Vol.2019-MUS-123 No.36
Vol.2019-SLP-127 No.36
2019/6/22

変調フィルタによる特徴量の抽出と機械学習を用いた音響シーン分類

小林 礼奈[1]　小澤 賢司[1,a)]　伊勢 友彦[2]

概要：音情報に基づいてシーンを分類できれば，自動車乗車中にその場に応じた音楽の自動選曲ができるなどの有用性がある。本研究では，音響シーン分類システムの構築を目指して，国際大会 DCASE 2018 の課題に取り組むこととした。そのために，大会により提供されたベースラインシステムのうち，音響特徴量抽出部を拡充した。具体的には，ベースラインシステムにおける対数メルバンドエネルギーの時系列を，変調フィルタバンクを通じて分析することで，音波形の振幅包絡における時間変動を特徴量として抽出した。その特徴量を，2 つの畳み込み層と 1 つの全結合層からなるニューラルネットワークの入力として 10 種のシーンの分類を試みたところ，ベースラインシステムに比べて 3.9%の性能向上を得た。

Acoustical Scene Classification Using Feature Extraction with Modulation Filters and Machine Learning

1. はじめに

著者らは，自動車運転席からの景観に応じて，音楽を自動選曲するシステムを構築してきた [1], [2]．そのシステムではビデオカメラで撮影した映像を用いて景観を分析していたが，より小規模なシステムとするためにはマイクロホンで収音した音によりシーンを自動分類できれば好都合である。

このような音響シーン分類には様々な需要があり，2013 年から国際大会 DCASE (Detection and Classification of Acoustic Scenes and Events) が開催されている [3], [4]。本研究では，音響シーン分類システムを構築するにあたり，DCASE 2018 の課題に取り組むこととした。

本研究では，音の特徴量を検討するにあたり，変調フィルタバンク [5], [6] からの出力に着目した。変調フィルタバンクとは，聴覚系内にあり，音の波形における振幅包絡の時間変動を分析する機構と考えられている。振幅包絡線のうち特定の周波数成分を通過させるフィルタ（変調フィルタ）が複数あり，それらの出力を統合することで音の特徴を捉えることができる。例えば，種々の環境音のテクスチャ（音色）が，それらの出力についての要約統計量で表現できる [7], [8]。そこで，このフィルタバンク出力を特徴量とすることが，シーン分類にも有効であると考えた。

[1] 山梨大学工学部コンピュータ理工学科
University of Yamanashi, Kofu, Yamanashi 400–8511, Japan
[2] アルプスアルパイン（株）
Alps Alpine, Co. Ltd.
[a)] ozawa@yamanashi.ac.jp

2. シーン分類の課題とシステム構成

2.1 分類課題

本研究では，DCASE 2018 の Challenge Task 1: Acoustic scene classification を対象とした。この課題は，空港・公園などを含む 10 種の音響シーンを分類するもので，学習に使用できるデータセットごとに 3 つのサブタスクに分かれている。

本研究ではサブタスク A に取り組むこととした。このサブタスクで用いられるデータセットは，音響シーンごとに 864 個，計 8640 個である。個々の音データは 10 s で，いずれもバイノーラル録音されたデータである。

2.2 自動分類システムの構成

2.2.1 システムの概要

当該課題については，大会の主催者からベースラインシステムが提供されている。そのシステムは，音データから特徴量を取り出し，正規化したものをニューラルネットワーク (NN: Neural network) へ入力して学習することで

Vol.2019-MUS-123 No.36
Vol.2019-SLP-127 No.36
2019/6/22

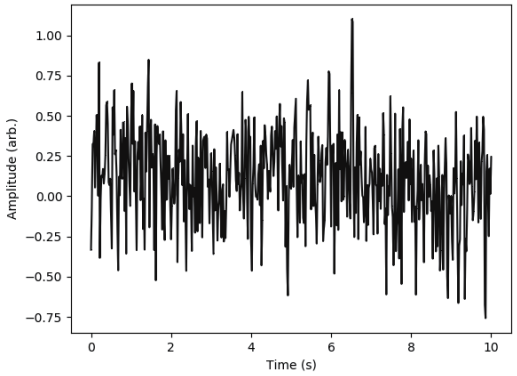

図 1 最も低い周波数帯域における対数メルバンドエネルギーの時系列データ

分類を行うものである。本稿では，そのうちの特徴量抽出方法を拡充し，システムの性能向上を目指す。

ベースラインシステムでは，対数メルバンドエネルギーを特徴量として抽出し，NN へ入力して学習を行っている。以下では，このベースラインシステムについて簡単に述べる。

2.2.2 対数メルバンドエネルギー

まず 10 s の音データを，40 ms ごとに STFT (Short-time Fourier transform) を用いて分析することで，500 点からなる時系列データを得る。そのデータをメルフィルタバンクへの入力とし，出力の対数をとることで対数メルバンドエネルギーを得る。

ベースラインシステムでは，0〜24000 Hz の範囲で 40 個のフィルタをもつメルフィルタバンクが使用されているため，ひとつの音データから時系列 40 個の特徴量を得る。例として，空港の音データから対数メルバンドエネルギーを抽出した場合に，最も低い周波数帯域から得られる振幅値の時系列データを図 1 に示す。なお，本研究では，メルフィルタバンクは，Python のパッケージである LibROSA に含まれる librosa.filters.mel を使用して構築した [9]。

2.2.3 ニューラルネットワーク

ベースラインの NN は，2 つの畳み込み層 (CNN: Convolutional NN) と 1 つの全結合層から構成されている。入力データのサイズは，メルフィルタバンクからの出力が 40 個で，それぞれが 500 個の時系列データを保持しているため，(40, 500) である。出力層には，10 種の音響シーンに分類するため，ソフトマックス関数が用いられている。

3. 特徴量抽出部分の拡充と性能評価

3.1 本研究で実装した特徴量抽出部

本研究で拡充した特徴量の抽出方法を図 2 に示す。本手法では，まずベースラインシステムと同様に対数メルバンドエネルギーを抽出し，その時系列を変調フィルタバンクを用いて分析する。本システムで使用する変調フィルタバンクは，0.06〜25 Hz の範囲を対数スケールで分割した値を

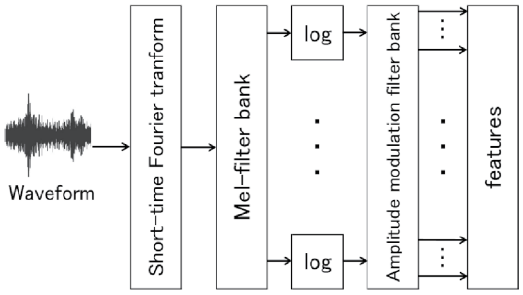

図 2 本研究で拡充した特徴量の抽出方法

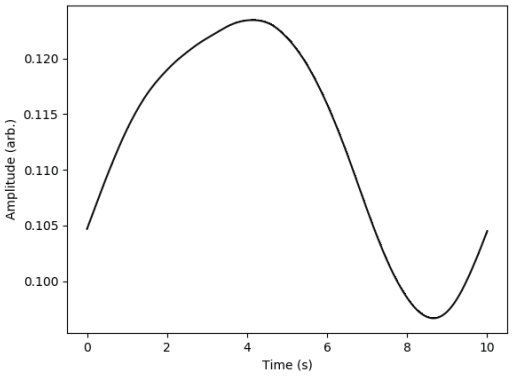

図 3 図 1 に示すデータを中心周波数 0.06 Hz の変調フィルタで分析したときに得られる特徴量の時系列データ

中心周波数とする，20 個のフィルタで構成した。この変調フィルタバンクは，Python のパッケージである pambox に含まれる pambox.central.EPSMModulationFilterbank を使用して構築した [10]。

この変調フィルタバンクに，対数メルバンドエネルギーを信号として入力することで，特徴量として用いる時系列を得た。例として，図 1 に示す波形を変調フィルタバンクで分析したとき，0.06 Hz を中心周波数とするフィルタによって得られた振幅値の時系列データを図 3 に示す。このように 40 個の時系列データの各々に対して，20 個の時系列データが出力される。すなわち，変調フィルタバンクからの出力は 800 個の時系列データとなる。

3.2 性能評価

性能評価は，ベースラインシステムに搭載されている開発モードを用いて行った。開発モードでは，学習用データセットとして用意されている，10 種の音響シーンごとに 864 個，合計 8640 個の音データのうち，7 割を学習，3 割をテストに用いて評価を行うモードである。なお，学習は，機械学習環境の構築をサポートするライブラリである TensorFlow GPU [11] を用いて行った。

まず，上記の手順で求めた全ての時系列データを特徴量として用いることが適当であるかを検討するため，予備的な検討を行った。その結果，必ずしも全ての特徴量を用いることは性能向上に寄与しないことが示された。そこで，

表 1　システムの分類性能の比較

	Accuracy (%)	
Scene label	Our system	Baseline
Airport	71.7	72.9
Bus	65.9	62.9
Metro	61.9	51.2
Metro station	61.1	55.4
Park	81.4	79.1
Public square	43.4	40.4
Shopping mall	59.6	49.6
Street pedestrian	54.9	50.0
Street traffic	81.5	80.5
Tram	54.3	55.1
Average	63.6	59.7

分類性能の向上に寄与する変調フィルタを選択するための検討を行った。その結果，20 個の変調フィルタ特徴のうち，0.06，1.4，9.4 Hz を中心周波数とする 3 つのフィルタから得られた時系列データを特徴量として用いることが適当であると判断した。このとき，ベースラインシステムで得られる 40 個の特徴量系列のそれぞれから 3 つの新たな特徴量系列を得ることになるため，NN の入力データサイズは (120, 500) である。

本システムの実行結果とベースラインシステムの結果を表 1 に示す。数値は，学習・テストの一連の流れを 10 回繰り返した結果の平均値を示している。平均値として，本システムはベースラインシステムの性能を 3.9 %上回ることができた。ただし，シーンごとの値を見てみると，全てのシーンにおいて性能が向上したわけではない。このことから，変調フィルタから抽出する特徴量にはさらに改善の余地があるものと考える。すなわち，文献 [7], [8] に基づけば，変調フィルタ出力においてフィルタ間の相互相関係数を含む要約統計量が音の特徴として重要であることが示されているので，これらを積極的に特徴量として用いた機械学習を行うことが有効であると考えている。

ところで，変調フィルタバンクを用いて特徴量の抽出を行った音響シーン分類システムとして，Moriz ら [12] のシステムがある。このシステムは DCASE 2016 Challenge Task1 に提出されたシステムであり，ベースラインを 10% 以上も上回る性能を示している。DCASE 2018 とは課題として用いられる音データも異なるので単純な比較は困難であるが，本研究のシステムにはさらに改良の余地があることが示唆される。

4. まとめと今後の発展

本研究では，特徴抽出部に変調フィルタを用いた音響シーン分類システムの開発を行った。結果として，ベースラインシステムの性能を 3.9 %上回ることができたが，システム性能のさらなる向上が必要であることも示された。具体的には，出力を特徴量として用いる変調フィルタの選択を再度検討すること，その時系列データの要約統計量も特徴量として利用するように変更することが有効であると考える。

参考文献

[1] Y. Kinoshita, T. Muto, K. Ozawa, and T. Ise, “Development of a *Kansei* evaluation model for the scenery in automobile driving,” Proc. of International Conference on *Kansei* Engineering and Emotion Research 2009 (KEER 2009), No. 14G–04 (6 pages in electronic proceedings), (2009).

[2] Y. Kinoshita, Y. Masaki, T. Muto, K. Ozawa, and T. Ise, “Scenery based *Kansei* music selection for car audio systems,” Proc. of the 13th IEEE International Symposium on Consumer Electronics (ISCE 2009), pp. 94–98 (2009).

[3] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange and M. D. Plumbley, “Detection and Classification of Acoustic Scenes and Events,” IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1733–1746 (2015).

[4] Detection and Classification of Acoustic Scenes and Events, (online), <http://dcase.community/> (2019.5.24).

[5] T. Dau, B. Kollmeier, and A. Kohlrausch, “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers,” J. Acoust. Soc. Amer., Vol. 102, pp. 2892–2905 (1997).

[6] T. Dau, B. Kollmeier, and A. Kohlrausch, “Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration,” J. Acoust. Soc. Amer., Vol. 102, pp. 2906–2919 (1997).

[7] J. H. McDermott and E. P. Simoncelli, “Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis,” Neuron 71, pp. 926–940 (2011).

[8] J. H. McDermott, M. Schemitsch, E. P. Simoncelli, “Summary statistics in auditor perception,” Nat. Neurosci., Vol. 16, pp. 493–498 (2013).

[9] librosa development team: Document: librosa.filters.mel, (online), <https://librosa.github.io/librosa/generated/librosa.filters.mel.html> (2019.2.27).

[10] Alexandre Chabot-Leclerc: pambox: Central auditory proccessing, (online), <https://pambox.readthedocs.io/en/latest/central/index.html> (2019.2.27).

[11] TensorFlow, (online), <https://www.tensorflow.org/> (2019/3/1).

[12] N. Moritz, J. Schröder, S. Goetze, J. Anemüller, and B. Kollmeier, “Acoustic scene classification using time-delay neural networks and amplitude modulation filter bank features,” Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE 2016), pp. 70–74 (2016).