打撃音の特徴抽出と材質の識別

47-186729 藤田 健斗
指導教員 佐々木 健 教授

Recently the environmental sound recognition using deep learning has attracted a lot of attention. However, the lack of deep learning is the difficulty to understand the recognition process and the requirement of a lot of data, which is one of the problems of environmental sound recognition. Therefore, this research aims to feature value extraction. In this research, the feature values for the impact sound with persistent components, mean and variance of the saliency of persistent components, were proposed. The classification accuracy using proposed feature values was 84.5 %, which is superior to the classification accuracy based on conventional feature values. Both the misrecognition between glass and pottery, and between wood and pottery and the robustness towards reverberation were the remaining problem.

Key words: Sound event detection, Impact sound, Material recognition

1 緒言

従来，環境音（音声や音楽以外の音）は雑音として除去するか，認識するとしても機械の故障診断等での限られた音が入力される状況に留まっていた．しかし，近年の計算機技術やパターン認識技術の発展に伴い，高齢者の見守りやビデオのシーン解析，ライフログ，音による監視と異常事態の検出など様々な課題への応用可能性のもと，環境音の認識が注目を集めてきている．多様な課題への応用を見据えているため，学習によって適宜様々な音に適用可能な環境音認識手法が研究されている．

近年の研究にて，スペクトログラムを入力とする畳み込みニューラルネットワーク（CNN）を応用した認識手法が多く考案され，高い精度が得られている[5]．しかし，深層学習を用いた手法は，非常に多くのデータが必要であることと，認識の根拠が不明瞭であることが欠点である．特に多くのデータが必要であることに関して，我々が日常的に耳にする，つまり入力されうる音響イベントの種類は数百～数千以上と非常に多く，ラベル付けにおいても音の開始時刻と終了時刻を正確につける必要がある．そのため，環境音認識においては質の良い大量のデータを集めるのは容易ではない．実際に性能評価に利用されているデータセットも音の種類は 17 種類や 50 種類と非常に少ない．そのため，ここ最近の研究では必要なデータ数を抑えるための工夫が行われている．

しかし，人為的に計算した特徴量を用いて識別を行うのであれば，モデル内部のパラメータは少なく済むため，必要な学習データ数も少なく済む．また，特徴量の大小に基づいた識別がなされるため，識別の根拠もある程度明瞭なものとなる．

音の特徴量の計算方法はどのような音を対象とするのかに依存する．雨音や薪の燃える音，風音などの sound texture と呼ばれる音の特徴量については知られており，それによる環境音の識別も行われている．しかし，それらの手法は持続のある突発的な音が含まれる場合を苦手としており，そのような音の特徴量は知られていない．

持続のある突発的な環境音は多く存在し，例えば，ガラスの割れる音（図 1 に示す）や踏切の打鐘式警報器，自転車のベルの音は異常音検出において重要であり，食器の音は食事中を表すためシーン解析やライフログにおいて重要である．それ以外にも，流しに落ちる水滴の音など，環境音の構成要素としても持続のある突発的な音は現れる．

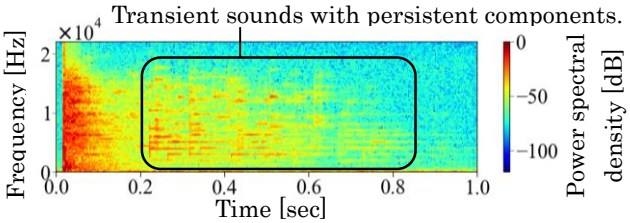

Fig. 1 The spectrogram of a sound of glass breakage.

突発音は日常的に頻発する．例えば，足音や物の落下音，ちょっとした衝突音である．そのような雑音となる突発音と区別して重要な意味を持つ突発音を認識するには，振幅の急激な変化だけでなく，音色による識別が必要である．人がそれらの音を聞き分ける場面から類推すると，木のような音，金属音，ガラスらしい音，のような材質による音色が重要な手がかりの一つなのではないかと考える．

そこで本研究では，持続のある突発的な音として，打撃由来の突発音を取り上げ，材質の違いを音から識別するための特徴抽出を目的とする．また，特定の形状に限定した材質の違いではなく，形状や接触状況の違いにある程度頑健な特徴量を目指す．

本研究では，RWCP（Real World Computing Partnership）実環境音声音響データベースを利用した．

2 持続成分の顕著さを表す特徴量

2.1 打撃由来の突発音

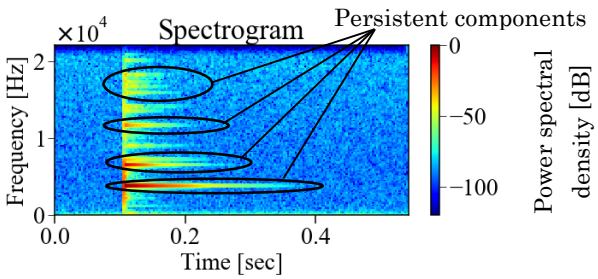

Fig. 2 The spectrogram of the sound of striking a glass. Black circle indicates the example of the area where persistent component can be seen.

打撃由来の突発音のスペクトログラムを図 2 に示す．特定の周波数で短時間持続する成分が複数確認できる（例えば，図中黒丸で囲われた領域内）．打撃由来の突発音は主にこのような複数の持続成分の組み合わせによって成り

立っている．この持続成分は音源物体の振動に対応し，その振動数は振動モードに，振幅は振動の初期値に，減衰の早さは減衰能に由来するため，持続成分は打撃由来の突発音を特徴づける要素であり，材質に関係する要素である．

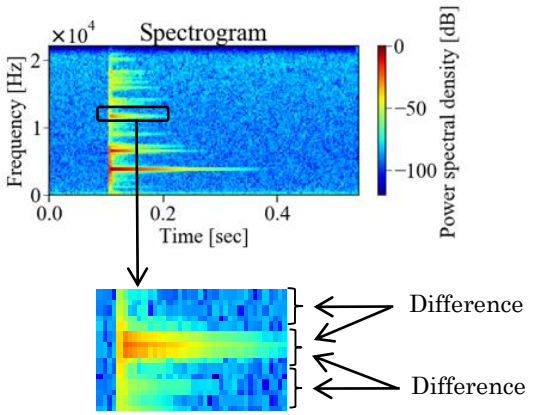

Fig. 3 A sound of striking a glass and one of its persistent components.

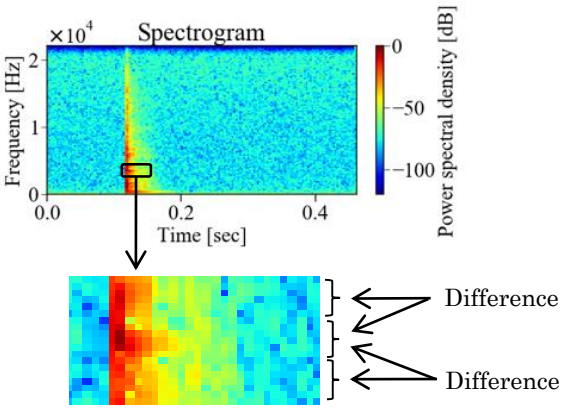

Fig. 4 A sound of striking a wood board and one of its persistent components.

2.2 持続成分の顕著さ

木板を叩く音やガラス容器での音の観察から，持続成分に当たる周波数帯での音のパワーと隣接する周波数帯での音のパワーの差が，恐らく減衰の早さに由来するものだが，特徴となるのではないかと考えた．図 3 にガラス瓶を叩く音のスペクトログラムを，図 4 に木版を叩く音のスペクトログラムを示している．図 3 では，持続成分に隣接する周波数帯は青く，パワーの差は大きくなっている．一方，図 4 では，隣接する周波数帯も赤や青になっており，パワーの差は小さい．

このように，ガラス由来の打撃音や陶器由来の打撃音ではパワーの差が大きいが，木板を叩く音や金属板を叩く音ではパワーの差が小さくなっていると考え，特徴抽出を行った．

計算の概要を図 5 に示す．まず短時間フーリエ変換で得られたスペクトログラムに対し，フィルタを畳み込む．これは，隣接領域との差分を取り，横線状の部分を強調することを意味する．持続成分がある部分に対応した領域ではフィルタの出力で正の値を持つが，背景雑音部分でも正の値を持つ部分が存在するため，雑音に対応した領域の除去を行う．さらに，純粋なパワーの大小による値の変化を打ち消すため，対応する時間・周波数でのスペクトログラムのパワーでフィルタから出力された値を割る．これによ

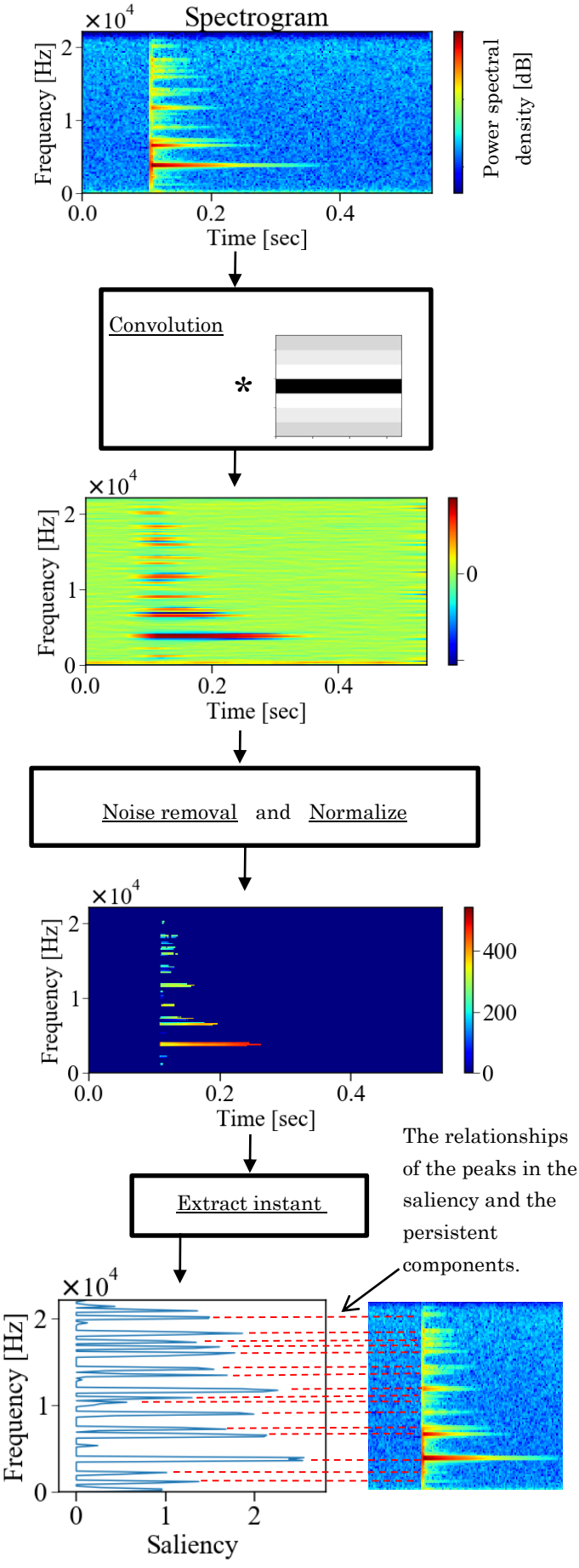

Fig. 5 Flow chart of the calculation of the extraction of the saliency of persistent components.

り，持続成分に対応した領域以外では値が 0 となり，持続成分に対応した領域では，フィルタにより隣接周波数との差分がとられた値となる．最後に，音の立ち上がりに相当するフレームを抽出する．

図 4 の最終段が抽出されたフレームでの計算された顕著さである．この図に示されている赤い破線はスペクトログラムで確認できる持続成分と，計算された顕著さのピークとの対応付けを示している．持続成分がうっすらとでも表れている部分では顕著さがピークとなっている，つまり値を持っていることが確認できる．

その後，0 でない部分の平均と分散を計算し 2 次元の特徴量とした．

使用しているフィルタは以下の式 1 で表される$h(x,y)$であり，平均が 0 であるため大域的なオフセットを除去し，急峻なピークのみを抽出している．

$$h(x,y)=\begin{cases}0 & if\ \ y\geq 3\\ -0.15 & if\ \ y=2\\ -0.35 & if\ \ y=1\\ 1 & if\ \ y=0\\ -0.35 & if\ \ y=-1\\ -0.15 & if\ \ y=-2\\ 0 & if\ \ y\leq -3\end{cases} \quad (1)$$

3 特徴量による識別

3.1 識別実験に用いたデータ詳細

計算した特徴量で打撃由来の突発音の材質の違いを識別できるのかを調べるため，RWCP-SSD に含まれる打撃由来の突発音を用いて識別実験を行なった．RWCP-SSD は国立情報学研究所より提供されている 105 種類約 9700 音が収録されている環境音のデータセットであり，無響室で録音されているのが特徴である．本研究では，収録されている音のうち，木由来の打撃音として木板を叩く音を，金属由来の打撃音として金属板を叩く音と金属製ボウルを叩く音を，ガラス由来の打撃音としてガラスコップを叩く音とガラス瓶を叩く音を，陶器由来の打撃音として陶器を叩く音を利用した．それ以外に，持続成分の現れない突発音として，プラケースを叩く音と手を叩く音を利用した．各約 300 個，合計 1740 個の音である．これらの音は，形状や叩き方を変化させて録音されている．

木，金属，ガラス，陶器由来の打撃音はそれぞれの材質の違いを識別できるかどうか，プラケースを叩く音と手を叩く音は持続成分の現れない音を除外することができるかどうかを調べるためのものである．

3.2 既存特徴量との比較と特徴選択

打撃由来の突発音に関する研究はいくらか存在し，特徴量もいくつか提案されている．それらの特徴量との比較として，持続成分の顕著さを含めた場合と含めずに既存特徴量のみを用いた場合との識別精度の比較を行った．

既存特徴量は先行研究 [1–3] で提案されている計 13 個を利用した．精度の計算には，音を 5 個のグループに分割し，一つをパラメータの調整用に使い，残り 4 つのグループを使った 4 分割交差検証を行なった．識別にはジニ係数に基づく決定木を用いた．

結果，既存特徴量のみを用いた場合では 73.1±4.8 %であったのに対し，持続成分の顕著さを含めた場合では 84.8±1.0 %となり，識別精度の向上が確認された．

Table. 2 Comparison of the classification accuracy.

Methods	Accuracy
Wang[4]	70.0±7.2 %
This work	84.5±4.7 %
Kumar[5]	91.7±4.8 %

Table. 1 Confusion matrix of the classification using proposed feature values and decision tree.

		Predicted label					
		metal	glass	plastic	clap	pottery	wood
True label	metal	95.9	0.3	0.3	0	3.4	0
	glass	0.6	80.6	0	0	18.8	0
	plastic	1.8	0	86.4	11.8	0	0
	clap	0	0	1.3	94.0	2	2.7
	pottery	3.3	27.1	0	0	55.8	13.7
	wood	0.8	0	0	0.4	7.9	90.8

Table. 3 Confusion matrix of the classification using the method proposed by Kumar et. al[5].

		Predicted label					
		metal	glass	plastic	clap	pottery	wood
True label	metal	90.6	7.2	0	0.9	0	1.3
	glass	0	96.2	0	0	3.8	0
	plastic	0	0	95.5	4.5	0	0
	clap	0	0	0	98.7	0	1.3
	pottery	3.3	23.3	0.4	0	68.8	4.2
	wood	0	0	0	1.2	0	98.8

次に，識別に有効な特徴量を炙り出すため，特徴量の取捨選択を実験的に行なったところ，持続成分の顕著さの平均と分散を含む 4 種類の特徴量を用いた識別で，84.5±4.7 %の精度が得られた．そのため，以後の実験ではこの 4 種類の特徴量を利用している．4 種類の特徴量のうち，持続成分の顕著さ以外は，Giordano らの研究 [3] で使われている tanϕ と呼ばれる特徴量と，大久保ら [1] が提案しているスペクトルの滑らかさと呼ばれる特徴量である．tanϕ は各周波数での減衰の傾きを平均したもの，スペクトルの滑らかさは振幅スペクトルを信号として捉えたときに，ある閾値以上の周波数を持つ成分に対する，閾値以下の周波数を持つ成分の割合を意味する．

決定木による識別を行っているため，特徴量による識別の過程を確認した．持続成分の顕著さに基づく分離を行っているノードについて調べると，持続成分の顕著さの分散は持続成分がはっきりと現れているガラスと陶器を分離することに，持続成分の顕著さの平均は木板と手を叩く音の分離，金属とプラケースを叩く音の分離に使われていた．木板を叩く音と手を叩く音の組も，金属を叩く音とプラケースを叩く音の組も，どちらも持続成分が多少なりとも現れているか否かが音の違いとなっている．

以上のことから，持続成分の顕著さの平均と分散が既存特徴量では捉えられていない特徴を抽出していること，当初の予測通り持続成分がどの程度はっきりと現れているかどうかによる識別が可能であることが確認された．

3.3 既存識別手法との比較

特徴量による識別と比較するため，RWCP-SSD で高い識別精度が得られているスペクトログラムの主成分分析に基づく手法[4]と，大規模データセットで学習した CNN を用いた転移学習により高い識別精度が得られている手法[5]での識別を行い識別精度の計算を行った．識別精度を表 1 に，特徴量を利用した識別の混同行列を表 2 に，転移学習に基づく識別の混同行列を表 3 に示す．

特徴量を利用する手法での識別精度は 84.5 %となり，Wang らの手法よりも精度は高いもが Kumar らの手法の 91.7 %よりは低い精度となっている．4 次元の特徴量を利用して高い精度が得られてはいるが，混同行列を比べると，この違いは主に木板を叩く音と陶器を叩く音の混同及びガラスを叩く音と陶器を叩く音の混同に起因している．

ガラスを叩く音と陶器を叩く音は実際に聞いてみても非常によく似ている音であり，特徴量を用いた識別でも転移学習に基づく識別でも誤認が発生しているため，そもそも識別が困難な音である．しかし，4 種類の特徴量による識別では転移学習による識別に劣っている差分は特徴量抽出の過程で抜け落ちた情報によるものであるため，改良によって識別性能が向上する可能性がある．

木板と陶器の混同の原因は，一部の木板において持続成分の顕著さの分散が大きな値を取っていることである．それにより陶器との混同が発生している．これら 2 種類の混同が課題として残されている．

3.4 反響の違いによる影響

ここまで，RWCP-SSD に収録された無響室で録音された音源を利用したが，実用においては様々な反響が畳み込まれることになる．そこで，無響室で録音された音で学習したモデルを用いて，会議室での反響が畳み込まれた音の識別を行った．音は，3.1～3.3 で用いている音と同じ音を利用し，学習時に利用した音と全く同じ音に反響を畳み込んだ音を試験に用いている．識別結果の混同行列を表 3 に示す．結果，多くの音が金属に識別され，識別精度が非常に低くなっている．

実際の環境での部屋等の違いによる反響の違いは，無響室と実際の環境との音の違いほど大きくはなく，無響室での音は人間も普段とは違って聞こえるものである．そのため，実際の環境で収集された音を学習に用いて構築した識別器を，異なる反響を持つ別の環境で利用した際にこれだけの精度の差が発生するとは限らない．しかし，反響によって特徴量が変化しこれだけ精度が変わりうることは，実際の環境での利用を検討する場合だけでなく，音の収集においても注意が必要な重要な問題である．

4 結論

本研究では，打撃に由来する持続を持つ突発音の特徴量として，持続成分の顕著さを計算しそれによる識別を行った．これにより，既存特徴量を用いた識別よりも高い精度が得られている．そして，特徴量は持続成分の顕著さの平均と分散，tan φ ,スペクトルの滑らかさの 4 次元の特徴量を使えばそれ以外の特徴量を使わずとも高い精度が得られた．これは，基本周波数や調波構造，1 kHz～3 kHz でのパワーの全体に対する比率のような，スペクトルの詳細な構造を必要としないことも意味している．

Table. 4 Confusion matrix of the classification of the sound convolved with the reverberation of a conference room.

		Predicted label					
		metal	glass	plastic	clap	pottery	wood
True label	metal	98.3	1.3	0.5	0	0	0
	glass	12.5	83.7	0	0	1.8	2.0
	plastic	83.3	0	16.0	0	0	0.7
	clap	67.9	0	1.1	0	0	31.1
	pottery	42.4	44.1	3.1	0	1.0	9.3
	wood	10.7	0	0.7	0	0	88.7

しかし，CNN を利用した手法が 91.7 %であるのに対して特徴量による識別は 84.5 %の精度と劣っており，その差は木と陶器の混同，及びガラスと陶器の混同であった．また，反響による影響も大きく受けることが確認された．

今後の展望として，ガラスと陶器の混同及び木と陶器の混同を抑えるために必要な情報を持つ特徴量が求められる．また，本研究で利用した特徴量によって，任意の形状のガラスや木でできた物体に由来する打撃音の識別が本研究の実験と同等の精度で可能だとは思えないため，実生活において遭遇する各材質での衝突や打撃による突発音（例えばガラスの割れる音における破片の音）をどこまで識別可能かどうかの検証が必要である．そのためには反響による影響に留意し場合によっては影響を除去する必要がある．

文献

1) S. Okubo, et al., “Recognition of Transient Environmental Sounds Based on Temporal and Frequency Features. *Int. J. Autom. Technol.* **13**, 803–809 (2019).

2) M. Aramaki, et al., “Controlling the Perceived Material in an Impact Sound Synthesizer,” *IEEE Trans. Audio. Speech. Lang. Processing* 19, 301–314 (2011).

3) B. L. Giordano, “Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates,” *Acoust. Soc. Am. J.* **119**, 1171 (2006).

4) W. X. Zhou, et al., “Large Scale Environmental Sound Classification Based on Efficient Feature Extraction,” in *Proceedings of the International Conference on Parallel Processing Workshops* 2016-September, 421–425, 2016.

5) A. Kumar, et al., “Knowledge Transfer from Weakly Labeled Audio Using Convolutional Neural Network for Sound Events and Scenes,” *2018 IEEE Int. Conf. Acoust. Speech Signal Process.* 326–330 (2017).