DNN-GMMを用いた音響イベント検出に関する研究

ソフトウェアサイエンス主専攻　201111407　湯原祥甫

指導教員　山田武志（システム情報系情報工学域）

牧野昭二（システム情報系情報工学域）

提出日2016年10月28日

1 研究背景・目的

私たちの日常は，大量の環境音によって囲まれている．環境音に含まれている情報を認識することを環境音認識といい，今日様々なシステムに対して環境音認識技術の利用が検討されている．環境音認識には大きく分けて2つの種類がある．足音や車の通過音などといった短時間に起こるイベント音を認識する音響イベント検出と，繁華街やオフィスなどといった環境音の録音された状況を認識する音響シーン認識である．

環境音認識技術の利用が検討されているシステムには様々なものがあり，ライフログや高齢者の見守りシステム，ハウスセキュリティシステム等，日常生活の様々な面に幅広く適用することができる．環境音認識を活用することのメリットとしては，映像では対応することのできない死角の情報を得ることができるというものがある．しかし，この技術を用いたシステムは未だ少なく，理由として，環境音認識が音声を含む多様な音を扱っているために性能を十分発揮することができないことが挙げられる．

こういった問題に対応するため，2016年9月には IEEE (The Institute of Electrical and Electronics Engineers) の AASP (Audio and Acoustic Signal Processing Technical Committee) が主催の，環境音認識技術を向上させることを目的とした DCASE (Detection and Classification of Acoustic Scenes and Event) 2016[1] というワークショップが開催された．ここでは環境音認識に関する4つのタスクが提示され，世界中の研究チームが与えられた共通のタスクに対する精度を競い合った．しかし，このワークショップにおけるタスクの1つである「Sound event detection in real life audio (実録音による音響イベントの分類)」における，最も高い精度を出したチームの研究[2]でも47.8%の検出率であり，さらなる精度の改善が求められる結果となった．

よって，本研究は環境音認識の精度の向上を目的とし，特に音響イベントの高精度な検出に焦点を当てる．この問題を解決するにあたり，音響シーン認識で高い精度を達成した，DNN (Deep Neural Network) と GMM (Gaussian Mixture Model) を組み合わせた手法[3]である DMM-GMM を取り入れ，精度の向上を目指す。

2 DCASE2016

本研究を進める上で， DCASE2016 のデータセットを用いて評価を行う．提示された4つのタスクのうち，

・Sound event detection in synthetic audio (合成音による音響イベントの分類)」
・Sound event detection in real life audio (実録音による音響イベントの分類)」

の2つが本研究において取り組むタスクである．

前者は，オフィスでの音響イベント検出に焦点が当てられている．咳・ドアの開閉・引き出しの開閉といったような11種類のイベントを，2分程度の人為的に作られたイベント音から識別する．このワークショップにおいて最も高い精度を出したチームは，特徴量抽出に VQT (Variable Q Transform)，識別器に NMF-MLD (Non-negative Matrix Factorization - Mixture of Local Dictionaries) を用い，80.2%の認識率を出している[4]．

後者は，屋内・屋外それぞれの環境において実際に収録された音響イベント検出の課題であり，前者とは評価データが実録音か，人為的に作られた音であるかの違いがある．最も高い精度であったチームの手法は，特徴量抽出が mel-energy，識別器が RNN (Recurrent Neural Network) であり，47.8%の認識率であった[2]．

3 提案手法

近年，ニューラルネットワークを多層構造にした DNN を識別器に用いた手法が画像認識や音声認識の分野で多数発表されており，従来よりも高い認識精度を達成している．同様に，本研究室における DMM-GMM を音響シーン認識に適用した研究[3]が，DCASE2016のタスクにおいても高い成果を発揮した．DMM-GMM は，従来 GMM を用いて確率分布を表現していた部分を DNN で置き換えたものである．

本研究では，DCASE2016 のタスクに対して DNN-GMM を適用し，従来手法と比べて精度がどれだけ向上するかを確認する．

4 進捗状況と今後の課題

現在は，DNN-GMM のシステム作成するにあたりKaldi[5] を用いて実装を行っている．Kaldi は C++ で書かれた音声認識用のツールキットであり，音声認識システム全体を作れるだけの豊富なスクリプトが用意されていることが特徴である．また，DNN 作成用のスクリプトも用意されているので，本研究においては最適であると言える．また，並行してDCASE2016で発表された既存研究の調査を進めている．

今後の予定としては，DNN-GMM システムの実装と実験，パラメータのチューニングをした後，DCASE2016 の既存方法と作成したシステムでどれだけ精度の差が出るかを確認し，システムの改善をすることで性能の向上を図る．

参考文献

[1]DCASE2016 Web Site, http://www.cs.tut.fi/sgn/arg/dcase2016/.

[2]Sharath Adavanne et al., "Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features", Proc.DCASE2016, Sept. 2016 (Web).

[3]Gen Takahashi et al., "Acoustic Scene Classification Using Deep Neural Network and Frame-Concatenated Acoustic Feature", Proc. DCASE2016, Sept. 2016 (Web).

[4]Tatsuya Komatsu et al., "Acoustic Event Detection Method Using Semi-Supervised Non-Negative Matrix Factorization with a Mixture of Local Dictionaries", Proc. DCASE2016, Sept. 2016 (Web).

[5]Kaldi Web Site, http://kaldi-asr.org/.