中間報告

空間情報を用いた音響シーン識別に関する研究

岡野 稜

情報科学類4年 マルチメディア研究室

指導教員　山田 武志 准教授

研究背景

人の動きや周囲の状況を環境音を使って自動認識しようとする取り組みが活発化

例）高齢者見守りシステム
高齢者の生活音の中の異常を検知し、対応の迅速化を図る

例）動画への自動タグ付け
ライフログとして動画に自動でタグ付け

環境音認識のコンペティションが行われている → DCASE

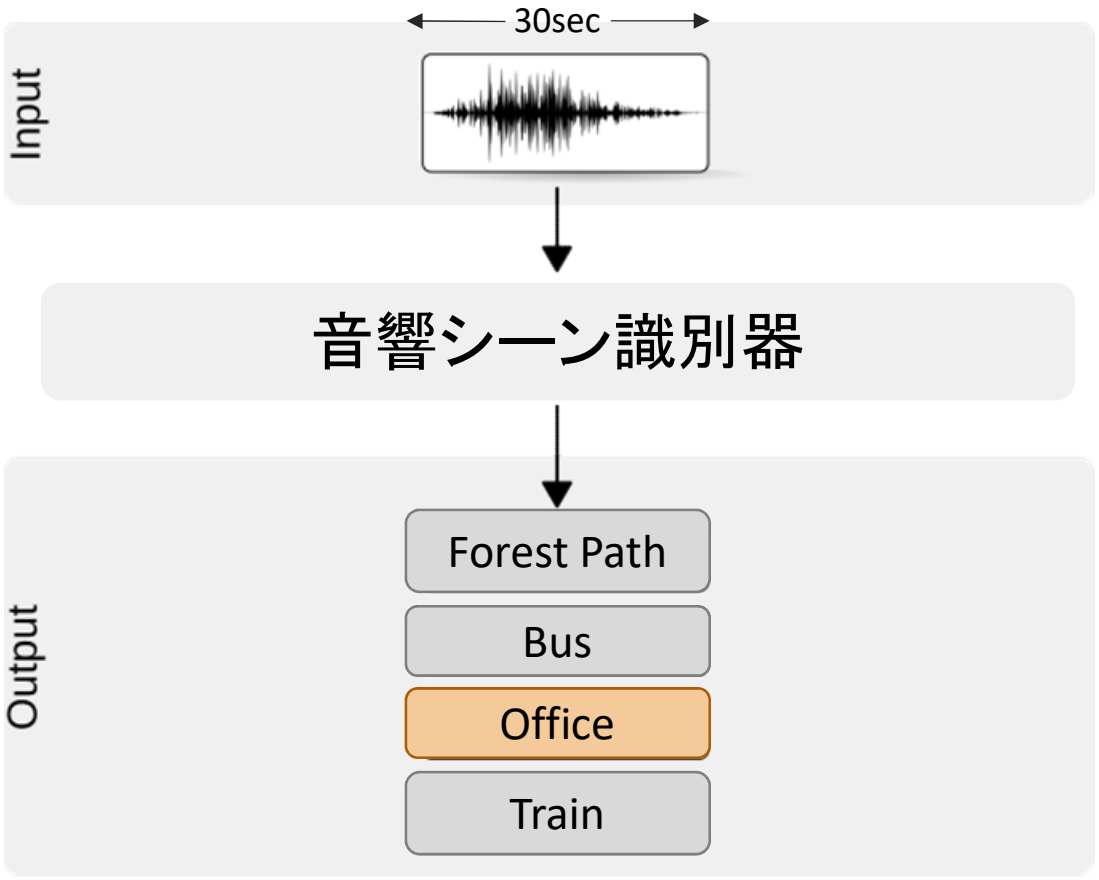
環境音認識

環境音認識は大きく2分野に分けられている

- 音響シーン識別
- 音響イベント検出

本研究では音響シーン識別に取り組む

音響シーン識別



入力音響信号からどのシーンなのかを認識する

従来研究

DCASE2016
古くから音響シーン識別で用いられている手法
→GMM-HMM[DCASE2016 baseline]

時間方向に着目した手法
→DNN-GMM + フレーム連結手法[高橋2016]

使用データ

実験に用いるデータはこちらを使用している

音響シーン識別
録音された音の環境を分類 DCASE2016[1]

	DCASE2016
分類シーン数	15シーン
データ数	1170 (15シーン×78個)
各セグメントの長さ	30sec

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/task-acoustic-scene-classification

研究目的

音響シーン識別において識別性能を向上させる

- 方針

  入力信号に対して空間的前処理を施す

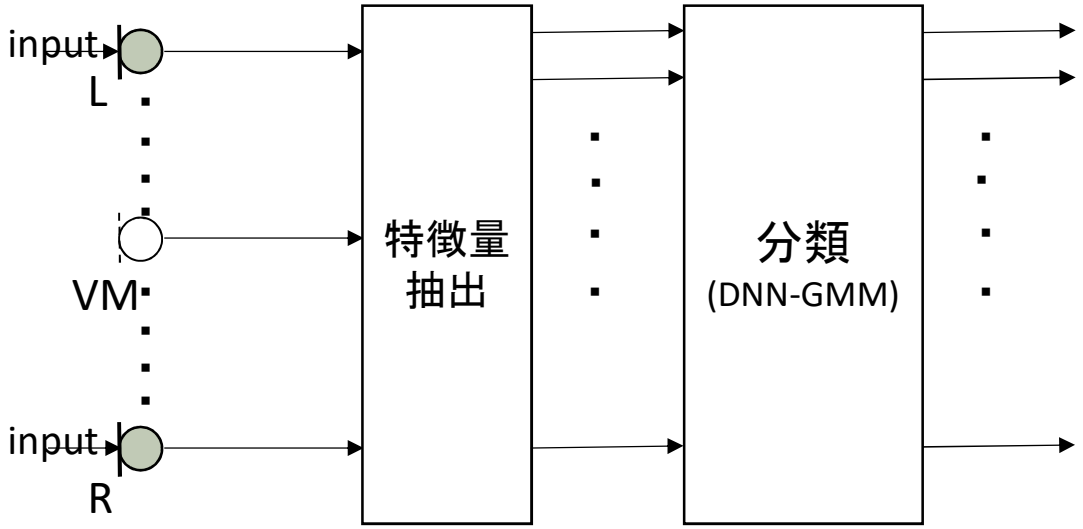
  先行研究として空間的前処理を施したGMM-HMMでは精度の向上を確認[湯原2016]

  DNNでは向上を観察できなかったため、本研究では新たにvirtual microphoneを用い空間的前処理を行う

提案手法

- Virtual Microphoneを用い、入力チャネル数を増やす

DNNの中でアレー信号処理を行っている可能性があるため
マイクの数を仮想的に増やすことで
識別精度の向上、学習を十分に行わない環境での精度の安定などを期待



Virtual Microphone[片平2014]

$$v = A_{V\beta}\exp(j\varphi_V)$$

$$A_{V\beta} = \begin{cases} \exp\left((1-\alpha)\log x_1 + \alpha\log x_2\right) & (\beta = 1) \\ \left((1-\alpha)A_1^{\beta-1} + \alpha A_2^{\beta-1}\right)^{\frac{1}{\beta-1}} & (otherwise) \end{cases} \quad ※_1$$

$$\varphi_V = (1-\alpha)\varphi_1 + \alpha\varphi_2 \quad ※_2$$

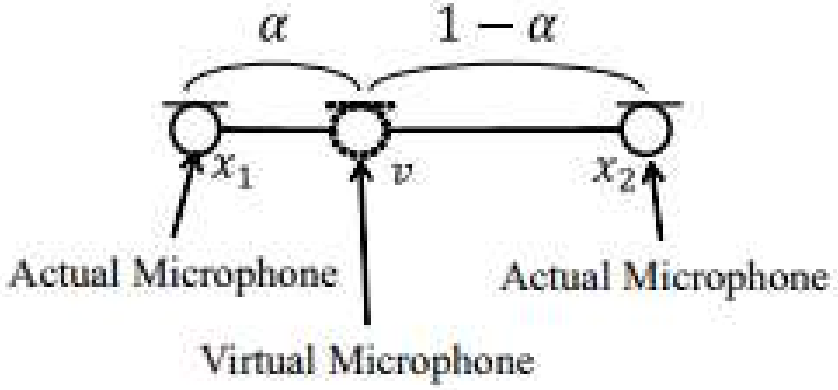
※$_1$βダイバージェンスを導入した非線形補間

※$_2$平面波の位相はvmの位置αに対して線形に変化するため線形補間

上記の式によって補間された信号が

virtual microphone信号として

入力に加わる



実験

2chのデータセットをDMM-GMMを用いて認識したものとヴァーチャルマイクを用いて3chにしたデータセットを同様に認識したものそれぞれの推定精度を比較

実験条件

	従来手法	提案手法
教師データ	DCASE2016 Development Dataset	
評価データ	DCASE2016 Evaluation Dataset	
分類器	DNN-GMM	
特徴量	MFCC+Δ+ΔΔ(それぞれ20次元,フレーム連結無し)	
ノード数	128,256,512,1024,2048	
隠れ層数	2,3,4,5,6	
シード値	10通りで平均を取る	
入力信号	Real2ch(L,R)	Real2ch(L,R)+VM

実験結果：平均識別精度(%)

従来手法2ch(L,R)

次元＼層	2	3	4	5	6
128	84.14	84.15	84.19	83.98	83.79
256	85.22	84.64	83.86	84.24	83.66
512	85.82	84.33	84.10	84.01	83.96
1024	84.87	84.10	84.21	83.77	83.91
2048	84.82	84.26	83.73	83.53	83.09

どちらも83%~85%の値を取っている

これらの表の値は10通りの
シード値での結果の平均値である

提案手法3ch(L,R,VM)

次元＼層	2	3	4	5	6
128	84.78	84.36	83.98	84.26	83.96
256	85.11	84.24	84.17	83.75	84.03
512	85.33	84.68	84.61	84.40	84.38
1024	85.33	84.75	84.70	84.10	84.17
2048	85.52	84.5	84.12	84.19	83.94

参考値

DCASE2016baseline	77.2%
DCASE2016最高値	89.7%
フレーム連結法[2016高橋]	85.6%

実験結果：提案手法と従来手法との差

次元＼層	2	3	4	5	6
128	0.630	0.210	-0.209	0.278	0.163
256	-0.104	-0.396	0.303	-0.485	0.373
512	-0.49	0.35	0.511	0.396	0.42
1024	0.465	**0.653**	0.49	0.327	0.254
2048	**0.692**	0.234	0.396	**0.657**	**0.844**

正：提案手法>従来手法
負：従来手法>提案手法

多くの学習データが必要な
大きな次元において精度が良い傾向がある

今後の計画

同様の調査を主音源・副音源分離[湯原2016] 手法でも行い、性能に差が生じるか確認

エポック数を減らした学習、学習データが少ない学習などを行い性能に影響があるかを比較

年間スケジュール

8月		**12月**	
9月		**1月**	卒業論文提出(1月26日)
10月		**2月**	卒業研究発表(2月16日)
11月		**3月**	