空間フィルタの自動推定による音響シーン識別の検討

ACOUSTIC SCENE CLASSIFICATION BASED ON ESTIMATION OF SPATIAL FILTER

大野泰已[1]　山田武志[1]　牧野昭二[1]
Taiki Ono　Takeshi Yamada　Shoji Makino

筑波大学大学院 システム情報工学研究科[1]
Graduate School of Systems and Information Engineering, University of Tsukuba

1　はじめに

音響シーン識別とは音響信号が収録された場所や状況を識別することであり，マルチメディア検索や推薦などの応用が期待されている．従来の音響シーン識別はモノラル信号を入力としていた．それに対して，最近は環境音認識の国際コンペティションであるDCASE(Detection and Classification of Acoustic Scenes and Events)のようにステレオ入力を前提とする手法が増えてきている．これは，空間情報を活用することによって識別精度の向上が期待できるからである．例えば，Hanらは Middle-Side処理（左右チャネルの加算信号と減算信号の生成），Tanabeらは Duongの音源分離手法を前処理として適用した[1][2]．しかし，音響シーン識別においては，どの音を強調・抑圧すれば良いのか（どのような空間フィルタを形成すれば良いのか）が自明ではないという問題がある．

そこで本稿では，個々の音響シーンの識別に適した空間フィルタを自動的に推定して識別を行う手法を提案する．提案手法は，空間フィルタモデルと音響シーン識別器からなり，これらを同じ損失関数のもとでEnd to Endに学習する．

2　提案手法

提案手法の概要を図1に示す．まず，マルチチャネル（図1ではステレオ）入力信号に対してSTFT（short-term Fourier transform）を適用し，各チャネルの周波数スペクトログラムを得る．次に，これを空間フィルタモデルに入力し，空間フィルタ出力である周波数スペクトログラムを得る．そして，音響特徴量を抽出して識別を行う．空間フィルタモデルと音響シーン識別器には様々なモデルを用いることが考えられるが，本稿では共にCNN（convolutional neural network）を適用する．これらを同じ損失関数のもとでEnd to Endに学習することにより，識別に適した空間フィルタが自動的に適用されるようになる．

3　実験

DCASE 2019 Task 1[3]のデータセット（ステレオ入力）を用いて識別実験を行った．提案手法における空間フィルタモデルの入力として振幅スペクトログラムを用いる場合と複素スペクトログラムを用いる場合を比較する．ここで，後者の場合はcomplex-valued CNN[4]を適用した．

実験結果を表1に示す．各手法の音響特徴量と識別器は同一である．「L or R」（モノラルチャネル）と「L+R」

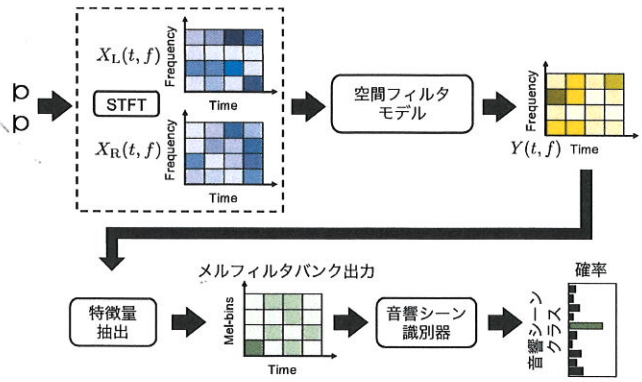

図1　提案手法の概要

表1　実験結果

手法	Accuracy
L or R（モノラルチャネル）	0.59
L+R（左右チャネルの加算信号）	0.60
提案手法（振幅スペクトログラム）	0.65
提案手法（複素スペクトログラム）	0.67

（左右チャネルの加算信号であり，正面方向を強調することに相当）を比較するとほとんど差がないことが分かる．一方，提案手法により最も高い識別精度が得られること，及び複素スペクトログラムの方が優れていることが確認できる．

4　おわりに

本稿では，識別に適した空間フィルタを自動的に推定して識別する手法を提案し，実験によりその有効性を確認した．今後は推定された空間フィルタの特性を解析する．

謝辞　本研究はJSPS科研費19H04131の助成を受けた．

参考文献

[1] Y. Han, and J. Park, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," DCASE Challenge Technical report, 2017.

[2] R. Tanabe, T. Endo, Y. Nikaido, T. Ichige, P. Nguyen, Y. Kawaguchi, and K. Hamada, "Multi-channel acoustic scene classification by blind dereverberation, data augmentation, and model ensembling," DCASE Challenge Technical report, 2018.

[3] http://dcase.community/challenge2019/task-acoustic-scene-classification.

[4] C. Trabelsi, O. Bilaniuk, D. Serdyuk, S. Subramanian, J. FelipeSantos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. Pal, "Deep complex networks," arXiv preprint, arXiv:1705.09792, 2017.