様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 ２５ 日現在

機関番号：８２６２６
研究種目：研究活動スタート支援
研究期間：2018～2019
課題番号：１８Ｈ０６４９０・１９Ｋ２１５５３

研究課題名（和文）深層ベイズ学習に基づく雑踏環境下でも頑健に動作する音源分離の教師なし学習

研究課題名（英文）Unsupervised neural source separation for crowded environments based on deep Bayesian learning

研究代表者

坂東　宜昭（Bando, Yoshiaki）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究員

研究者番号：４０８２８１６７

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：本研究の目的は、単チャネル音源分離を教師データのない大量の多チャネル混合音信号から学習する枠組みを構築することである。マイクアレイで観測される音の空間情報に着目し、分離された音源信号の空間的な尤もらしさに基づき音源分離を教師なし学習する。まず、混合複素ガウスモデルの変分償却推論に基づく音源分離の深層ベイズ学習を確立した。さらに、視聴覚統合に基づく教師なし音源定位へ拡張し、拡散性雑音が存在し音源数が未知の雑踏環境でも動作する手法を開発した。

研究成果の学術的意義や社会的意義
従来の統計モデルの知見と、独立して研究されがちな近年の深層学習の知見を統合し、音源分離の教師なし学習の枠組みを実現した。これまで、深層学習に基づく音源分離には、個別の音源信号である正解データを大量に準備する必要があり、実世界の様々な音源を分離するには原理上限界があった。そこで、統計的信号処理で研究されてきたブラインド信号処理の知見を深層学習に導入することで、教師データを用いずとも音源分離を学習できる枠組みを実現した。

研究成果の概要（英文）：This study aims to develop a framework that can train neural source separation by using multichannel audio signals without any supervision. We utilized the spatial information observed by a microphone array and trained the source separation based on the spatial likelihood of the separated signals. We first developed a deep Bayesian method that trains a neural source separation based on a complex Gaussian mixture model. This approach is then extended to an audio-visual source localization method that can deal with the diffuse noise and the unknown number of sound sources, which are problematic for recognizing the crowded real-world environments.

研究分野：統計的音響信号処理

キーワード：音環境認識　音源分離　深層ベイズ学習　償却変分推論

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ロボットなどの知的システムが周囲の環境に合わせて的確に応答・行動するには、計算機による聴覚機能が不可欠である。特に、観測した混合音から個別の音源信号を抽出する音源分離は、各音イベントの「どこで」「どんな」を推定する音源定位・識別の基盤技術として重要である。日常生活での使用に耐える音環境理解システムには、音源が無数に存在し動的に変動する雑踏環境での頑健な動作が不可欠である。研究開始当初時期には、以下の２つのアプローチが高い性能を発揮していたが、雑踏環境下では未だ課題が残っていた。

- **深層学習に基づく単チャネル音源分離**：Deep clustering (DPCL)や permutation invariant training (PIT)といった深層学習に基づく手法が、音声や音楽信号の分離といった応用で、各音源スペクトログラムの時間周波数構造に着目し圧倒的な性能を達成している。しかし、既存の深層学習の枠組みは、膨大な音源信号の正解データを教師として必要とし、実世界のあらゆる音源信号を収集することは事実上不可能で、雑踏環境といった一般の音源分離は困難という課題があった。
- **統計モデルに基づく多チャネル音源分離**：空間情報に着目する多チャネル音源分離では、物理モデルに基づく確率的生成モデルの逆問題を解くアプローチが学習データを必要とせず高い性能を発揮している。独立低ランク行列分析(ILRMA)や混合複素ガウスモデル(CGMM)が知られている。しかし、限られた観測データから高い性能を達成するためにこれらのモデルは複雑化の一途を辿っており、膨大な計算コストや近似誤差による性能限界といった課題がある

２．研究の目的

本研究の目的は、単チャネル音源分離を教師データのない大量の多チャネル混合音信号から学習する枠組みを構築することである。多チャネル信号から「空間的な独立性」を、大量の混合音データから「時間的な独立性」を持つ音源信号の時間周波数構造を学習する。

特に、「教師データを用いずとも個別の音源信号を高速・高精度に聞き分ける機能を計算機で実現する方法は何か」を学術的な問いとして設定し、研究を進めた。人間は音源の教師データがなくとも、日常生活のなかで個別の音源をリアルタイムに認識する能力を獲得している。また、片耳(単チャネル)でもある程度聞き分ける能力を持っている。申請者は、空間的・時間的に独立な音源信号の特徴を学習すれば、このような音源分離機能が獲得できるのではという観点から、本問題に取り組んだ。

３．研究の方法

雑踏環境下での大規模録音データを用いて音源分離を教師なし学習するアルゴリズムの確立し音環境理解システムを構築するため、以下の 2 つのサブテーマに取り組んだ。

- **教師なし深層ベイズ学習の基礎アルゴリズム確立**

音源分離を教師なし学習する基本アルゴリズムを確立し、シミュレーション混合データを用いて基本性能を確認する。ベイズ生成モデルの事後分布を深層ニューラルネットワーク(DNN)で推論する枠組みである償却変分推論に基づく学習アルゴリズムを導出する。ブラインド信号処理に用いられる混合複素ガウスモデルに基づく混合音の空間的なベイズ生成モデルを構築し、DNN で推定された分離結果(事後分布)を最適化する。

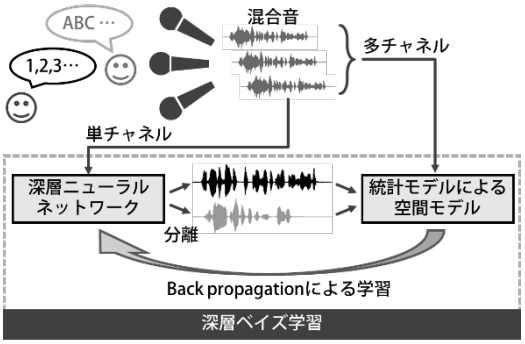

図 1 深層ベイズ学習の概要

- **雑踏環境に対応する深層ベイズ学習への拡張**

雑踏環境での音源分離に必要な「未知音源数」「拡散性雑音」を扱えるベイズモデルを構築し、深層ベイズ学習の更新式を導出する。「未知音源数」の対処は、ノンパラメトリック・ベイズによる無限混合モデルを打切り近似し、back propagation 可能な更新式を導出する。「拡散性雑音」に対しては、チャネル間の無相関性を仮定した音源項を導入し、背景雑音として推定する。

４．研究成果

本研究では、深層ベイズ学習に基づく音環境認識に関わる、主に以下の 2 つの成果が得られた。

(1) 混合複素ガウスモデルに基づく音源分離の深層ベイズ学習

2 人の音声が混合された混合音のみから、個別の音声信号を抽出する DNN を教師なし学習する枠組みを実現した。ブラインド音源分離で広く用いられている統計的な空間モデルは、時間周波数(TF)領域での瞬時混合過程を仮定しており、周波数ごとに音源インデックスが揃わないパーミュテーション問題が存在する。パーミュテーション問題の解決には、音源信号を制約(低ランク性など)する方法と、音源到来方向(DoA)を同時推論する方法が知られている。本研究では、各音源の TF マスクと DoA を潜在変数にもつ混合複素ガウスモデル(CGMM)に基づきコスト関数を導出し，TF マスクを推定する分離 DNN と DoA を推定する定位 DNN を学習する(図 2)．DoA を同時推

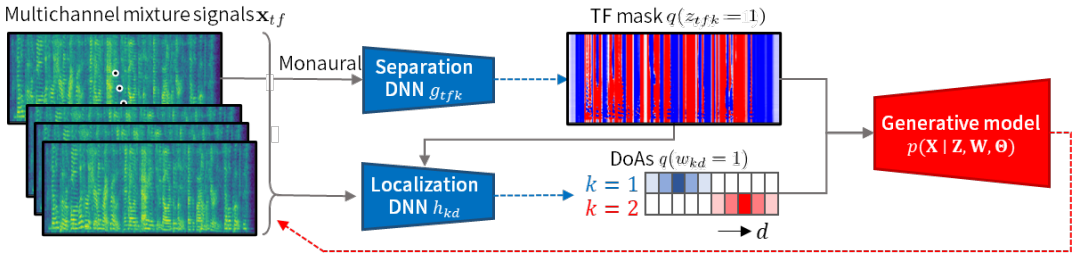

図 2 音源定位と分離を一挙に推論する深層ベイズ学習の概要

論することで，空間モデルのパーミュテーション問題を統一的な枠組みで解決できる．具体的には、推定された TF マスクと DoA が、それぞれの事後分布を表していると仮定し、この事後分布と真の事後分布とのカルバック・ライブラーダイバージェンスを最小化するように学習する。この学習は、CGMM の周辺尤度の変分下限の最大化として定式化され、このような学習は償却変分推論と呼ばれる。本手法を評価するため、2 つの音声を混合した多チャネル混合音を数値的に生成し、分離・定位 DNN の学習を行った。表 1 のように信号対歪比(SDR)が 5.3 dB 程度の性能を達成し、多チャネル混合音のみから単チャネル音源分離が学習できることを示した。さらに学習済みの分離 DNN は、単チャネル音源分離として動作するだけでなく、局所解の多い CGMM の多チャネル分離アルゴリズムに良い初期値を与えることができる。シミュレーション混合音を用いた評価により、従来の初期化法より SDR が 0.9dB 改善することを確認した。これらの内容は、査読付き国際会議 IEEE International Workshop on Machine Learning for Signal Processing において発表した。

表 1 音声分離の SDR

Method	Init.	# of mics. M train	test	SDR [dB]
EM-cGMM	g_{tfk}	4	4	**10.6 ± 4.2**
EM-cGMM	(19)–(20)	–	4	9.7 ± 5.0
AVI-cGMM	–	4	1	5.3 ± 4.5
AuxIVA+	–	–	4	9.9 ± 4.4
AuxIVA	–	–	2	5.6 ± 4.0
PIT	–	1	1	7.7 ± 4.5
DPCL	–	1	1	6.9 ± 4.7

(2) 視聴覚統合に基づく教師なし音源定位への拡張

視聴覚統合は音源定位・分離などの性能を改善できるうえ、音響シーン分析に画像情報を併用でき多面的な拡張が期待できる。本研究では、深層ベイズ学習による音源分離法を視聴覚統合へと拡張した。深層学習による教師なし視聴覚統合として、近年、画像と単チャネル音響信号のペアデータからそれらの共起関係を学習することで、画像中の音源位置を推定する DNN を教師なし(自己教師あり)学習する枠組みが注目されている。しかし、これらの多くは音源を視覚的に特定できることを前提としており、口元がはっきり見えない複数の人間が同時に喋っている映像といった、視覚的に不鮮明なデータには適用できない。本研究では、多チャネル信号に含まれる空間情報を用いて、画像情報のみでは音源を弁別できない映像データからでも頑健に動作する自己教師あり学習の枠組みを確立した。具体的には、全方位画像から音源方向を推定する DNN を、CGMM に基づく空間モデルを用いて深層ベイズ学習した。本枠組みでは、音源分離に用いる TF マスクは周辺化し、DoA のみ推論する。特に、DoA の表現を、音源分離の枠組みで用いた混合モデルによる定式化ではなく、音源信号の空間相関行列(SCM)を、事前に計測したステアリングベクトルで構成される基底 SCM の線形和で表現する因子分解モデルで定式化することで、拡散性雑音を陽に表現できるようになった。屋内で撮影された全方位画像と人間の写真を合成したシミュレーション視聴覚データを用いた実験で、提案法は教師データを用いずとも個別の人間領域を検出できることを確認した(図 4)。 さらに、特にマイク数が少ない条件で、本手法は従来の音源定位法 (MUSIC 法)の性能を上回ることを確認した。この枠組では、混合音に含まれる音源数を事前に指定する必要があった。そこで、画像から推定された音源「候補」が、実際に音を発しているかを音響信号から検証する 2 段構成にした枠組み(図 5)を開発した。本枠組みでは、ディリクレ事前分布モデルを用いて音源数を同時推定する。雑踏である科学館で収録した 16 チャネル混合音と全方位映像から、提案法が一定の性能で音環境を認識できていることを確認した。本成果は、現在、査読付き国際会議に投稿中である。

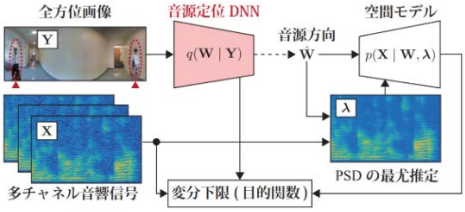

図 3 視聴覚統合の概要

図 4 音源物体(人間)の検出結果例

表 2 音源定位性能 (F 値)

マイク数	$M=2$		$M=4$		$M=6$	
音源数	$N=2$	$N=3$	$N=2$	$N=3$	$N=2$	$N=3$
MUSIC 法	–	–	0.69	0.63	**0.82**	**0.76**
提案手法	**0.62**	**0.62**	**0.78**	**0.76**	0.76	**0.76**

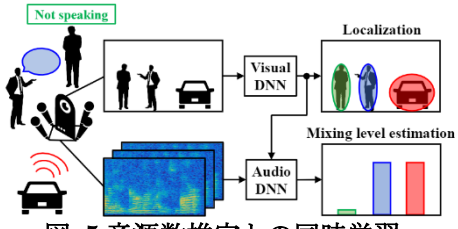

図 5 音源数推定との同時学習

これらの研究活動を評価され、視聴覚統合に関して国内学会での招待講演を 2 件行った。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　2件／うち国際学会　1件）

１．発表者名 坂東宜昭，佐々木洋子
２．発表標題 深層ベイズ学習に基づく単チャネル音源分離の教師なし学習
３．学会等名 第35回情報論的学習理論と機械学習（IBISML）研究会
４．発表年 2018年

１．発表者名 Yoshiaki Bando, Yoko Sasaki, Kazuyoshi Yoshii
２．発表標題 Deep Bayesian Unsupervised Source Separation Based on a Complex Gaussian Mixture Model
３．学会等名 IEEE International Workshop on Machine Learning for Signal Processing（国際学会）
４．発表年 2019年

１．発表者名 坂東 宜昭，佐々木 洋子，吉井 和佳
２．発表標題 混合複素ガウスモデルに基づく深層ベイズ音源分離
３．学会等名 音学シンポジウム
４．発表年 2019年

１．発表者名 升山 義紀，坂東 宜昭，大西 正輝，矢田部 浩平，及川 靖広
２．発表標題 全方位画像と多チャネル音響信号を用いた自己教師あり深層音源定位
３．学会等名 日本音響学会2020年春季研究発表会
４．発表年 2020年

１．発表者名
坂東 宜昭

２．発表標題
（招待講演）視聴覚情報の深層ベイズ学習に基づく音響シーン分析

３．学会等名
日本音響学会2020年春季研究発表会（招待講演）

４．発表年
2020年

１．発表者名
坂東 宜昭

２．発表標題
3次元・マルチモーダル音環境認識

３．学会等名
第25回 画像センシングシンポジウム（招待講演）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

研究代表者ホームページ
https://ybando.jp

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考