモデル・役割に応じた視線制御モデル及び頭部・眼球を共に動かすモデルがロボットの振る舞いが最も自然であることが示された．しかし，その他の対話区間では，有意差はみられなかったことから、どのような条件において、提案法が効果的であるのかを今後詳しく調べる必要がある．

6 考察

評価実験の結果から，視線逸らしと役割に応じた視線配分を実装し，頭部動作だけでなく眼球も動かしたモデルがロボットの振る舞いを自然に見せることに有効であることが示された．しかしながら，本実験では視線逸らしと役割に応じた視線配分の 2 要因のうちどちらが強く影響を与えたのか，あるいは，両者がそろって初めて自然な振る舞いとなるのかを明らかにすることはできなかった．そのため，より細分化して再実験を行う必要がある．また，ロボットの振る舞いは対話状況や対話者間の関係・対話者の心的状態に大きく影響を受けることが考えられる．対話の役割ごとのモデルの効果は今後詳細に調べる必要があり、今後の予定とする．

7 まとめ

本研究では，対話役割に応じた視線パターンと視線を逸らすモデルがロボットの自然な振る舞いに影響を与えることを示した．現在，人間酷似型ロボットであるアンドロイド ERICA にも実装を試みており，より人に近いロボットでも自然に振舞えるようなモデルへと改良している．また，現在のモデルは乱数に依存しすぎているため，対話内容などから考えて明らかに不適切な動きをしないような条件付きモデルを作成することを今後の課題とする．

謝辞

本研究は、 JSPS 科研費 JP20H05576 の助成を受けたものである．データ分析にご協力いただいた村瀬妙子氏、 中西京子氏、中山祐佳氏に感謝する．

参考文献

[1] K. Ogawa, S. Nishio, K. Koda, K. Taura, T. Minato, C. T. Ishii, and H. Ishiguro. Telenoid: Telepresence android for communication. In *ACM SIGGRAPH 2011 Emerging Technologies*, SIGGRAPH ’11, New York, NY, USA, 2011. Association for Computing Machinery.

[2] A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, Vol. 26, pp. 22–63, 1967.

[3] M. Argyle and M. Cook. Gaze and mutual gaze. 1976.

[4] M. Shimada, Y. Yoshikawa, M. Asada, N. Saiwaki, and H. Ishiguro. Effects of observing eye contact between a robot and another person. *International Journal of Social Robotics*, Vol. 3, pp. 143–154, 2011.

[5] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pp. 7–55. Elsevier, 1978.

[6] J. Lee and S. Marsella. Modeling side participants and bystanders: The importance of being a laugh track. In *International Workshop on Intelligent Virtual Agents*, pp. 240–247. Springer, 2011.

[7] Y. I. Nakano, T. Yoshino, M. Yatsushiro, and Y. Takase. Generating robot gaze on the basis of participation roles and dominance estimation in multiparty interaction. *ACM Trans. Interact. Intell. Syst.*, Vol. 5, No. 4, December 2015.

[8] R. Ishii, K. Otsuka, S. Kumano, R. Higashinaka, and J. Tomita. Analyzing gaze behavior and dialogue act during turn-taking for estimating empathy skill level. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2018. Association for Computing Machinery.

[9] R. Ishii, K. Otsuka, S. Kumano, M. Matsuda, and J. Yamato. Predicting next speaker and timing from gaze transition patterns in multi-party meetings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, New York, NY, USA, 2013. Association for Computing Machinery.

[10] 徳永弘子, 湯浅将英, 武川直樹. 3 人会話における発話交替時の視線行動分析：聞き手の立場から見た発話・非発話の戦略. 電子情報通信学会技術研究報告. HCS, ヒューマンコミュニケーション基礎, Vol. 106, No. 268, pp. 23–28, sep 2006.

[11] 榎本美香, 伝康晴. 3 人会話における参与役割の交替に関わる非言語行動の分析 (テーマ:一般). 言語・音声理解と対話処理研究会, Vol. 38, pp. 25–30, jul 2003.

[12] T. Arimoto, Y. Yoshikawa, and H. Ishiguro. Multiple-robot conversational patterns for concealing incoherent responses. *International Journal of Social Robotics*, Vol. 10, No. 5, pp. 583–593, 2018.

[13] K. Sakai, F. Dalla Libera, Y. Yoshikawa, and H. Ishiguro. Generation of bystander robot actions based on analysis of relative probability of human actions. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 21, No. 4, pp. 686–696, 2017.

[14] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, Vol. 1, No. 2, pp. 1–33, 2012.

[15] T. Pejsa, S. Andrist, M. Gleicher, and B. Mutlu. Gaze and attention management for embodied conversational agents. *ACM Trans. Interact. Intell. Syst.*, Vol. 5, No. 1, March 2015.

障害物検知のための
測域センサ取り付け角度に関する一考察

About the mounting angle of the Laser Range Finder for detecting obstacle

藤井　穂尊[1]　鈴木　勇貴[1]　植村　渉[1]

Hotaka Fujii[1], Yuki Suzuki[1], and Wataru Uemura[1]

龍谷大学[1]
Ryukoku University[1]

Abstract: 近年，少子高齢化による働き手の不足や，人件費の高騰から，様々な業種でロボットによる自動化の需要が高まっている．ショッピングセンター等の警備員不足に対応するための警備ロボットがある．このような移動式ロボットに欠かせない機能の一つとして障害物回避がある．通常はロボットの前方上部に測域センサを取り付けて障害物を検知し，回避する．しかしながら，LRF を地面と水平に取り付けているため，LRF の取り付け位置より背の低い物体は検知することができず，衝突する危険性が存在する．そこで本研究では，LRF をロボットの前方に向けて傾けて取り付けることで，背の低い物体を検知する方法を提案し，角度と検知範囲の関係性について議論する．

1 はじめに

近年，少子高齢化による働き手の不足や，人件費の高騰から様々な業種でロボットによる自動化の需要が高まっている．弊研究室では産学連携活動の一環として，地元企業と共同で警備業務の人手不足の解消を目的として，警備ロボットの試作機を開発した(図 1). この警備ロボットは老人ホーム等の施設の夜間警備を想定した移動式ロボットである．

このような移動式ロボットに欠かせない機能の一つとして障害物回避がある．試作した警備ロボットは障害物回避のために，他の移動式ロボットでも広く利用されている，測域センサ(Laser Range Finder : LRF)を用いて物体を検知し回避する．LRF は内部のレーザーを照射する装置を回転しながら，レーザーが反射してくるまでの時間を測定することにより，距離と回転角を取得する装置である．

通常のLRFはロボット前方上部に地面と水平に取り付け，物体の検知に用いる．この際，LRF を取り付けた高さより，背の低い物体はレーザーが照射されず検知できないため，衝突する危険が存在する．

そこで，本研究では LRF をロボット前方の床に向けて傾けて取り付けることで，通常の LRF の取り付け方では検知できなかった背の低い物体を検知することを検討する．

図 1 試作した警備ロボット

2 移動式ロボットと LRF

本章では，本研究で扱う移動式ロボット Robotino[1]と，LRF について述べる．

[1] 連絡先：龍谷大学先端理工学部電子情報通信課程
滋賀県大津市瀬田大江町横谷 1-5
E-mail: wataru@rins.mail.ryukoku.ac.jp

2.1 移動式ロボット：Robotino

本研究で使用するロボットは Festo 社が販売している，全方位移動ロボット Robotino[1](図 2)である．1 章で述べた警備ロボットは Robotino を基にサーモカメラをはじめとするカメラやセンサを搭載し，ホイール部分にサスペンションなどの機構を組み込んで試作したものである．

図 2　Robotino

2.2 URG-04LX-UG01

本研究で使用する LRF は北陽電機が販売する，URG-04LX-UG01[2]である．1 章で述べた警備ロボットで使っている LRF と同種のものである．外形を図 3 に，加えて仕様を表 1 に示す．

表 1 URG-04LX-UG01 仕様

測距範囲	距離：5.6m
	角度：240°
測距精度	0.06~1m：±30mm
	1~4m：±3%
測距分解能	距離：約1mm
角度分解能	0.36°

図 3 URG-04LX-UG01

3 障害物の検知方法

本章では LRF を床に向けて傾けた際の障害物の検知方法について述べる．図 4 に LRF を水平面と異なる角度で取り付けた時の検知範囲を示す．

図 4 中の点 A にて床を検知する．LRF のレーザーの始点を点 O として，LRF の測域を上から見た図が図 5 である．

$\overrightarrow{|AO|}$は床までの LRF の測定距離である．LRF のレーザーが床に照射する時，直線 BC は点 O から床までの距離を測定した点群の集まりとなる．そこで Robotino の直径と同じ 500mm の幅に何も無ければ進行可能であると考え$\overrightarrow{BC}$の目標値を 500mm で定義する．そこでこの$\overrightarrow{BC}$が床であることがわかればレーザーの範囲に障害物が無いと判断できる．そこで最初にあらかじめ床までの距離$\overrightarrow{|AO|}$を測定しておく．$\overrightarrow{BC}$の点群の集まりが直線かつ，あらかじめ測定した$\overrightarrow{|AO|}$より小さい測定値が出なければ$\overrightarrow{BC}$の直線は床であると判定できる．すなわち直線を検出できなかった場合，または直線を検出したとしてもあらかじめ測定した床までの距離$\overrightarrow{|AO|}$より近くに存在する場合はロボットの前方に障害物あるいは壁が存在すると判定する．

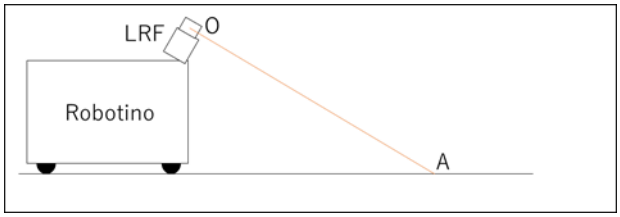


図 4 LRF を水平面と異なる角度で
取り付けた時の検知範囲

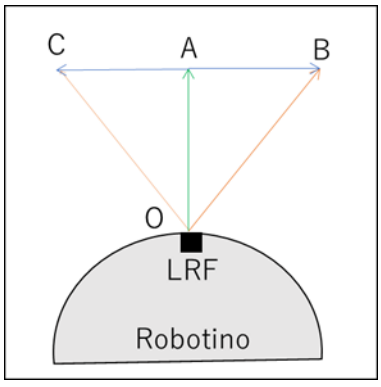


図 5 図 4 を上から見た図

4 実験

実際に 3 章の検知方法で障害物を検知できるかを確認するために，実験を行う．

実験にあたって，LRF を 30 から 60 度の範囲で Robotino 前方に傾ける．傾ける角度を 10 度ステッ

プで変更しながら実験を行う．Robotino に LRF を傾けて取り付けるために 3D プリンタで台座を作成した．図 6 に作成した台座と LRF を台座に取り付けた様子を示す．これを地面から 21.5cm の高さに取り付ける．
障害物として 11[cm]×7[cm]×5[cm](幅×奥行×高さ)の小箱を設置する．これを Robotino の正面に置き，LRF を傾ける角度ごとに，小箱を図 5 中の線分 BC 上にあたる位置で左右に動かしながら，検知できる範囲を測定する．これを床の上で行う．
図 7 に実験風景，そして表 2 と図 8 に実験結果を示す．

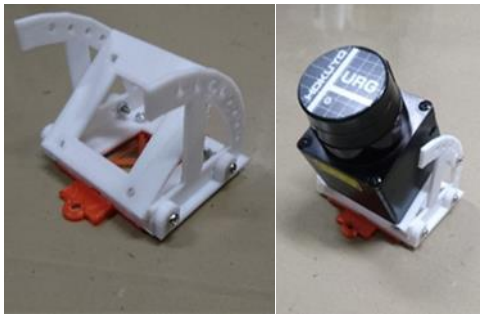


図 6 実験に使用した LRF 台座

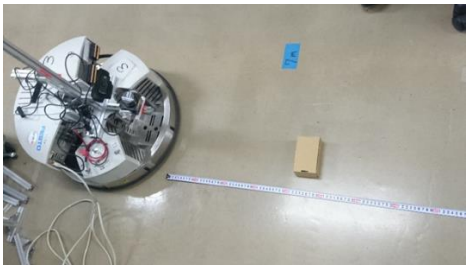


図 7 実験風景

実験結果より，検知範囲はいずれも理論値の 500mm を上回る結果となった．

表 2 実験結果

LRFを傾けた角度	30度	40度	50度	60度
検知範囲(床)[mm]	558	563	548	552

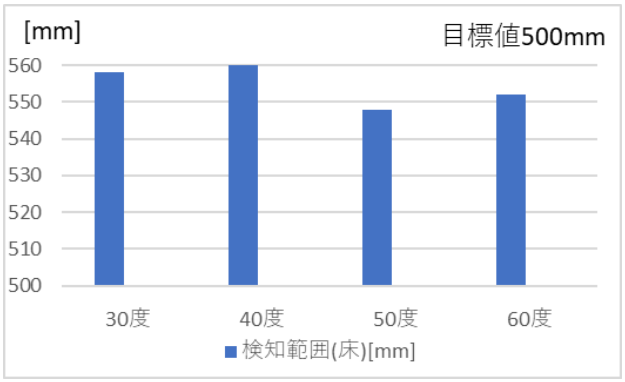


図 6 実験結果

5 おわりに

近年，少子高齢化や人件費の高騰により，様々な業種でロボットによる自動化の需要が高まっている．本研究では警備ロボットのような移動式ロボットに注目し，測域センサをロボットの前方に傾けて取り付けることで，従来の地面と水平にセンサを取り付ける方法では検知できなかった障害物を検知することを試み，その効果について測定した．測定の結果 Robotino の直径と同じ 500mm の範囲であれば障害物を検知できることがわかった．

今後の課題として実際にロボットに実装することを考えて，障害物の有無の判定だけではなく，位置の検出を行うことが考えられる．

参考文献

[1] 移動式ロボット Robotino
https://www.festo-didactic.jp/jp-ja/news/robotino.htm
(2020 年 1 月 16 日 閲覧)
[2] データ出力タイプ/URG-04LX-UG01
https://www.hokuyo-aut.co.jp/search/single.php?serial=17
(2020 年 11 月 9 日 閲覧)

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-057-19 (11/21)

タグマーカーを用いた自律移動ロボット間の自己位置推定に関する一考察

About localization between autonomous mobile robots using tag markers

鈴木　勇貴[1]　植村　渉[1]

Yuki Suzuki[1] and Wataru Uemura[1]

[1]龍谷大学
[1]Ryukoku University

Abstract: 近年，少子高齢化による労働者不足や，人件費の高騰から様々な業界でロボットによる自動化の需要が高まっている．ロボットの自律移動のためには自己位置の推定が必要である．主に距離センサを用いた自己位置推定の手法が使用させているが，港など周囲に形状の特徴物が少ない環境では，自己位置推定の精度が低下する．また，カメラを用いた自己位置推定の手法もあるが，ロバスト性が低いという課題と画像特徴が少ないことから同様に低下する．解決策として，タグマーカーをカメラで認識することで位置推定が可能になるが，移動はタグマーカーが設置されている範囲に限定されてしまう．本研究では，複数台の自律移動ロボットにタグマーカーを取り付け，他の自律移動ロボットのタグマーカーを認識することで位置推定をし，協調的に動くことで活動範囲を広げていく自律移動ロボット間の自己位置推定の手法を提案する．

1 はじめに

近年，少子高齢化による労働者不足や，人件費の高騰から様々な業界でロボットによる自動化の需要が高まっている．ロボットの自律移動のためには自己位置の推定が必要である．自己位置の推定には主に距離センサやカメラを用いる．距離センサではLiDARSLAM[1]が，カメラではVisual SLAM[2]が主に自己位置推定の手法に使用されている．港など周囲に形状の特徴物が少ない環境で使用する場合，距離センサは特徴点の低下により．自己位置指定の精度が低下する．また，カメラでもロバスト性が低いという課題と画像特徴が少ないことから同様に低下する．解決策として，タグマーカーをカメラで認識することで位置推定が可能になるが，移動はタグマーカーが設置されている範囲に限定されてしまう．

本研究では，複数台の自律移動ロボットにタグマーカーを取り付け，他の自律移動ロボットのタグマーカーを認識することで位置推定をし，協調的に動くことで活動範囲を広げていく自律移動ロボット間の自己位置推定の手法を提案する．

2 自律移動ロボットとタグマーカー

本章では，本研究で扱う自律移動ロボットRobotino[3]とタグマーカーについて述べる．

2.1 自律移動ロボット：Robotino

本研究で使用するロボットはFesto社が販売している，全方位移動ロボットRobotino(図１)である．また，Robotinoのコントロールプログラムは，robotinoview[4]を使用して作成および実行することができる．本研究では，robotinoviewからロボットの車輪やステアリングの回転角度から計算し，それぞれの移動距離を求め，その累積からロボットの位置をする推定する手法のオドメトリの情報を取得する．

図 1 Robotino

2.2 タグマーカー

タグマーカーは，カメラ一台による撮影でカメラとの相対位置と姿勢と ID 番号を認識できる平面パターンである画像計測ツールである．様々なタグマーカーが提案されているが，本研究では図２の 6x6 のドットで情報を表現しているタグマーカーを使用する．外円の２つのドットは黒で固定され，それ以外の白黒のパターンで ID 番号を表現する．

図 2 タグマーカー

2.2.1 タグマーカーを利用した位置推定

タグマーカーの大きさと回転角，そして正方形からのゆがみ具合から，仮想的な三次元空間上の位置と姿勢が計算できる[5]．また，位置と姿勢はカメラ座標系Cから見たタグマーカー座標系Mの位置と姿勢を表す同次変換行列 ${}^{C}H_{M}$ として表すことができ，この逆行列 ${}^{M}H_{c}$ を求めることで，タグマーカー座標におけるカメラの位置と姿勢が得られる．タグマーカーは標準的なカメラで認識が可能なため，複数台のロボット間でシステムの共有が可能である．

3 提案手法

本章では，タグマーカーを用いた自律移動ロボット間の自己位置推定手法とロボットの実装方法を述べる．

3.1 タグマーカーを用いた自律移動ロボット間の自己位置推定手法

n 台の自律移動ロボットから，基準位置となる自律移動ロボットを一台選び $Robotino_0$ とする．この $Robotino_0$ に対して, $Robotino_m$ が移動する$(1 \leqq m < n)$. 全ての自律移動ロボットには，タグマーカーとカメラが設置されている．$Robotino_m$ の移動範囲は $Robotino_0$ のタグマーカーの位置や姿勢が取得を行える範囲である．$Robotino_m$ の移動後，$Robotino_m$ のカメラが $Robotino_0$ のタグマーカーの位置と姿勢を取得し，$Robotino_0$ の位置座標と合わせることで自己位置推定を行う．また同様に n 台の自律移動ロボットから，基準位置となる自律移動ロボットを一台選び $Robotino_0$ とし，$Robotino_m$ が移動し活動範囲を広げる．

3.2 ロボットの実装方法

自律移動ロボットである Robotino に設置するタグマーカーやカメラを示す．タグマーカーは一台の Robotino に対して３つ使用する．これは，$Robotino_0$ に選ばれた場合，$Robotino_m$ の最大移動範囲を考えると全方位にタグマーカーが見える必要があるためと $Robotino_m$ のタグマーカーを利用した位置推定はタグマーカーのゆがみ具合が大きいほど姿勢の推定の精度が上がる特徴があるためである．図３に実装した Robotino を示す．今回使用するカメラは，全方位カメラではないため $Robotino_0$ のタグマーカーを取得するために $Robotino_m$ がタグマーカーの見える向きに回転する必要がある．

図３実装した Robotino(右；横，左；前)

4 実験

本章は提案手法と従来法の自己位置推定を比較して評価を行う．従来法の自己位置推定としてオドメトリを使用する．

4. 1 実験方法

実験環境としては港など周囲に形状の特徴物が少ない環境で行うべきですが，大学にそのような環境がないため，大学の廊下で実験を行う．実験方法はそれぞれ自己位置を行い，指定した位置に移動し，

指定した位置とロボットの位置の差を評価基準とする．提案法では Robotino を 2 台使用する．図 4 に移動経路と指定位置を示す．

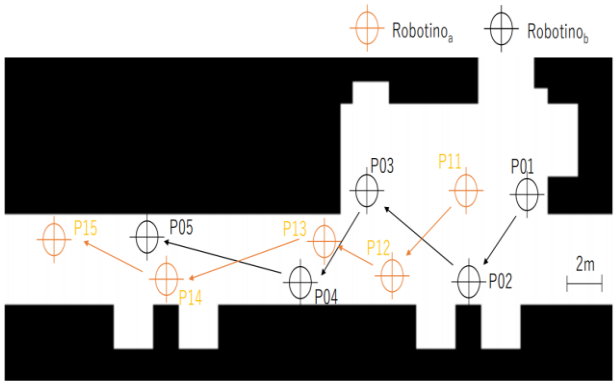


図 4 移動経路と指定位置

4.2 実験結果

図 5 に実験結果を示す．P02 と P12 の平均を Pa に P03 と P13 の平均を Pb に P04 と P14 の平均を Pc に P05 と P15 の平均を Pd に示す．縦軸は指定した位置と Robotino の距離の差を示す．

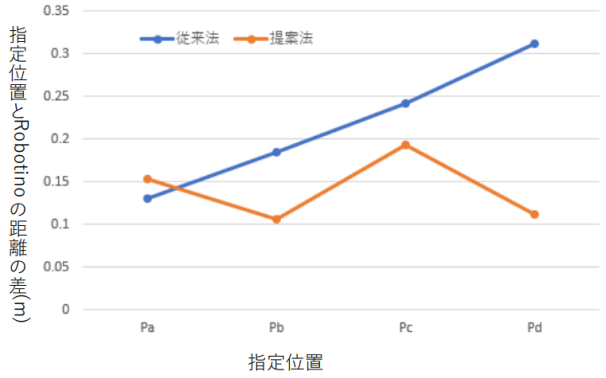


図 5 実験結果

5 まとめ

近年，少子高齢化による労働者不足や，人件費の高騰から様々な業界でロボットによる自動化の需要が高まっている．ロボットの自律移動のためには自己位置の推定が必要である．従来法の自己位置推定では，港など周囲に形状の特徴物が少ない環境では，自己位置推定の精度が低下する．

本研究では，複数台の自律移動ロボットにタグマーカーを取り付け，他の自律移動ロボットのタグマーカーを認識することで位置推定をし，協調的に動くことで活動範囲を広げていく自律移動ロボット間の自己位置推定を提案し，評価を行った．周囲に形状の特徴物が少ない環境での，提案法の自己位置推定に期待できる．

参考文献

[1] W. Hess et al., “Real-time loop closure in 2D LIDAR SLAM,” In Proceedings of the IEEE International Conference on Robotics and Automation, 2016

[2] M. Yokozuka et al., “VITAMIN-E: Visual tracking and mapping with extremely dense feature points,” In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[3] 移動ロボット Robotino
https://www.festo-didactic.jp/jp-ja/news/robotino.htm?fbid=anAuamEuNTYwLjE4LjE2LjI2MTA
(2020 年 11 月 10 日 閲覧)

[4] Robotinoview
https://www.festo-didactic.com/int-en/services/robotino/programming/robotino-view/?fbid=aW50LmVuLjU1Ny4xNy4zNC4xNDI2 (2020 年 11 月 10 日 閲覧)

[5] Kato, H., Billinghurst, M. (1999) Marker Tracking and HMD Calibration for a videobased Augmented Reality Conferencing System. In Proceedings of the 2nd International Workshop on Augmented Reality (IWAR 99). October, San Francisco, USA.

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-057-20 (11/21)

VRヘッドセットを用いたサッカー審判体験システム

An Experience System of Soccer Referee using VR Head Set

秋山英久[1*]　田中雄大[1]　齋藤涼太[1]　荒牧重登[1]
Hidehisa Akiyama[1], Yudai Tanaka[1], Ryota Saitou[1], Shigeto Aramaki[1]

[1] 福岡大学
[1] Fukuoka University

Abstract: The analysis of sports data has become increasingly popular in order to strengthen teams and players themselves. On the other hand, referees also have an important role to play in the fairness of the game. This paper proposes an experience system in order to provide a low-cost practice environment for soccer referees. The system assumes not only a practice environment but also a data collection system for soccer referees. In the experiment, we collect behavioral data from subjects and analyze them based on some evaluation criteria.

1　はじめに

スポーツにおいて公平かつスムーズに試合を実行するためには，選手だけでなく，審判も重要な役割を持つ．審判の行動データの収集と分析を進めることは，審判の育成や自動化に繋がると期待できる．審判の育成に関しては，審判の練習環境の不足も問題となっている．特にサッカーのようなチームスポーツにおいて，審判の練習環境を提供することは容易ではない．結果として，審判として充分な練習機会を得られないまま，大きな試合に臨むことになるかもしれない．

本研究では, これらの問題を解決するために，Virtual Reality デバイスとサッカーシミュレータを用いることで，サッカーの審判体験とデータ収集を行うシステムを開発した．このシステムは，審判にとっての練習環境を低コストで提供することも想定している．本稿では，Virtual Reality デバイスとしてヘッドマウントディスプレイ型のデバイスを使用し，サッカーシミュレーション環境として RoboCup サッカーシミュレータを用いた．開発したシステムを用いて被験者実験を行い，行動データの分析を行う．

2　関連研究

スポーツデータの分析では，個々の選手の能力やチームの強さを高めることを想定した取り組みが多い [6, 1]．このような試みは特にプロサッカーの試合データ分析で盛んであり，様々な分析対象と手法が提案されている [7, 8, 9]．

一方で，スポーツでは審判も重要な役割を持つ．審判は，選手以上にその競技を理解し，選手とともに試合へ参加し，そして，選手と協働して試合を進行しなければならない．より良い審判を育成することはスポーツにおいて重要な課題であり，審判に関するデータ収集と分析の試みが進められている [2, 3, 12]．

スポーツデータの可視化と分析を効率良く行う手法として，Virtual Reality（VR）環境は一つの有望なアプローチである．田尻らは VR 環境を用いた学習の実用性と有効性について述べている [11]．空間認識能力の習得に関しては，その習得度合いに個人差が現れるとされているが，VR 環境で空間認識を疑似体験することで，学習者の個人差を小さくすることができると期待されている [5, 10]．

近年はヘッドマウント型ディスプレイを用いた没入型の VR 環境が安価で利用可能になっている．本研究では，ヘッドマウント型ディスプレイを用いたサッカー審判体験システムを開発する．

3　サッカー審判体験システム

3.1　システム概要

本研究で開発したサッカー審判体験システムでは，使用者が装着するデバイスとして，ヘッドセット型の Virutal Realty デバイス（以下，VR ヘッドセットと呼ぶ）とゲームパッドを用いる．システム使用者は，シミュレータが実行する試合を仮想フィールド上の審判

*連絡先：福岡大学工学部
〒 814–0180 福岡市城南区七隈 8-19-1
E-mail: akym@fukuoka-u.ac.jp

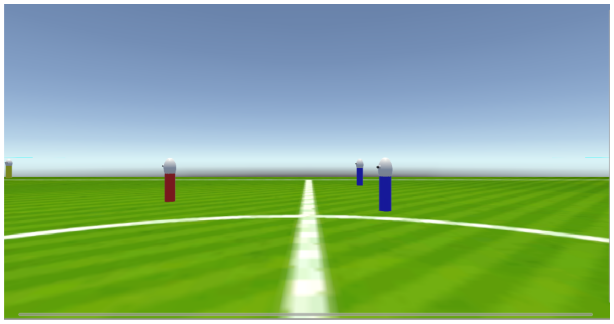


図 1: VR ヘッドセット装着者に表示される画面の例.

図 2: 本稿のシステムで使用する VR ヘッドセット.

視点で観察しつつ，仮想フィールド上で審判としての移動行動を実行する．

本稿で実装したシステムでは，体験する対象を主審とする．実際のサッカーでは，主審以外に副審として線審を複数配置する場合がある．本システムでは，自動で移動する線審を仮想空間内に配置する．

3.1.1 VR ヘッドセット

VR ヘッドセット内部の画面には，審判視点での仮想フィールドの状態が表示される．図 1 は，本システム使用者に表示される画面の例である．図にはサッカーフィールドに加えて審判の視界内に存在する選手オブジェクトが表示されている．選手オブジェクトと同様に，審判の体となるオブジェクトを仮想空間に作成しており，審判オブジェクトの頭の位置にカメラオブジェクトが設置される．このカメラオブジェクトが仮想空間内での装着者の視点であり，カメラの方向が仮想空間内での視野方向となる．実際に VR ヘッドセットを装着して画面を観ると，視覚的には自分の周囲に仮想空間が広がっているように認識される．これによって，仮想空間に没入した感覚でサッカーフィールドを観察することができる．

本システムでは，VR ヘッドセットとして VIVE 社の VIVE Pro（図 2）を使用し，プログラム開発はゲームエンジンである Unity を用いて行う．VIVE Pro はヘッドセット本体にジャイロスコープを内蔵しており，装着者の頭の姿勢をリアルタイムに検出可能である．

3.1.2 サッカーシミュレータ

本システムが担うのは審判視点の画面表示と審判の操作部分のみであり，サッカーのシミュレーション自体は別のシミュレーションシステムを用いる．本稿では，バックエンドのサッカーシミュレータとして RoboCup サッカーシミュレータ [4] を用いる．RoboCup サッカーシミュレータは，人間サッカーとほぼ同等のルールで 11 対 11 のサッカーを実現するシミュレーションシステムである．シミュレーションの空間は連続であるが，時間モデルは離散時間である．1 シミュレーションサイクルは 0.1 秒であり，各選手はシミュレーションサイクルごとに行動を決定し，物体の状態もシミュレーションサイクルごとに更新される．

RoboCup サッカーシミュレータには自動審判機能が既に組み込まれているが，本稿のシステムを接続して審判として介入可能とすることを想定している．ただし，現在の本システムの実装ではシミュレーション内容をリアルタイムに追従できず，シミュレータが記録したログファイルを再生することしかできない．そのため，現時点では，システム使用者は審判視点で試合を観察するのみであり，審判として試合に干渉することはできない．

3.1.3 審判の移動

現実の人間のサッカーでは，審判は試合中に自律的にフィールド上を移動している．より現実に近い体験を得るためには，我々のシステムでも審判が仮想フィールド上を自由に移動できる必要がある．本稿では，仮想フィールド上での審判の移動をゲームパッドで行い，ゲームパッドのスティックを傾けることで審判を任意の方向へ移動可能とする仕様とした．審判の運動モデルは簡略化し，移動スピードを一定とする．

3.1.4 審判の視野方向の変更

審判は仮想空間内のカメラを視野として持っており，VR ヘッドセット装着者はカメラが捉える範囲の物体しか認識できない．そのため，周囲の状況をより正確に認識するには，カメラの方向を変えることで視野方向を変更し，能動的に周囲の情報を収集しなければならない．本システムでは，装着者の頭の向きを仮想空間内でのカメラの方向へ反映し，審判の視野方向の変更として扱う．装着車の頭の向きは VIVE Pro によって測定可能である．VIVE Pro は装着者の視線を検出することはできないため，装着者の頭の向きを装着者が注目する方向とする．

3.1.5 審判の状態記録

本システムは仮想フィールド上の審判の状態を自動的に記録する．仮想フィールド上での審判の位置座標，視野方向，ゲームパッドでの操作内容を試合時間と同期させて記録する．ボールや選手の位置情報をシミュレーションログから取得できるため，ある時点の審判の行動に対応した仮想フィールドの状態を容易に取り出すことができる．

3.1.6 自動線審

現在の本稿のシステムでは主審の体験を想定しており，線審は自動で移動するオブジェクトを仮想フィールド上に 2 体配置する．ただし，線審オブジェクトは各チームの陣地に対応させて 1 体ずつ配置し，それぞれが異なるサイドラインを担当する．各線審オブジェクトは，各チームのオフサイドラインの移動に合わせて，フィールドの担当陣地側のサイドラインに沿って移動する．

3.2 審判の評価方法

審判の行動を分析するために，記録された審判の状態を用いて，審判としての妥当性を評価する．本システムを用いて仮想審判として行動した結果に対して，いくつかの評価指標に基づいて審判としての評価を行う機能を実装した．本稿では，評価指標として以下の 4 項目を設定する．これらはすべて，サッカー審判として明らかに妥当ではない状態であり，審判としての行動を失敗したとみなせる状態である．

1. 選手との接触 (c_1)
2. ボールとの接触 (c_2)
3. パスコースの妨害 (c_3)
4. 主審と線審でボールを挟めていない (c_4)

c_i をそれぞれの状態が試合中に発生した回数とする．各状態が発生したかどうかは試合のシミュレーションサイクルごとに自動的に検出する．全項目が罰則対象となる状態であるため，これらの値は負の評価値として扱うべきである．本稿では式 1 を審判の評価値 V_r とする:

$$V_r = -\sum c_i \tag{1}$$

各項目の詳細な定義を以下で説明する．

3.2.1 選手との接触

RoboCup サッカーシミュレータ上の選手は，半径 $0.3m$ の円としてモデル化されている．本システムで表示する仮想空間上では，シミュレータ上の大きさよりも拡大し，半径 $0.5m$ の円柱として選手オブジェクトの 3 次元モデルを作成した．同様に，審判オブジェクトも半径 $0.5m$ の円柱として 3 次元モデル化する．このモデルに基づいて，Unity の衝突判定機能によって審判と選手の接触判定を行う．

審判と選手のいずれも，行動時に体が傾くことはなく常に直立しており，常に地面に接しているものとする．いずれも円柱モデルであるため，地面平面上での審判と選手の中心位置の距離が $1.0m$ 以下の場合に衝突判定が発生し，審判は選手と接触したとみなされる．審判と選手の接触判定は，シミュレーションサイクルごとにすべての選手に対して行う．

3.2.2 ボールとの接触

選手との接触判定と同様に，Unity による衝突判定によって審判とボールの接触判定を行う．RoboCup サッカーシミュレータではボールは半径 $0.085m$ の円としてモデル化されているが，本システムでは半径 $0.5m$ の球として 3 次元モデルを作成した．シミュレータの仕様上ボールが空中を飛ぶことは無いため，ボールは常に地面に接しているものとする．よって，地面平面上での審判とボールの中心位置の距離が $1.0m$ 以下の場合に衝突判定が発生し，審判はボールと接触したとみなされる．審判とボールの接触判定は，シミュレーションサイクルごとに行う．

3.2.3 パスコースの妨害

審判が選手間のパスコースを妨害しているかどうかを判定し，罰則として評価に反映する．選手がボールをキック可能な状態である場合に，他の味方選手へのパスコース上に審判が位置していればパスコース妨害と判定する．ボールを持つ選手と他の味方選手とを結ぶ平面上の線分を考え，その線分と審判との平面上の距離が $1.0m$ 以下の場合に審判がパスコース上に位置していると判定する．パスコースの妨害判定は，シミュレーションサイクルごとに行う．

3.2.4 主審と線審でボールを挟めていない

審判が 1 人だけであれば，ボールが審判の死角に位置してしまう場合がある．より正確な判定を行うために，現実のプロサッカーの試合では主審に加えて 2 人の線審が配置され，3 人の審判からボールに対する死角

が同時に発生しないようにすることが求められる．通常は，いずれかの線審とボールを挟む位置に移動することが主審には求められる．

本システムでは 2 体の自動線審を追加しており，システム使用者は主審として振る舞うことを想定している．よって，システム使用者は，いずれかの線審の位置に合わせてボールを挟めるように移動する必要がある．主審の位置を R，線審の位置を L，ボールの位置を B とすると，ボールを中心とした角度 $\angle RBL$ が 90° 以下の場合に，本システムでは主審とその線審でボールを挟めていないと判定する．この角度判定を 2 体の線審に対して行い，いずれの線審ともボールを挟めていない場合に主審の位置が妥当ではないと判定する．この主審位置の妥当性判定は，シミュレーションサイクルごとに行う．

4 実験

実装したシステムを用いて被験者実験を行う．複数の被験者にシステムを使用してもらい，被験者に対して審判としての評価を行うとともに，その行動内容のログを収集して分析を行う．

4.1 実験設定

10 名の被験者に対してデータ収集と分析を行う．すべての被験者は 20 歳代の男子大学生である．ただし，半数の 5 名は実際の人間サッカーで審判の経験を持つ．被験者は，本システムの評価項目について事前に説明を受けてから実験に臨む．今回の実験では審判は試合に干渉せず，再生される試合を審判視点で観察するだけであるため，全ての被験者に対して同一の試合を再生する．審判の初期位置はフィールド中央付近のボールに干渉しない位置とし，試合の再生開始後は被験者が任意に移動行動を可能とする．

被験者からのデータ収集時間は試合再生開始から約 3 分間とし，シミュレーション上は 1800 サイクル分の試合再生を行う．RoboCup サッカーシミュレータでは 1 サイクル 0.1 秒で 6000 サイクルを 1 試合としており，本来であればロスタイムを含めると 1 試合は 10 分強である．しかし，VR ヘッドセットを装着した状態での長時間の集中が困難であり，経過時間とともに審判としての行動に影響が発生したため，データを収集する時間を短縮した．

4.2 実験結果

表 1 に各被験者の評価結果を示す．被験者 E1 から E5 は人間サッカーでの審判の経験を持ち，N1 から N5 は未経験者である．

表 1: 各被験者の評価値

被験者 ID	評価値 (V_r)
E1	-63
E2	-68
E3	-60
E4	-108
E5	-63
N1	-145
N2	-102
N3	-107
N4	-70
N5	-139

図 3, 4, 5, 6 は評価項目ごとの結果である．. 選手との接触とボールとの接触を比較すると，選手と接触する頻度の方がより高いことが分かる．被験者は選手よりもボールにより注目していると推定され，特に，実際に審判の経験を有する被験者はボール接触の罰則に対してより良い結果を示している．選手との接触回数については全被験者でばらつきが大きく，平均的には経験の有無による違いは観られなかった．パスコースの妨害回数については，審判経験者のほうが平均的には良い結果を示しているものの，大きな違いは観られない．審判経験の有無がもっとも大きく現れたのは，線審とボールを挟む位置取りに関する評価項目である．

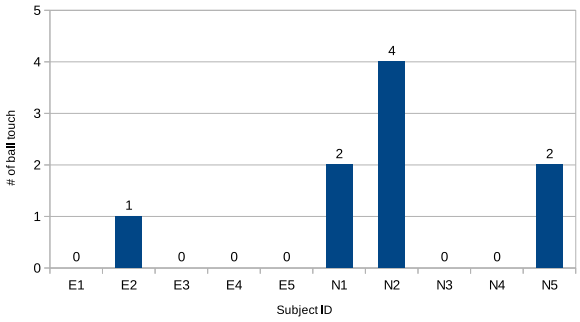


図 3: 各被験者のボールとの接触回数.

被験者に対する評価値と以下の 2 つの値との関係を調査した．

- 被験者が視界からボールを外した回数.
- ボールの移動距離に対する審判の移動距離の割合.

ボールの移動距離と審判の移動距離の関係については，先行研究として小林らの報告がある [12]．小林らは，良いパフォーマンスを示す審判の移動距離はボールの移動距離に対して 45% 程度であることを，実際のサッカーの試合データより示している．

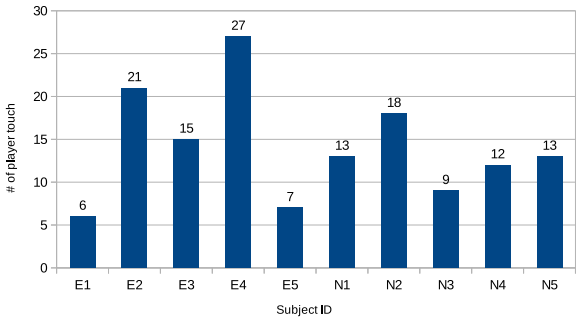


図 4: 各被験者の選手との接触回数.

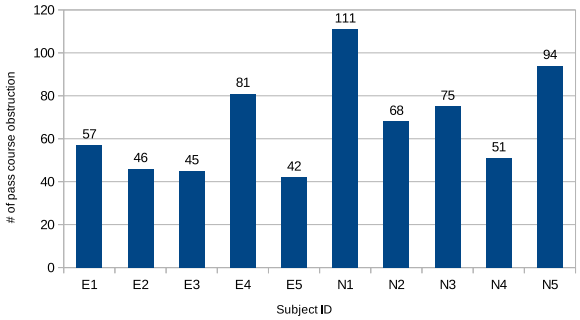


図 5: 各被験者のパスコース妨害回数.

図 7 は，各被験者の評価値と視界からボールを外した回数をプロットしたものである．これらの相関係数は 0.244 となり，強い相関は観られなかった．一方，図 8 は，各被験者の評価値と移動距離の割合との関係をプロットしたものである．これらの相関係数は −0.942 となり，強い負の相関が観られた．

4.3 考察

今回の実験結果からは，審判の経験を持つ被験者は，平均的には未経験の被験者よりも高い評価を得ている．そして，有経験者は，より妥当な位置取りができていることが分かる．図 9 と図 10 は，最良評価の被験者と最低評価の被験者の移動位置を平面上にそれぞれプロットしたものである．グラフの原点は仮想サッカーフィールドの中央である．これらのプロット図から，失敗が少ない審判は移動範囲が比較的小さく，フィールドの中央付近でより多くの時間を過ごしていることが分かる．これはボールの移動に追従しすぎていないことを意味し，図 8 の結果と一致する．

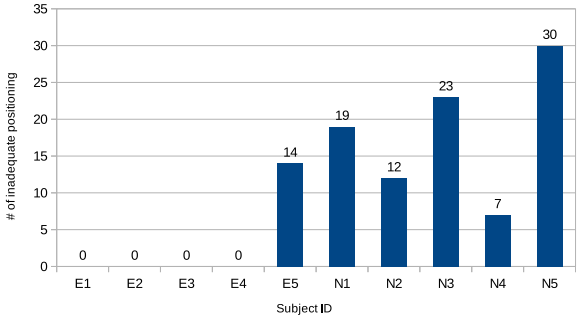


図 6: 各被験者が線審とボールを挟めていなかった回数.

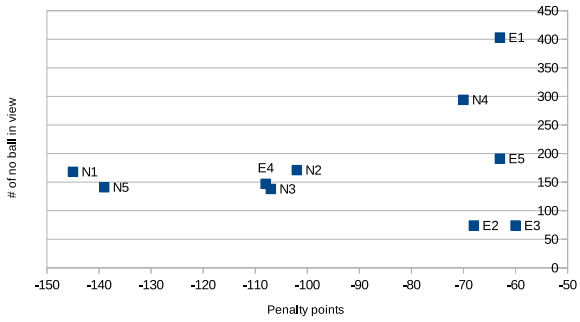


図 7: 各被験者の評価値とボールを視界から外した回数.

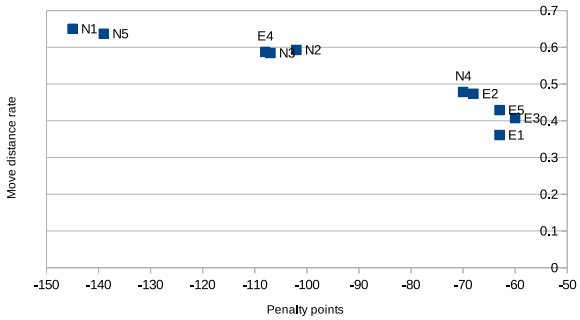


図 8: 各被験者の評価値とボールに対する移動距離の割合.

ボールを視界から外した回数と評価値との間には，強い相関は観られなかった．しかしながら，視野方向変更の意図が被験者によって異なる可能性がある．例えば，失敗が少なかった被験者は，周囲の状況を確認することを意図して視野方向を頻繁に変更していたと予想される．より多くのデータを収集し，各被験者が何に注目していたかをより詳細に分析する必要があるだろう．

5 まとめ

本稿では，VR ヘッドセットデバイスを用いたサッカー審判体験システムを開発した．開発したシステムを用いた被験者実験を行い，仮想空間上での審判の行動データの収集と分析を行った．実験結果から，審判

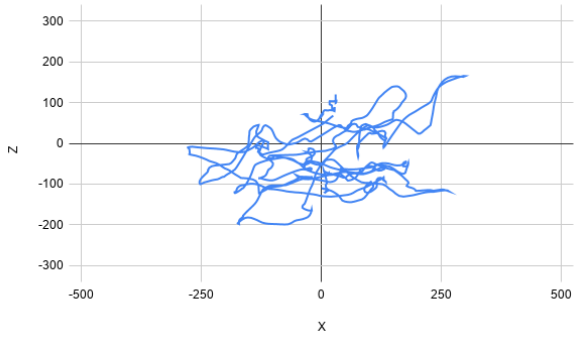


図 9: 評価値がもっとも高かった被験者の移動位置プロット図.

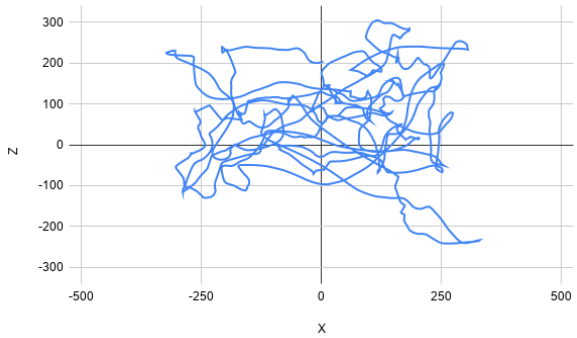


図 10: 評価値がもっとも低かった被験者の移動位置プロット図.

の経験を有する被験者は，未経験の被験者に比べて位置取り行動での失敗が少ないことが示唆された．

現在のシステムでは，審判としての判定に必要な情報を観察できていたどうかを評価できていない．この評価指標を反映するために，システム使用者がリアルタイムに試合に参加して，審判として判定を下すことができる機能を追加する必要がある．人間が下した判定と自動審判が下した判定を比較することで，審判の評価精度を高めることができる．その結果を用いて，審判視点での移動行動をより詳細に分析することが今後の課題である．

参考文献

[1] Brooks, J., Kerr, M., Guttag, J., “Developing a data-driven player ranking in soccer using predictive model weights”, Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[2] Downward, P., Jones, M., “Effects of crowd size on referee decisions: Analysis of the FA Cup”, Journal of Sports Sciences, 25(14), pp.1541–1545, 2007.

[3] Hancock, D. J., Ste-Marie, D. M., “Gaze behaviors and decision making accuracy of higher- and lower-level ice hockey referees” Psychology of Sport and Exercise, 14(1), pp.66–71, 2013.

[4] Itsuki Noda, Hitoshi Matsubara, Kazuo Hiraki, Ian Frank : “Soccer server: A tool for research on multiagent systems”, *Applied Artificial Intelligence*, Vol. 12, No. 2-3, pp. 233–250 (1998)

[5] Dan Mikami, Kosuke Takahashi, Naoki Saijo, Mariko Isogawa, Toshitaka Kimura, and Hideaki Kimata, “Virtual Reality-based Sports Training System and Its Application to Baseball”, NTT Technical Review, vol.16, no.3, 2018.

[6] Hughes, M., Bartlett, R., The use of performance indicators in performance analysis, Journal of Sports Sciences, 20(10), pp.739–754, 2002.

[7] H. Sarmento, R. Marcelino, M. T. Anguera, J. CampaniÇo, N. Matos, and J. C. LeitÃo: “Match analysis in football: a systematic review”, Journal Sports Science, vol. 32, no. 20, pp. 1831–1843, 2014.

[8] Shaw, L., Glickman, M.: “Dynamic analysis of team strategy in professional football”. Barça Sports Analytics Summit, pp.1–13, 2019.

[9] B. Spencer, M. Hawkey, and S. Robertson: “Using contextual player movement and spatial control to analyse player passing trends in football”, Barça Sport. Analytics summit, pp. 1–12, 2019.

[10] 権藤聡志, 樽川香澄, 井上智雄, 岡田謙一: “スポーツの試合を再現した仮想空間を複数視点で提示する戦略分析支援システムの提案”, 情報処理学会研究報告, vol.2013-DCC-3, no.1, pp.1–5, 2013.

[11] 田尻圭佑, 瀬戸崎典夫: “HMD を用いた 3 次元ジェスチャ操作による没入型天体教材の開発”, 日本教育工学会論文誌 40 巻 ,Suppl. 号 ,pp.193–196, 2016.

[12] 小林久幸, 瀬戸進, 宮村茂紀, 川合悟, 瀬戸就一: “サッカーの級別主審の移動距離とボールの移動距離に関する研究”, 日本体育学会大会号, 44B 巻 ,p.726, 1993.

実機自律移動ロボット競技大会における無選手試合の提案と課題

About an Autonomous Robot Competition without Participants at the Competition Venue

植村　渉[1]*

[1] 龍谷大学
[1] Ryukoku University

Abstract: Covid-19 の影響により，世界的にロボットの競技大会が中止や延期となっている．一方でロボットの自律移動や自動制御を競う大会は，競技が始まると選手はロボットに触る必要が無いため，会場にいなくても競技できる可能性がある．しかし，ロボットの調整，特にハードウエアの調整には，実際にロボットを触る必要があり，遠隔での実施は難しい．そこで本研究ではロボットを大会会場に送り込み，現地運営スタッフが選手の代わりにロボットを触ることで大会を実施する無選手試合を提案し，その問題点や課題を検討する．また，同条件下で 2020 年 9 月 19 日から 22 日にかけて実施した RoboCup JapanOpen 2020 Logistics League についても報告し，with コロナのニューノーマルな世界でのロボット競技大会の形を模索する．

1　はじめに

世界的なロボット技術のデファクト･スタンダード化を推進する研究として，National Institute of Standards and Technology（NIST:米国標準技術研究所）では Joe らによりロボットアームによるパーツの取り付けやケーブルの取り回しのタスクが提案されている [1]．現在の工場で必要とされている加工技術に加え，これからの工場の組立作業で必要となる技術が含まれている．また，横小路らによる組立タスク [2] では，工場のラインの最適化を想定しており，作成する製品が変わったときの段取り替え時間（リードタイム）を減らし，究極には一品物をラインで作ることを目標としてタスクを設定している．そのためには，作成する製品の仕様が直前に変わる仕組み（サプライズ）が用意されており，現状の教示を中心とした固定的なロボットアームの動作では対応できず，柔軟な動きができるロボットが求められている．しかし，COVID-19 の影響により，実機ロボットの大会の中止が相次いでいる．

自律移動ロボットの世界大会である RoboCup（2020 年 6 月フランス，一年延期）[3] や，若年者ものづくり競技大会のロボットソフト組込職種（2020 年 7 月，中止）[4]，東京オリンピックに対してロボットのオリンピックを開催しようという目的で企画された World Robot Summit（2020 年 10 月，翌年に延期）[5] などである．大会で用いるロボットが共通仕様の場合は，プログラムを会場に送ることで競技を実施する大会がある．2020 年 9 月に AWS DeepRacer フィジカルリモートレース [6] が実施された．この大会は，車型の共通仕様のロボットに対して，参加者のプログラムを動かし，ソフトウエアで制御する仕組みである．そして，各参加者は，持ち時間 4 分の中で，人手や強化学習などでの調整を経て，コースを走る車を仕上げ，最速の周回ラップを競う．ハードウエアが共通仕様であり，ハードウエアに対する調整が不要なため，遠隔で実施することができた大会である．各チームがロボットを作り込んでいる場合，選手がロボットを触って調整する必要がある．技能五輪移動式ロボット職種（2020 年 11 月，無観客試合で開催予定）[7] は 1 チーム 2 名という人数制限があるため，三密を避けることで開催する予定である．しかし，人数制限のない大会では会場での三密を避けることができないため，各チームが自分の研究室で協議を行い，それを中継することで大会として扱う．ミニ四駆 AI 大会（2020 年 9 月実施，中継）[8] や International Conference on Intelligent Robots and Systems (IROS)2020 Open Cloud Robot Table Organization Challenge（2020 年 10 月，中継）[9] などである．国際大会は，国外の選手が開催国に入国する必要があるが，2020 年 10 月現在，日本では入国者には少なくとも 2 週間の隔離を必要とするため，実質，国外の選手が日本の国際大会に参加することができない．一方で，IROS のように各チームを中継で結んで競技することも可能であるが，選手やロボットの動きを全て把握することができず，不正が行われる危険性がある．また，競技直前にチームへ通知するサプライズ

*連絡先：龍谷大学先端理工学部電子情報通信課程
〒 520-2194 滋賀県大津市瀬田大江町横谷 1-5
E-mail: wataru@rins.ryukoku.ac.jp

の部品を扱うことも難しい．

本研究では，そのような条件下において自律移動ロボットの実機競技の実現性について検討する．具体的には，RoboCup Logistics League[10, 16, 17] を対象とする．この競技大会では，移動式ロボットの部分は全チーム共通であるが，物を掴むためのハンドやグリッパー部分は各チームが作る必要があり，それに関連するセンサー類もチーム毎に異なる．ロボットシステムの土台部分は各チーム統一されているが，上に載っている物がチーム毎に異なるため，ハードウエアが共通の競技と個別の協議の両方の特徴を持つ．遠隔操作に関連する対象としては，選手，審判，そして運営の 3 者になるため，それらの組み合わせにより，開催に必要な項目を検討する．

本節では，本研究で対象とする競技大会である RoboCup Logistics League (RCLL) について説明する．RCLL は工場のオートメーション化を想定した自動搬送車のプランニングを主とした競技である．作成する製品を構成する要素としてワークを用い，ベース，リング，キャップの 3 つの素材で構成される．それぞれ色が数種類あり，製品のオプションによってリングの数が変わる．現在は 0 個から 2 個までの注文が用意されてい.．それらを加工する機器としてフエスト社製の Modular Production Systemt（MPS）を用いる．MPS は，モジュラー単位で装置を組み付けることができるシステムになっており，工場のラインと加工機器を想定して装置が組み付けられている．MPS は 5 種類あり，いくつか同じ種類の MPS が配置されているが，作業対象が異なるため，同一の MPS は存在しない．いずれの MPS にも，素材や製品を扱うためのベルトコンベアが中央に置かれており，MPS の種類によって，ベルトコンベアを経由してロボットからワークを受け取ったり，加工したりし，その作業結果のワークを再度ベルトコンベアに置いて，ロボットがそれを受け取るという仕組みになっている．RCLL では，移動部は各チーム共通で，例えばモーターの種類を変更するなどの手を加えることは許されていないが，センサーの追加やワークを掴むためのグリッパーの追加は各チームに任されている．

2020 年現在のルールではフィールドは 14m × 8m の広さで，試合の開始時に各チームに対して 7 台，計 14 台の Modular Production Systemt（MPS）を配置する．工場のラインが頻繁に変わることを想定しており，MPS の配置は試合ごとにランダムに決定される．これらの MPS を区別するために側面にマーカーが付けられている．探索フェーズでは，頻繁な段取り替えへの対応を課題としており，ロボットに MPS の場所や書類の情報は伝えられておらず，それを調べることで得点となる．ロボットはマーカーを見ることでどの場所のどの向きにどの種類の MPS が置かれているかを把握して，審判のプログラムに伝える．

審判のプログラム（RefBox）は，各チームのロボットに対して競技に必要な注文情報などを伝え，ロボットは自分の位置情報などをビーコンとして返信する．これらの通信には ProtoBuff[11] を利用しており，RefBox から両チームに配信する情報は平文で送るが，チームから送信する情報は相手チームに見られると困るため各チームに配布した鍵情報で暗号化して送る．MPS の制御も RefBox の仕事であり，ロボットから「ワークを置いた」とか「黒色のベースを提出して欲しい」といった要求を受け取り，MPS にその指示を出す中継器の仕事も行う．また，ロボットが指示を間違えたときや，定期的なタイミングで MPS は利用休止状態になるが，その指示も RefBox が扱う．なお，RefBox とロボット間の通信には．ProtoBuff を用い，RefBox と MPS 間の通信には OPCUA のプロトコルが使われている．

このように Logistics League は，ハードウエアの作成から通信プロトコルの利用，そしてソフトウエアの作成と幅広い技術が求めらる．それゆえ，新規参入が難しく，参加チーム数少ない．そのため，Logistics League の技術を要素に分解し，それぞれを競うことで，新規チームの参入を促す競技が提案されている．これを Technical Entry Challenge (TEC) と呼ぶ．TEC では，要素技術として 5 つのスキルに分解し，それぞれに対して 3 通りの難易度を設定している（表 1 参照）．5 つのスキルは，1) 移動（Driving），2) 位置（Positioning），3) 認識（Detecting），4) 把持（Gripping），5) 通信（Communication）となっている．各チーム，競技開始前にスキルを選択し，挑戦する．成功したスキルは，選べなくなるが，失敗したときは再度挑戦することができる．

2020 年 9 月に龍谷大学で実施した RoboCup Japan Open 2020 Logistics League [14] は，この TEC を扱っている．若年者ものづくり競技大会に Robotino で参加していた 3 高校 4 チームが初参加し，龍谷大学の BabyTigers-R[15] と合わせて 5 チームで技術を競った．

2　遠隔対応

まず，遠隔で実施するにあたり会場である龍谷大学と各参加チームとをインターネット経由でつなぐ必要がある．ネットワークの接続図を図 1 に示す．ロボットシステムそのものの組み立てを含むハードウエアの調整や，会場のネットワークへの接続に関しては会場での人手が必要となる．それらの作業を会場にいる運営スタッフが代行することで，無選手試合を実現する．各チームと大会会場とをインターネットでつなぎ，ビデオ会議などを利用して選手が会場スタッフに指示を出し，会場での調整を行う方法である．

表 1: RoboCup Logistics League Technical Entry Challenge の課題 [13]

Level	Points	Topics				
		Driving	Positioning	Detecting	Gripping	Communication
3	50	Nbr 13	Nbr 23	Nbr 33	Nbr 43	Nbr 53
2	30	Nbr 12	Nbr 22	Nbr 32	Nbr 42	Nbr 52
1	10	Nbr 11	Nbr 21	Nbr 31	Nbr 41	Nbr 51

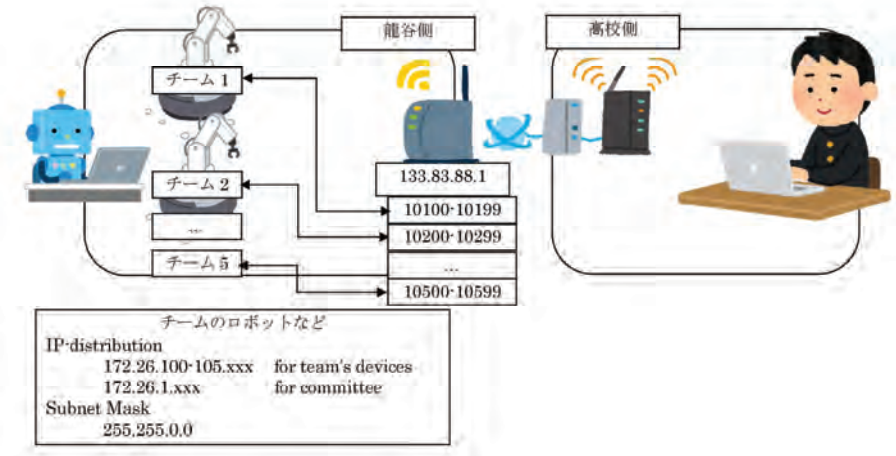


図 1: 大会会場と各チームとの間のネットワーク接続イメージ図

なお，所属する高校（県）によるインターネット利用ポリシーもそれぞれ異なり，ソフトウエアの制限や使えるネットワークプロトコルやポートの制限もバラバラで，それらに対する解除申請も一律に進める事ができなかった．

オンライン化で考慮すべきことは，選手・運営・審判のそれぞれの視点で考える必要があるが，実際はそれぞれの 2 組間（もしくは 3 組間）でのやり取りが主となる．それらの項目を表 2 に示す．

また，2020 年 9 月に実施した大会では，会場のフィールド周囲に 8 つのカメラを設置し，オンライン会議システムである zoom にそれぞれのカメラを接続した 2. 選手も同様に zoom に接続し，話者を選択する形でカメラのつながっているパソコンを選ぶことで見たい視点を選べるようにした．今後は Virtual Reality（VR）を利用し，選手があたかも会場にいるように周囲を見渡せるようにすることで，調整が行いやすくなると考えられる．移動式ロボットに VR 用のカメラを設置することで，会場内を自由に移動する方法も可能であると考える．

3 まとめ

ロボット技術は特に工業の分野において発展しているが，それぞれの分野に特化しているため，技術を客観的に測ることが難しくなっている．そのため，工業分野で必要とされる標準的な技術を盛り込んだ課題を用意することにより，各技術の相対的な比較ができるようになると共に，この先の技術目標を見据えることができるようになる．このように世界的なロボット技術のデファクト・スタンダードが必要とされている中で，我が国発の目標設定に基づいた技術の競争の場（大会）を設定することによりデファクト・スタンダード化を推進するプロジェクトが進んでいる．一方で，COVID-19 の世界的な流行によって，ロボットを始めとする競技

図 2: 競技フィールドの周囲にカメラを設置し，オンライン会議システムに接続．競技者はカメラのアカウントを選択することで画面を大きくすることが可能．

表 2: オンライン大会の視点での選手・運営・審判の相互関係

■ 選手-運営間のやり取り（選手から運営への要求事項）

- ハードウエア（ロボットやセンサ，取り付け機器）の調整
  - 自由に操作できるカメラ（現在は現地スタッフが手足や目となって対応）
  - センサの感度，取り付け位置，ロボット設置位置，ケーブルなどの調整
  - ハードを触ることによる調整
- 現地の状況の確認
  - センサのノイズ源の確認
  - ロボットとフィールド上の物体との位置関係の確認（あとちょっと右　など）
- ネットワーク（不調）への対応
  - 無線 LAN が途切れてロボットと通信できない（他チームの影響）
  - テレビ電話における快適な通話（ノートパソコンのマイクにノイズが乗る）
  - 遠隔操縦の方法（ロボットに直接ログイン，zoom などの遠隔操作機能の利用）

■ 選手-審判官のやり取り（審判から選手への要求事項）

- 非公開の課題の扱い
  - ランダム配置やサプライズ情報を選手に見せない方法
  - 選手が不正をしないことを確認する方法
- 制限時間の必要性（選手ではなく運営が作業をするため）

■ 運営-審判官のやり取り

- 両者ともに現地にいるため従来と変わらない．ただし，課題が見つかる可能性はある．

大会は軒並み中止になったり延期になったりしている．各チームでロボットを動かして，それを中継することで競技を行う大会もあるが，公平性の担保が難しい．このような COVID-19 の影響下において，実機ロボットの大会を開催するために，無選手試合を行った．ロボットを会場に送り，選手は会場の運営スタッフとやりとりして遠隔から参加することで，公平性を確保したロボット競技大会となる．しかし，選手が遠隔で参加するため，今までの大会と同等の質を保証することは大変難しい．これに対して，選手，運営，審判のそれぞれの立場で今までの大会と同じ質を保証する工夫が必要である．それらの組み合わせを元に，必要となる要素を検討した．

これからのニューコロナの世界では，実機ロボットの大会は実際に行うにしても会場では実施できないため，各研究室でロボットを動かしてライブ配信（動画配信）するなどの方法になっていくと考えられる．会場で実機を動かしながら大会を実施するためにも，本研究の要素項目が重要となる．もしコロナが収束しても，海外の会場に選手全員で移動しなくても参加できる方法となるため，需要のあるロボットの大会の実施方法になると考えられる．

参考文献

[1] Joe Falco, Kenneth Kimble, Karl Van Wyk, Elena Messina, Yu Sun, Mizuho Shibata, Wataru Uemura, and Yasuyoshi Yokokohji. “Benchmarking Protocols for Evaluating Small Parts Robotic AssemblySystems”. IEEE Robotics and Automation Letters, Vol. 5, pp. 883–889, (2020).

[2] Yasuyoshi Yokokohji, Yoshihiro Kawai, Mizuho Shibata, Yasumichi Aiyama, Shinya Kotosaka, Wataru Uemura, Akio Noda, Hiroki Dobashi, Takeshi Sakaguchi, and Kazuhito Yokoi. “Assembly Challenge: a robot competition of the Industrial Robotics Category,World Robot Summit – summary of the pre-competition in 2018”. Advanced Robotics, Vol. 33, pp. 876–899, (2019).

[3] ロボカップ日本委員会ニュース，“RoboCup2020 Bordeaux（ロボカップ世界大会ボルドー）は 1 年延期されます”, http://www.robocup.or.jp/news/entry-133.html, (2020 年 11 月 09 日 閲覧)

[4] 厚生労働省報道発表資料，“第 15 回「若年者ものづくり競技大会」の開催を中止します”,

https://www.mhlw.go.jp/stf/newpage_10848.html,
(2020 年 11 月 09 日 閲覧)

[5] 経済産業省ニュースリリース,
“World Robot Summit 2020（ワールドロボットサミット 2020）の開催を延期します”,
https://www.meti.go.jp/press/2020/04/20200417001/20200417001.html,
(2020 年 11 月 09 日 閲覧)

[6] aws, “AWS DeepRacer フィジカルリモートレース”,
https://pages.awscloud.com/ DeepRacerPhysicalRemoteRace_01.LandingPage.html,
(2020 年 11 月 09 日 閲覧)

[7] 厚生労働省報道発表資料, “「第 58 回技能五輪全国大会」と「第 40 回全国アビリンピック（全国障害者技能競技大会)」を無観客で 11 月に開催します”,
https://www.mhlw.go.jp/stf/newpage_13026.html,
(2020 年 11 月 09 日 閲覧)

[8] ミニ四駆 AI 大会, “ミニ四駆 AI 大会 in FSS2020”,
https://sites.google.com/site/ai4wdcar/home/taikai/fss2020?authuser=0,
(2020 年 11 月 09 日 閲覧)

[9] International Conference on Intelligent Robots and Systems (IROS), “Robotic Grasping and Manipulation Competition”,
https://rpal.cse.usf.edu/competition_iros2020/,
(2020 年 11 月 09 日 閲覧)

[10] RoboCup Logistics League,
https://ll.robocup.org/home/,
(2020 年 11 月 09 日 閲覧)

[11] RoboCup Logistics League, Referee Box
http://www.robocup-logistics.org/refbox (2020 年 11 月 09 日 閲覧)

[12] Tim Niemueller, “Referee Box for the RoboCup Logistics League Integration Manual 2014”,
http://www.robocup-logistics.org/refbox/llsf-refbox-manual-2015.pdf&?attredirects=0&d=1
(2020 年 11 月 09 日 閲覧)

[13] “Technical Entry Challenge”,
https://vega.elec.ryukoku.ac.jp/trac/wiki/robocupLogisticsLeague/JapanOpen2020/rulebook
(2020 年 11 月 09 日 閲覧)

[14] “ロボカップ“無観客・無選手” 龍谷大でロジスティクスリーグ”, 日刊工業新聞 (2020/9/18 05:00),
https://www.nikkan.co.jp/articles/view/571675
(2020 年 11 月 09 日 閲覧)

[15] BabyTigers - R,
https://vega.elec.ryukoku.ac.jp/trac/wiki/BabyTigers-R
(2020 年 11 月 09 日 閲覧)

[16] 植村 渉, “RoboCup Logistics League におけるフィールド内の障害物検知に関する一考察’, 人工知能学会第 54 回 SIG-Challenge 研究会, pp. 28–30, (2020).

[17] 山北善輝, 辻和輝, 植村渉. “RoboCup Logistics League 用通信プログラムを搭載した組込機器の作成と評価”. 人工知能学会第 53 回 SIG-Challenge 研究会, pp. 14–17, (2019).

〒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/`