2Q-05

視聴覚統合に基づく音源定位と音区間検出の自己教師あり学習

升山 義紀[1,2]　坂東 宜昭[2]　佐々木 洋子[2]　大西 正輝[2]　矢田部 浩平[1]　及川 靖広[1]

[1]早稲田大学 基幹理工学研究科 表現工学専攻　[2]産業技術総合研究所

1. はじめに

ロボットが周囲の音環境を理解し行動するには，どんな音がいつどこから聞こえてくるか認識することが重要である．その際，音源物体の視覚的な情報も利用すればより多面的に音環境を理解できる [1]．会議映像など音源の種類が限定された環境では，深層ニューラルネットワーク (DNN) を用いた手法が高い性能を実現している [2]．しかし，様々な種類の音源が存在する実環境で収録された学習データに対し，各音源の位置や区間を表す教師データを作成するには膨大なコストがかかる．

この課題を解決するために，教師データの代わりに音響信号と画像の共起関係を利用して DNN を自己教師あり学習する手法が注目されている [1]．特に，多チャネル音響信号の空間情報を活用した手法は，頑健に複数の音源物体を定位できる [3]．統計的な空間モデルに基づき学習することで，画像内の物体が音を発生しているかいないかを判別でき，音源数が未知でも定位できる．

本研究では，空間モデルを用いた従来の学習法を拡張し，音源物体を定位する DNN と同時に，各物体がいつ音を発しているか推定する DNN を自己教師あり学習する方法を提案する．具体的には，従来手法では時不変であった混合比を時変に再定式化することで，音区間検出に対応する．図 1 のように，多チャネル音響信号の生成モデルの償却変分推論に基づき，音源方向推定 DNN と音区間検出 DNN を教師データなしに学習できる．

2. 多チャネル視聴覚自己教師あり学習

全方位画像 $\mathbf{Y}$ と多チャネル音響信号 $\mathbf{X}$ から各音源の方向と区間を推定するために，音源方向を表す重み $\mathbf{W}$ と音区間に対応する時変の混合比 $\boldsymbol{\pi}$ を潜在変数にもつ混合複素ガウスモデル (CGMM) [4] を定式化する．DNN は各潜在変数の事後分布を推論するように学習する．

2.1 多チャネル混合音の生成モデル

音源の方向情報を利用した CGMM は，音源数が未知でも頑健に音源定位や音源分離できるため [5]，実環境での多チャネル音響信号処理において広く活用されている．CGMM ではまず，最大 $K_{\max}$ 個の音源信号を M チャネルのマイクアレイで観測した混合音 $\mathbf{x}_{tf} \in \mathbb{C}^M$ が以下の混合複素ガウス分布に従うと仮定する．

$$\mathbf{x}_{tf} \sim \sum_{k=1}^{K_{\max}} \pi_{tk} \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \lambda_{tfk} \mathbf{H}_{fk}\right) \tag{1}$$

ただし，$\pi_{tk} \in [0,1]$ は $\sum_k \pi_{tk} = 1$ を満たし，$\lambda_{tfk} \in \mathbb{R}_+$ は時間周波数ビンごとのパワースペクトル密度，$\mathbf{H}_{fk} \in \mathbb{S}_+^{M \times M}$ は空間相関行列，t と f は時間フレームと周波数ビンのインデックスを表す．

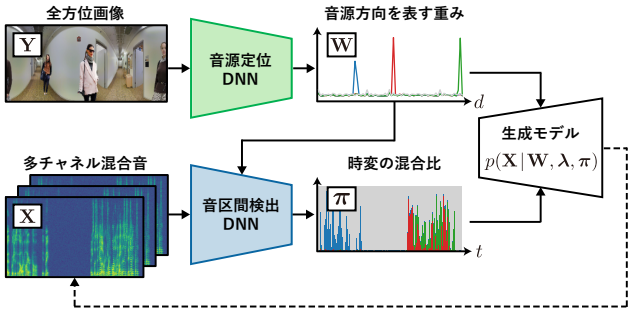


図 1: 音源定位 DNN と音区間検出 DNN の自己教師あり学習法の概要．

混合比を時間フレームごとにスパースにし，各音源の区間を推定するために，$\boldsymbol{\pi}$ にディリクレ事前分布をおく．

$$[\pi_{t1}, \ldots, \pi_{tK_{\max}}]^{\mathsf{T}} \sim \mathrm{Dir}\left(\alpha_0, \ldots, \alpha_0\right) \tag{2}$$

ただし，$\alpha_0 \in \mathbb{R}_+$ はハイパーパラメータである．各音源の方向と空間相関行列を関係づけて音源方向を推定するために，空間相関行列 $\mathbf{H}_{fk}$ を事前計算した音源の方向候補 d ごとの基底 $\mathbf{G}_{fd} \in \mathbb{S}_+^{M \times M}$ の重み付き和で表す．

$$\mathbf{H}_{fk} = \sum_{d=1}^{D} w_{kd} \mathbf{G}_{fd} \tag{3}$$

ここで，$w_{kd} \in \mathbb{R}_+$ が大きい値をとった場合，音源 k は方向 d に存在すると考える．重み w_{kd} を推論する際の正則化として対数正規事前分布 $w_{kd} \sim \mathcal{LN}\left(0, \sigma_0^2\right)$ を置く．

2.2 音源定位 DNN と音区間検出 DNN の学習

提案手法では，図 1 のように $\mathbf{W}$ の事後分布を推論する音源定位 DNN と，$\boldsymbol{\pi}$ の事後分布を推論する音区間検出 DNN を学習する．具体的には，各 DNN の出力を以下の近似事後分布のパラメータとみなす．

$$q(\mathbf{W} \mid \mathbf{Y}) = \prod_k \prod_d \mathcal{LN}\left(\hat{w}_{kd}, \sigma_k^2\right) \tag{4}$$

$$q(\boldsymbol{\pi} \mid \mathbf{W}, \mathbf{X}) = \prod_{t=1}^{T} \mathrm{Dir}\left(\beta\hat{\pi}_{t1}, \ldots, \beta\hat{\pi}_{tK_{\max}}\right) \tag{5}$$

ただし，$\hat{w}_{kd} \in \mathbb{R}$，$\hat{\pi}_{tk} \in [0,1]$ はそれぞれ音源定位 DNN と音区間検出 DNN の出力であり，$\sum_k \hat{\pi}_{tk} = 1$ を満たす．σ_k と β は入力に非依存なパラメータである．

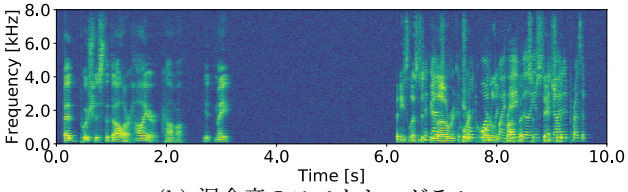
DNN は近似事後分布 $q(\mathbf{W}, \boldsymbol{\pi} \mid \mathbf{X}, \mathbf{Y})$ と真の事後分布 $p(\mathbf{W}, \boldsymbol{\pi} \mid \mathbf{X}, \boldsymbol{\lambda})$ 間のカルバック・ライブラーダイバージェンスを最小化するように学習する．この最小化は，以下の変分下限 $\mathcal{L}$ の最大化に対応する．

$$\mathcal{L} = \mathbb{E}_q \left[\log p(\mathbf{X} \mid \mathbf{W}, \boldsymbol{\lambda}, \boldsymbol{\pi})\right] - \mathbb{KL}\left[q(\mathbf{W}, \boldsymbol{\pi}) \,|\, p(\mathbf{W}, \boldsymbol{\pi})\right] \tag{6}$$

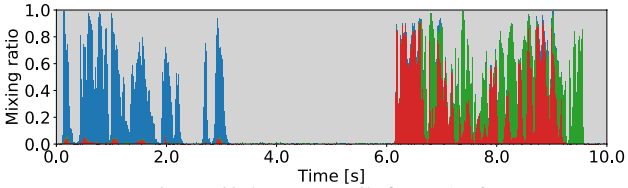
式 (6) は解析的に計算困難なため，[3] と同様にモンテカルロ近似して最大化する．

Audio-visual self-supervised learning for sound source localization and activity detection: *Y. Masuyama, Y. Bando, Y. Sasaki, M. Onishi, K. Yatabe and Y. Oikawa*

(a) 全方位画像と音源定位 DNN の中間層の出力



(b) 混合音のスペクトログラム



(c) 音区間検出 DNN が推定した混合比

図 2: DNN への入出力の例．青，赤，緑，灰の各色は CGMM の音源クラス k に対応 (灰色は雑音に相当)．

3. 評価実験

人物画像を複数含む全方位画像と多チャネル混合音のシミュレーションデータで提案手法を評価した．

3.1 データセットと実験条件

全方位画像と多チャネル混合音のペアを学習用 20000 組，検証用 1000 組，評価用 1000 組生成した．ただし，生成に利用した各音源の方向と区間の情報は学習に用いない．音源信号は，WSJ0 コーパスの読み上げ音声から 3 秒分切り出し，10 秒間のランダムな区間に配置し生成した．多チャネル混合音は，音源信号にインパルス応答を畳み込んで生成した．インパルス応答は鏡像法を用いて生成し，部屋の大きさは 5 m × 5 m × 3 m とし，残響時間は 0.2 s から 0.5 s の範囲で変動させた．半径 10 cm の $M \in \{2, 4, 6\}$ チャネルの円形マイクアレイを用いた．音源数 K は混合音ごとに $\{2, 3\}$ からランダムに選択し，マイクアレイとの距離 0.5 m から 2.0 m の範囲内でランダムに配置した．全方位画像は 2D-3D-S データセットの室内全方位画像と Clothing Co-Parsing データセットの人物画像を用い，図 2-(a) のように生成した．

音源定位 DNN には [3] と同様の DNN を用い，音区間検出 DNN には [3] で時不変の混合比を推定していた DNN の時間方向のプーリングを除去した DNN を用いた．これら DNN の学習には Sharpness-aware minimization を用い，学習率は 0.0004 とした．全方位画像 $\mathbf{Y}$ のサイズは 88×288 で，短時間フーリエ変換には 1024 点のハン窓を用い，シフト幅は 256 点とした．提案手法における CGMM のクラス数 $K_{\max}$ は実際の音源数 K よりも多い 4 とした．また，水平方向 5° 間隔の平面波を仮定し基底 $\mathbf{G}_{fd}$ を計算した．

音源定位性能は SRP-PHAT，MUSIC，IWMM-CGMM と比較し，音区間検出は IWMM-CGMM と比較した．提案手法と従来手法のハイパーパラメータは検証データを用いて最適化した．

表 1: 音源定位性能 (F 値)

マイク数	$M=2$		$M=4$		$M=6$	
音源数	$K=2$	$K=3$	$K=2$	$K=3$	$K=2$	$K=3$
SRP-PHAT	0.58	0.52	0.77	0.73	0.88	0.82
MUSIC	0.48	0.44	0.82	0.80	**0.95**	**0.91**
IWMM-CGMM	0.39	0.41	**0.90**	**0.87**	0.93	**0.91**
提案手法	**0.83**	**0.76**	0.88	0.82	0.82	0.80

表 2: 音区間検出性能 (F 値)

マイク数	$M=2$		$M=4$		$M=6$	
音源数	$K=2$	$K=3$	$K=2$	$K=3$	$K=2$	$K=3$
IWMM-CGMM	0.86	**0.78**	0.85	0.73	0.86	0.74
提案手法	**0.88**	0.77	**0.91**	**0.84**	**0.92**	**0.85**

3.2 実験結果

音源定位性能を表 1 に示す．多チャネル音響信号のみを用いる従来手法は $M = 6$ では高い性能を実現したが，マイク数の減少につれて大きく性能が劣化している．全方位画像を用いる提案手法はマイク数の変動に対し安定して定位できており，$M = 2$ では従来手法の性能を上回っている．また図 2–(a) に示すように，音源定位 DNN は複数の人物画像を区別して反応できている．音区間検出性能を表 2 に示す．マイク数 2 で 3 音源の場合を除き，提案手法は IWMM-CGMM を上回った．推定された混合比 $\hat{\boldsymbol{\pi}}$ の例を図 2–(c) に示す．人物画像に反応した音源クラス (青，赤，緑) は音声のある区間でのみ混合比が大きな値を取っている．これらの結果は，提案手法によって画像内の音源の方位と区間を推定する DNN を教師データを用いずに学習でき，マイク数が少ない場合でも安定して動作できることを示している．

4. むすび

多チャネル混合音と全方位画像を用い，音源物体の定位および音区間検出を行う DNN の自己教師あり学習法を提案した．今後は音響イベント検出のために，音源の種類を判別する DNN を同時に学習する方法を検討する．

謝辞 本研究の一部は国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託事業によって行われた．

参考文献

[1] A. Owens *et al.* Audio-visual scene analysis with self-supervised multisensory features. In *Proc. of ECCV*, 631–648, 2018.

[2] T. Yoshioka *et al.* Advances in online audio-visual meeting transcription. In *Proc. IEEE ASRU*, 276–283, 2019.

[3] Y. Masuyama *et al.* Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling. In *Proc. of IEEE/RSJ IROS*, 4848–4854, 2020.

[4] T. Higuchi *et al.* Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In *Proc. of IEEE ICASSP*, 5210–5214, 2016.

[5] T. Otsuka *et al.* Bayesian nonparametrics for microphone array processing. *IEEE/ACM Trans. on ASLP*, 22(2):493–504, 2014.