音響イベント検出における空間情報の活用に関する研究

ソフトウェアサイエンス主専攻　201411409 溝口和輝

指導教員　山田武志（システム情報系）

牧野昭二（システム情報系）

提出日2017年10月30日

1. 研究背景

私たちの日常は大量かつ多様な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取り出すことを環境音認識といい、高齢者の見守りシステムや動画の自動タグ付けなどの様々なシステムに対して適用が検討されている。環境音認識は大きく分けて 2 つの分野に分けられる。足音や車の通過音などの短時間に起こる個々の音響イベントを検出する音響イベント検出と、繁華街やオフィスなどの環境音の録音された状況を識別する音響シーン識別である。しかし、複数の音が同時に存在するような現実の環境下で、これらを確実に実行するのは依然として困難である。

この問題に対処するため、2017 年 11 月に IEEE の Audio and Acoustic Signal Processing Technical Committee（AASP）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）2017[1]というワークショップが昨年に引き続き開催される。このワークショップでは DCASE2017 Challenge として音響イベント検出や音響シーン認識を含む 4 つのタスクが設定されており、参加者は共通のデータセットを用いることで自身の開発した手法の性能を比較評価出来るようになっている。

2. 研究目的

本研究では、特に音響イベント検出の性能を改善することを目的とする。この目的を達成するにあたり、ステレオ録音から得られる空間情報を効果的に活用する手法について検討する。昨年開催された DCASE2016 において、入力としてステレオ録音を用いた手法が総合的に最も良い精度を出している[2]ものの、空間情報を十分に活用しているとは言い難い。また DCASE2017 において提出された手法のうち、ステレオ録音をモノラル化して利用したものが依然として全体の 8 割を占めている。これらのことから空間情報をより効果的に活用する方法論を見出すことが急務である。

3. 研究方法

本研究では DCASE2017 のデータセットを用いた比較評価を行う。本研究で取り組むのは、設定された 4 つのタスクのうち「Sound event detection in real life audio（実生活における音響イベント検出）」である。図 1 に示すように、このタスクではステレオ録音データを入力とし各音響イベントの開始時間と終了時間、及びその音響イベントの名称を表すラベルを出力することが求められる。録音は実際の道路沿いにて行われ、音響イベントのラベルはブレー

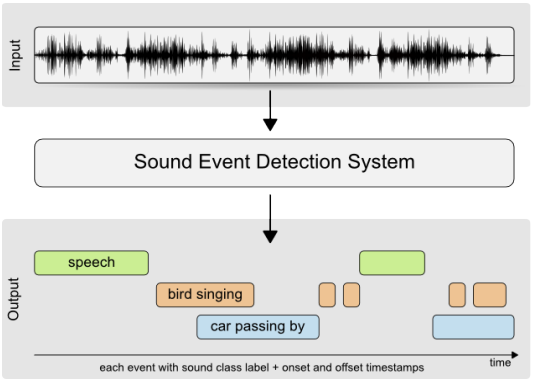

図 1 音響イベント検出の概略

キ音、車、子供、大型車両、歩行音、会話音の 6 種類がある。

DCASE2017 ではベースラインとなるシステムが公開されており、参加者はこのベースラインシステムを改良するか、あるいは新しいシステムを独自に開発することによりベースラインシステムを上回る検出精度を達成することを目指す。DCASE2017 のベースラインシステムはまず左右チャネルの信号を平均してモノラル化し、音響特徴量として対数メルフィルタバンク出力を抽出する。そして全結合型ニューラルネットワークを用いて短時間フレーム毎に音響イベントを検出する[3]。このベースラインシステムの検出精度は 42.8%である。また、本タスクにおいて最も高い精度を達成した手法は、音響特徴量として Scatteringtransform と Clustering を、識別器に Neuroevolution を用いた手法であり、検出精度は 44.9%である[4]。

音響シーン識別や音響イベント検出においては、対数メルフィルタバンク出力やメルケプストラム係数などの信号の位相情報を含まない音響特徴量を用いるのが一般的である。従って、ニューラルネットの内部では位相情報を活用した処理、特にステレオ録音から得られる空間的処理を行うことは原理的にできないことになる。一方、DCASE2016 に提出された手法の中には入力に左右チャネルの音の到来時間差を用いたものがあるが、到来時間差を用いなかった場合と比較して検出精度が 9.9%下がってしまうという結果が報告された[2]。

これらのことから本研究では、DCASE2017 のベースラインシステムにフロントエンドを追加して空間情報を抽出し、それを特徴量としてニューラルネットワークの入力とする手法を提案する。具体的にはどのような空間的前処理を行い、どのような空間情報を抽出するのが適切なのかを比較検討する。

4. 進捗状況

現在は DCASE2017 ベースラインシステムに各種フロントエンドを追加する実装を行っている。今後は、現在の実装が終わり次第各種フロントエンドを用いた場合の検出精度を比較検討する予定である。

5. 参考文献

[1] DCASE2017WebSite, http://www.cs.tut.fi/sgn/arg/dcase2017/
[2] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertila, Toni Heittola, Tuomas Virtanen,”Sound Event Detection In Multichannel Audio Using Spatial And Harmonic Features,” DCASE2016, 2016
[3] https://github.com/TUT-ARG/DCASE2017-baseline-system
[4] Christian Kroos and Mark D. Plumbley, ”Neuroevolution For Sound Event Detection In Real Life Audio: A Pilot Study,” DCASE2017,2017