東京大学大学院新領域創成科学研究科
人間環境学専攻

2019 年度

修士論文

打撃音の特徴抽出と材質の識別

2020 年 1 月 24 日提出

指導教員　佐々木　健　教授　　印

学籍番号　47186729

藤田　健斗

目次

第1章　序論

1.1 研究背景

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，高齢者の見守りサービスのためのライフログの作成や異常事態の検出，ビデオのシーン解析，音による室内の監視や公共空間の監視[1]，聴覚障害者のための警告音識別[2]，ロボットによる周辺環境の認識[3], [4]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おり，ここ数年では一部実用化されるようになってきている．

以上の需要を満たすため，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．

1.2 環境音認識の分類

音の大きさ，音の高さ，音色が音の感覚に関する三要素と言われる[5]．音の大きさは音圧に，音の高さは周波数に関係するが，それ以外の音の違いが音色と呼ばれている．環境音認識は音声や音楽以外の環境音を，その音色の違いから識別・検出することを目的としている．

環境音認識には様々な種類の課題が存在するが，大石[6]は環境音認識の分類を，音の特定度と時間的参照範囲の二つの要素に基づいて行なっている．

1.2.1 音の特定度の違い

音の特定度はその認識課題において音の再現性がどの程度なのかを意味する．「特定度が高い」とは，携帯電話の通知音や家電製品のアラーム音などの毎回決まった音を発する音に対して，データベースの音と照合して一致した場合に検出するような，再現性の高い音を対象とする課題を意味する．例えば，聴覚障碍者を支援するために踏み切りの警報音や緊急車両のサイレンを認識するという課題[2]は音の特定度が高い課題である．

一方，「特定度が低い」とは，例えば生活音や足音の認識のように再現性が低くスペクトルや時間変化の構造が完全に一致することが非常に珍しいような音を扱う認識課題を意味する．特定度の低い課題ではデータのばらつきを，分類に関係のないばらつきと分類に関係のあるばらつきに分ける必要がある．例えば，足音を認識する課題は，分類に関係のないばらつき（足音であれば靴や床面，個人による違い）が大きいため，音の特定度が低い課題である．

音の特定度が異なる問題に同じ手法を適用した際に識別性能が低下する場合がある（2.3 節参照）ことは注意が必要である．

1.2.2 時間的参照範囲の違い

録音された周囲の音をもとに，録音時の環境をオフィス，道路，レストラン等に識別する課題は音響シーン解析と呼ばれる．一方，キーボードの音やコピー機の音，車の走行音，歩行者の足音等の個々の音を識別する課題は音響イベント検出と呼ばれる．いずれも音の特定度は低いが，音響シーン解析は対象となるシーンはある程度長時間の音であるのに対し，音響イベント検出は数秒の音をもとに識別する．単純に識別器に入力するデータ長が異なるため，手法を変える必要性が出てくる可能性があるとともに，音響シーン解析では複数の雑多な音がまとめて入力される点も異なる課題である．

1.2.3 入力と出力の条件の違い

音が重複して発生している場合とそうでない場合とでは課題の質が異なるため，異なる手法が必要となる．課題の種類は概ね次の 3 つに分けられる．

1. 音の重なりの無い単一の音のみが含まれた音ファイルに対して 1 つのラベルを割り当てる．
2. 複数の音が重複して含まれうる音ファイルに対して対応する複数のラベルを割り当てる．
3. 複数の音が重複して含まれうる音ファイルに対して対応する複数のラベルを割り当てるとともに，各ラベルに対して音の開始時刻，音の終了時刻も出力する．

例えば打音検査で自発的に音を発生させるように，音が発生すると知っているならば，周囲が静かでありさえすれば 1 番の課題設定のまま実用まで可能だが，多くの場合実用化するには適当な長さの録音を行い，適当な長さであるがために他の音が入り込むだけでなく，対象とする音ではない音しか入っていない可能性すらある録音ファイルから認識を行うため，２番又は 3 番の課題に取り組まなければならない．しかし，1 番の課題は 2 番及び 3 番の課題の基礎となり得るため，研究が行われている．

1.2.4 対応するデータセット

公開されているデータセットを利用すれば，他の手法と直接比較を行うことができる．環境音認識に関するデータセットは複数存在するが，その中でもよく使われているものは RWCP-SSD[7]と ESC-50[8]，及び競争型ワークショップである DCASE（Detection and Classification of Acoustic Scene and Events）で使われたデータセット[9]–[11]である．

RWCP-SSD は時間的参照範囲が短い音響イベント検出のためのものであり，音の特定度は使い方によって調整できるようになっている．ESC-50 は時間的参照範囲が短く特定度の低い音響イベント検出のためのものである．RWCP-SSSD と ESC-50 のどちらも一つの音ファイルに対して単一の音のみが含まれ，そのまま使う場合 1.2.3 節 1 番の課題に相当する．DCASE では音響シーン解析と音響イベント検出の両方がそれぞれ取り組まれている．音響イベント検出は特定度が低く，1.2.3 節 3 番目の課題となっている．

1.3 先行研究

環境音認識が応用できる可能性のある課題として，ライフログの作成と異常事態検出による高齢者の見守りサービス，ビデオのシーン解析，音による室内の監視や公共空間の監視などが存在するが，どれも音の特定度が低い課題であるため，音の特定度の低い環境音認識は需要が高い．

中でも，音響イベント検出と音響シーン解析はどちらも音の特定度が低い課題として多く研究されている．音響シーン解析は音をもとに周辺環境がオフィスなのか，駅なのか，家の中なのか，街中なのか，交通量の多い道路なのかを識別するものであるが，オフィスであればキーボードの音や紙の音，駅であれば話し声や足音，放送の音，家の中であれば犬の鳴き声など，常に鳴っているわけではないが部分部分で発生する特定のイベントに紐づいた音をもとに判別することになる．そして，個々のイベントに紐づいた音に対してのラベル付けはされておらず，音の変化を察知して個々のイベントを抽出し教師なしで分類するということが行われている[12], [13]．一般に，教師なしの認識よりも教師ありの認識の方が簡単であるが，個々のイベントに紐づいた音の教師ありの認識にあたる音響イベント検出でも未だ課題は多く残されている．

以上のことから，本研究では音響イベント検出を目標とするため，次に音響イベント検出に関する先行研究をまとめる．

1.3.1 深層学習による識別

近年の研究により，ESC-50 及び DCASE での識別において，メルスケールの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を用いた手法が複数提案され，他の手法と比べて高い識別精度が得られている[14], [15]．メルスケールは実験的に決められたもので，1000 Hz を基準とし，何倍高く聞こえるか，何倍低く聞こえるかによって人の感じる音の高さを表すとされる．以下の式（1-1）で Hz をメルに変換できる[16].

$$f\ [\mathrm{mel}] = \frac{1000}{\log_{10} 2} \log_{10}\left(\frac{f\ [\mathrm{Hz}]}{1000} + 1\right) \tag{1-1}$$

メルスケールのスペクトログラムとは，周波数軸の量子化をメルスケールにしたスペクトログラムである．何故メルスケールのスペクトログラムが良いのか，メルスケールのスペクトログラムよりも優れた表現が存在するのかは不明だが，線形周波数でのスペクトログラムや後述する MFCC（Mel Frequency Cepstrum Coefficients）を入力する手法よりも高い識別精度が得られている．

1.2.3 節 2 番及び 3 番の課題では，音の開始時刻と終了時刻の推測も必要となるため，フレーム毎に予測結果を出力できるリカレントニューラルネットワークを応用した手法が提案され

ている[17].

また，入力をメルスケールのスペクトログラムではなく音圧波形とした方法も提案されており，メルスケールのスペクトログラムを入力した手法に精度は劣るものの，それに匹敵する高い識別精度となっている[18], [19]. これらの研究では学習されたモデル内部で波形からメルスケールに似た中心周波数を持つフィルタ群が畳み込まれている．

スペクトログラムの計算は短時間フーリエ変換に基づいて行われている．短時間フーリエ変換では，波形を短時間（通常数十ミリ秒）のフレームに分割した後に各フレームでフーリエ変換を行うが，周波数の性質上，時間分解能と周波数分解能にトレードオフの関係（不確定性原理と呼ばれる）が存在し[20]，フレームが短いほど時間分解能は高いが周波数分解能は低くなり，逆にフレームが長いほど時間分解能は低いが周波数分解能は高くなる．よりはっきりと識別するために高い周波数分解能が必要か，高い時間分解能が必要かは音によって異なるため，複数のフレーム長で計算した複数のスペクトログラムを使うことで識別性能が向上することが確認されている[21]. 類似の現象として，波形を入力とした手法でも一つの畳み込みフィルタで取り込む波形の長さを 3 段階設けることで識別性能が向上することが確認されている[18].

以上のように，深層学習を用いて特徴抽出から識別までを学習モデルに任せてしまう方法は様々な手法が提案され，環境音の識別において一定の成果を上げている．しかし，深層学習の欠点は大きく 2 つ存在する．

一つ目は，非常に多くの教師データが必要なことである．例えば Wang ら[22]は合計 7.9 時間のラベル付けされた音データを用いているが，過学習によって識別精度が低下しており，より多くのデータが必要と主張している．しかし，環境音認識においては質の良い大量の教師データを集めるのは容易では無い．というのも，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〜数千以上とも言われており[23]，一個 10 秒の音各 10 個を集めるとすると，計 244 時間のデータとなる．それに加え，音の開始時刻や終了時刻も教師データとして必要となる場合はそのズレは短時間フーリエ変換におけるフレーム等の分析ウィンドウ（通常数十ミリ秒）以下に納める必要がある．更にラベル付けでは，例えば「コップの音」と「マグカップの音」は同じ音なのか，のように音の分類境界は曖昧なものであるため，前もって統一するか一つの音に対して複数人でラベル付けを行う等の対策が必要となる．以上のことから，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．この問題に対しては現在の深層学習に関する技術でも対応策が存在し，転移学習を用いた識別モデル[15]や二つの音を足し合わせてから学習することによりデータ数を嵩増しする学習方法[24]，弱教師あり学習で識別を行なう手法[10]，開始時刻と終了時刻のラベル付けを緩和した学習手法[22]，ラベルの付いていない動画と画像認識をもとに環境音認識を学習する手法[25]等が模索されている．

二つ目は，識別の根拠が現状の深層学習に関する技術では不明瞭なことである．これによる問題として，実用上の問題以外にも，手法の問題点を改良しようにも憶測と手探りで進めるし

かないことが挙げられる．例えば，複数の音が重なり合っていない 50 種類の音が含まれる ESC-50 データセットでは識別精度が最も高い手法で 86.5%となっており，他にも 80%を超える手法が複数提案されている一方，音の重複が起こる DCASE においては，出力の形式が異なるため直接の比較は難しいが，10 種類の音の識別で F 値が約 40%と低くなっている．音の重複が問題点であることは確かだが，識別の過程が不明瞭であるため，音の重複によって計算上のどこに問題が発生しているかは不確かである．

多次元のデータをより低次元のデータに変換し，その過程で識別に関係のない情報を取り除き識別に関係のある情報のみを抽出する処理を特徴抽出と呼ぶ．また，特徴抽出で得られた値を特徴量と呼ぶが，特徴量の大小によって識別を行う方法であればその識別の根拠はある程度理解できるものとなる．また，深層学習に比べ特徴表現の学習の必要性が減る分，必要な教師データの数も少なく済むことが期待できる．そのため，深層学習による識別の欠点を緩和することができる．

特徴抽出を実行するには，音を見分ける際にどこを見れば良いのか，及びその特徴要素の数値表現を得るための計算方法，そしてある特徴量で識別できる音かどうか，つまり音を見分けるポイントが同じかどうか，という識別の観点での音の分類について，の 3 つの知識が必要となる．特徴表現に基づいた識別は，根拠の明瞭さと必要な教師データ数の少なさで優れるだけでなく，それにより得られる知識は深層学習における手法の改良やデータセットの構築へのヒントを与える可能性もある．しかし，環境音の特徴表現については十分に研究されていない．

1.3.2 音響特徴量を用いた識別

音響イベント検出や音響シーン解析において，深層学習より以前には MFCC（Mel Frequency Cepstrum Coefficients）と呼ばれる特徴量を信号から計算し，識別器を用いて識別するということが行われており[26], [27]， MFCC を入力とした手法は識別精度の比較対象としてもよく用いられる．MFCC は音声認識でよく使われるもので，その計算の概要は音のスペクトルを離散コサイン変換で変換するというものである．通常，MFCC を特徴量とする場合には低次の係数のみを利用する．低次の係数のみを利用することでスペクトルにローパスフィルタを適用しスペクトルの概形（スペクトル包絡）を得るとともに次元削減を行うことが期待できる．図 1-1 に音声のスペクトルと低次の MFCC 係数の逆フーリエ変換で得られたスペクトル包絡の一例を示す．スペクトル包絡ではスペクトルの細かい変動が取り除かれていることが確認できる．

音声は声帯によるパルス列に声道による共鳴が畳み込まれることで生成されるものであり，信号のフーリエ変換で得られたスペクトルをさらにフーリエ変換することで声帯によるパルスの周期と声道による共鳴の周波数特性を分離することが期待できる[16]．そのため，音声認識において MFCC が使われる．環境音では音声のように，音源にフィルタが畳み込まれた音

というのは珍しいが，環境音認識においても MFCC を入力とする識別が行われており，入力に MFCC を単独で用いることである程度の精度が得られているだけでなく，他の単独で用いた場合に MFCC よりも高い識別精度が得られる特徴量に追加で MFCC を用いることによる精度の向上も確認されている[28], [29]．そのため，MFCC によるスペクトル包絡は音の識別に有効に働く一方，MFCC によるスペクトル包絡以外にも識別に有用な特徴が存在するはずである．

音の特徴表現は対象とする音の種類によって変化する．例えば，弦楽器や管楽器，音声は基本周波数とその定数倍の倍音成分からなる調波構造を持っている．図 1-2 にホイッスルの音のスペクトログラムの一例を，図 1-3 に音声のスペクトログラムの一例を示す．調波構造を持つ音は「調波構造を持つこと」が特徴の一つであり，それを表す harmonicity と呼ばれる特徴量を計算することで，背景雑音である環境音を取り除き音声区間のみを抽出できる[30]．また，弦楽器では倍音成分が多いほど音が豊かに聞こえ，楽器の音色を特徴付けるとされている[31]．しかし，harmonicity や倍音成分の数は環境音のような調波構造を持たない音同士の違いとなると意味をなさない．このように，音の特徴表現は対象とする音の種類によって変化する．そのため，音の特徴表現をまとめるには音の分類がまず必要である．

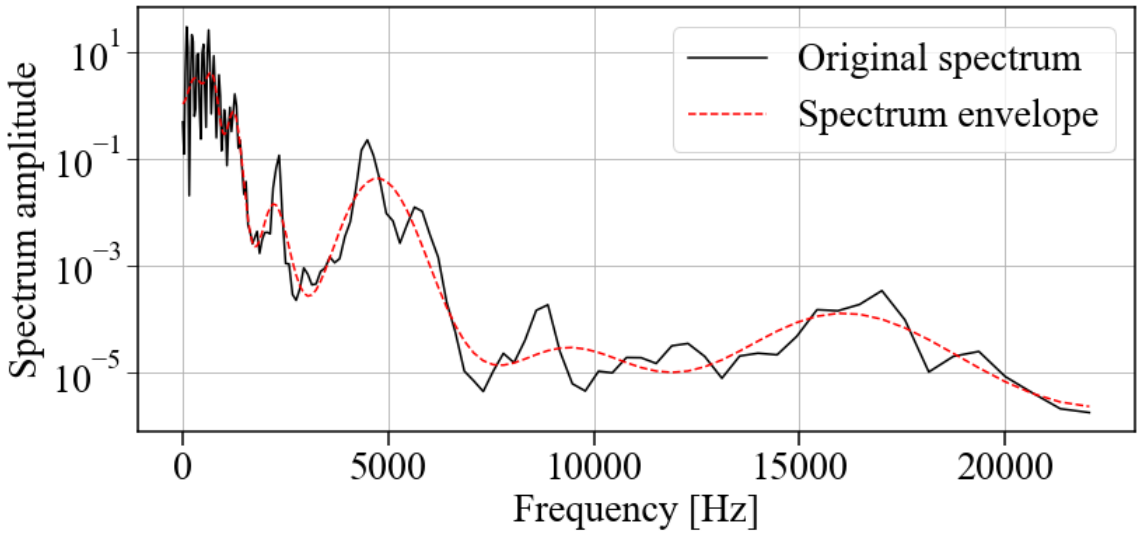


図 1-1 音声のスペクトルと MFCC をもとに計算されたスペクトル包絡．スペクトル包絡では細かな変動が除去されていることが確認できる．

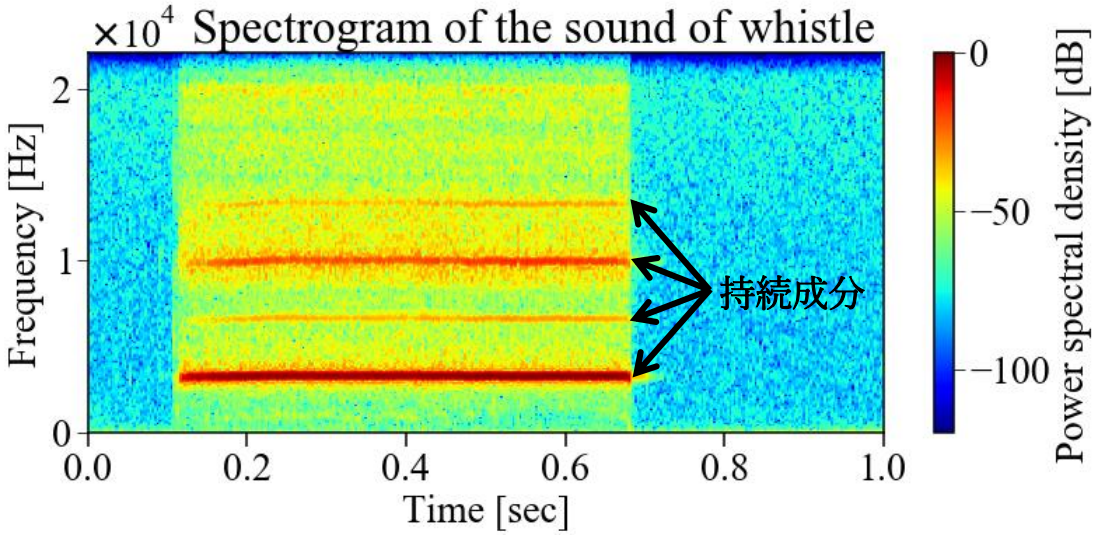


図 1-2 ホイッスルの音のスペクトログラム．3300 Hz，6700 Hz，10000 Hz，13300 Hz に持続成分が現れている．

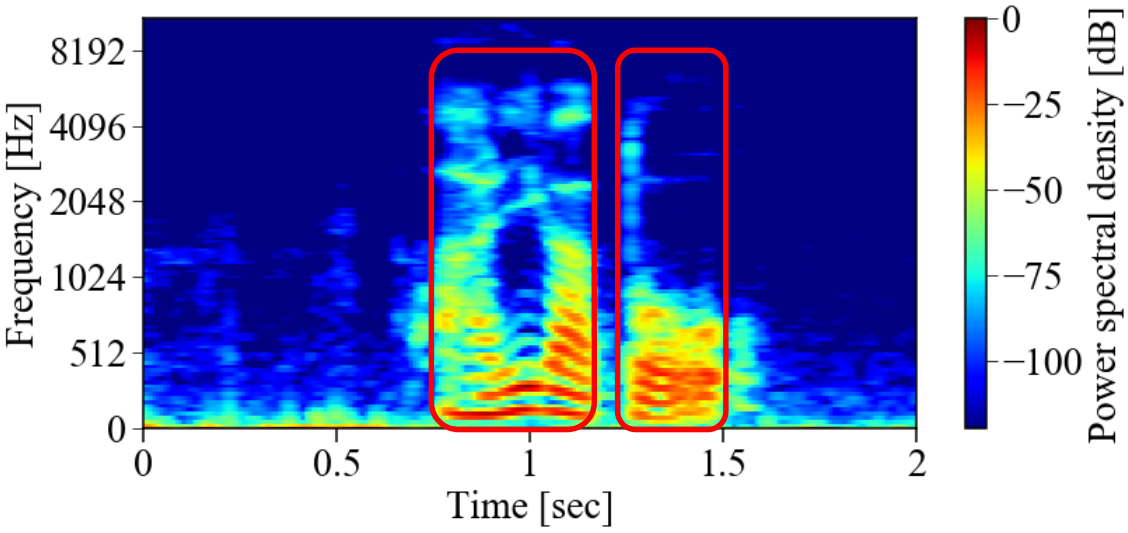


図 1-3 音声のスペクトログラム．赤枠内の音で調波構造が確認できる．

1.4 環境音の種類

音の分類方法は意味による分類（動物の鳴き声，家事の音等）や，発生要因による分類（衝突，摩擦，渦等）など様々な基準があり得るが，その一つとして，スペクトルの様子に基づく分類が考えられる．音をスペクトログラムで表すと，パワーの強い部分が密集して様々な形を形成するが，密集傾向にはある程度パターンが存在する．Saint-Arnaud[32]はそのような密集傾向に対して，”click”，”harmonic”，”noise patch” と名付けている．click は短時間かつ幅広い周波数に広がっている音を意味し，衝突音等で現れる．harmonic は特定の周波数に集中し長く続く音を意味し，金属の打撃音などで確認できる．調波構造を表す harmonicity と紛らわしいため，本論文では持続成分と呼ぶ．noise patch は時間方向と周波数方向のどちらにも広がっている音であり，例えば，スプレーの音や手を摺る音が挙げられる．全てではないが，環境音の中にはこれら 3 つを組み合わせたようなスペクトログラムとなっている音は多い．図 1-4 図 1-5 に一例を示す．これら 3 つを組み合わせた音として，次の Sound texture や打撃由来の突発音が挙げられる．

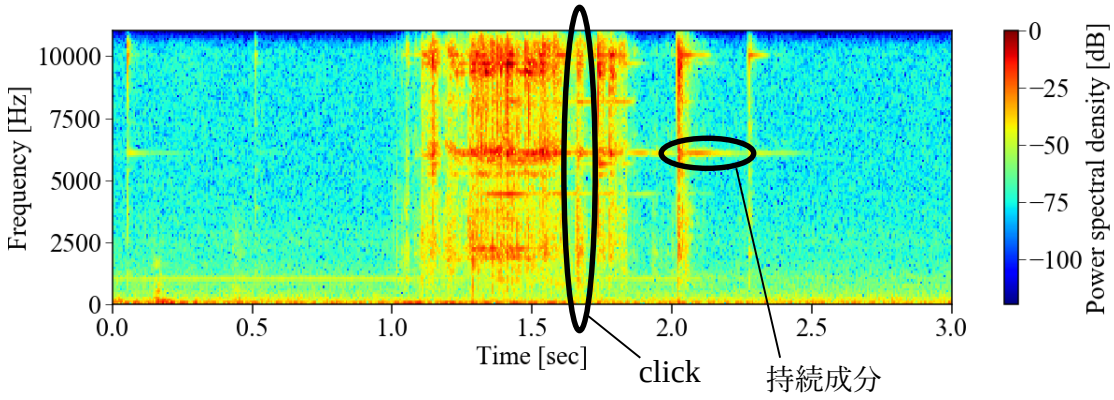


図 1-4 鍵束を取り出すときの音.

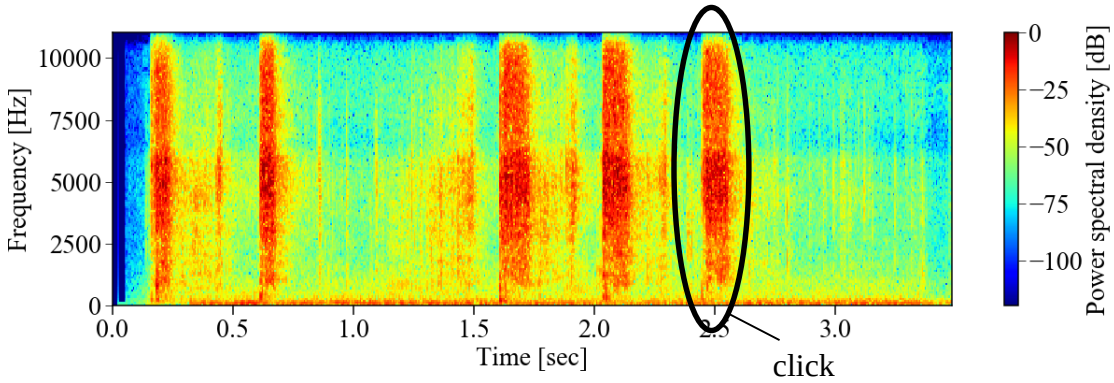


図 1-5 スリッパを履き，踵を擦りながら歩く足音.

1.4.1 Sound texture

McDermott らは Sount texture と呼ばれる種類の音の特徴表現を考案している[33]. Sound texture は複数の click や noise patch が集まってできた音であり，かつ個々の音やその発生タイミングなどが大きく異なっているけれども集合全体を見ると音が非常によく似ているような音を指している．このような音として，例えば水の流れる続ける音やファンの音，掃除機の音が挙げられる．図 1-6 に水の流れる音の波形とスペクトログラムの例を示す．図中では click が多数確認できるとともに，約 1000 Hz 以下では音が密集しているために noise patch のようになっている．個々の click のスペクトルはそれぞれ一致しておらず，3〜12 秒では click が密集しているようにも見えるが，全体を通して水音に聞こえ，5~10 秒も 10~15 秒もどちらもよく似た水の音として聞こえる．

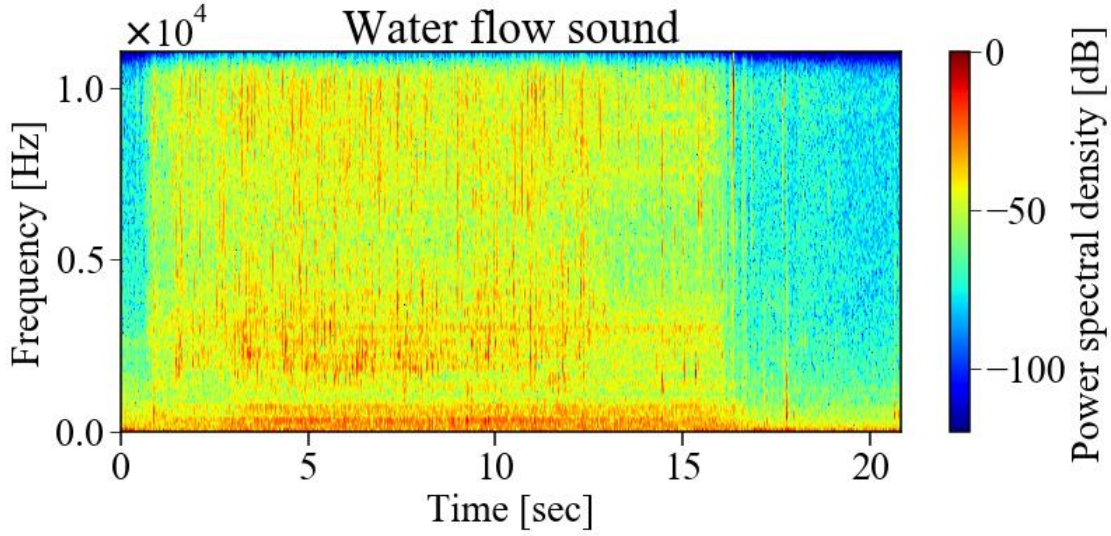


図 1-6 水の流れる音のスペクトログラムの一例.

Sound texture の特徴表現として McDermott らは音から計算された統計量をもとにその音を再現している[33]. 被験者実験にて音のリアルさを 7 段階評価した結果，雨音や川音，虫の鳴き声，ラジオの雑音，風音や大勢での拍手の音などは評価の平均が 6 以上である一方，教会の鐘の音，銅を叩く音，打鐘式の踏切警報機の音，風鈴の音のような持続成分がはっきり現れる突発音や犬の鳴き声，人の話し声，音楽は 1〜2 の評価となっている．突発音がはっきり現れる音でもヘリコプターの音は 5.67 と比較的高い評価を得ているが，この音は Gygi らの研究[34]にて，人間が聞き分ける際にはスペクトルは関係なく振幅の時間変化だけで識別できるとされている音である．時間変化の順番の情報を落とすように統計量を用いているために，持続成分がはっきり表れる突発音や犬の鳴き声，人の話し声，音楽において評価が低くなっていると思われる．

McDermott らの特徴量は，苦手な音はあるものの，雨音や虫の鳴き声のような環境音として頻出の音について人が聞いても遜色ない音が合成できる特徴量となっている．

人が聞いて遜色ない音が合成できるとしても，音を識別できる特徴量であるとは限らない．McDermott らの特徴抽出法で計算された特徴量をもとに環境音を識別する研究も行われており[29], [35], 特に Villamizar ら[35]はこの特徴抽出の計算をハードウェア実装し，更に線形カーネルによるサポートベクトルマシンを用いた識別で深層学習に比べパラメータ数を 1〜3 桁減らしつつ，深層学習による識別に匹敵する精度が得られている．別の研究では[36]，環境音ではないが，無声音である/θ/，/s/，/sh/，/f/の識別において最新の手法と同程度の精度が得られたとしている．

しかし上述の通り，個々の突発音がはっきりと含まれる音はこの方法ではうまく合成することができず，別の音に聞こえてしまう．特徴量が同じであるにも関わらず別の音に聞こえるということは，人間が聴き間違いを犯していない限り，その特徴量では区別できない音の組が存在することになる．そのため，個々の突発音がはっきりと現れる音は McDermott らの特徴抽出法では扱えない種類の音であり，識別のためには異なる特徴量が必要となる．

1.4.2 突発音

短時間で急激に立ち上がり，短時間で減衰する音を突発音と呼ぶ．手を叩く音や銃声，キーボードの打鍵音，足音などがその例である．突発音が単体で意味のある音，何らかのイベントが起きていることを表す音であることは滅多にないが，キーボードの打鍵音や足音のように突発音が集まって何らかの音を形作ることは多い．家庭内の監視で重要な音となるガラスの割れる音も，複数の突発音が集まってできる音である．突発音の中にも，打撃音や破裂音など複数存在するが，本論文では特に打撃音を取り上げる．

図 1-7 に打撃音の例としてガラスコップを叩いたときの音の波形とスペクトログラムを示す．0.1 秒にて音が急激に立ち上がり，その後速やかに減衰している．立ち上がりで突然振幅が変化するため，その不連続性によりスペクトログラム上では全周波数でパワーが大きくなり，click が形成されている．減衰部では特定の周波数で発生する減衰振動によって，その周波数では音が他の周波数に比べてゆっくりと減衰する．それにより，音源物体の振動に由来する持続成分が形成されている．そして，この持続成分のパワーは指数的に減衰する．

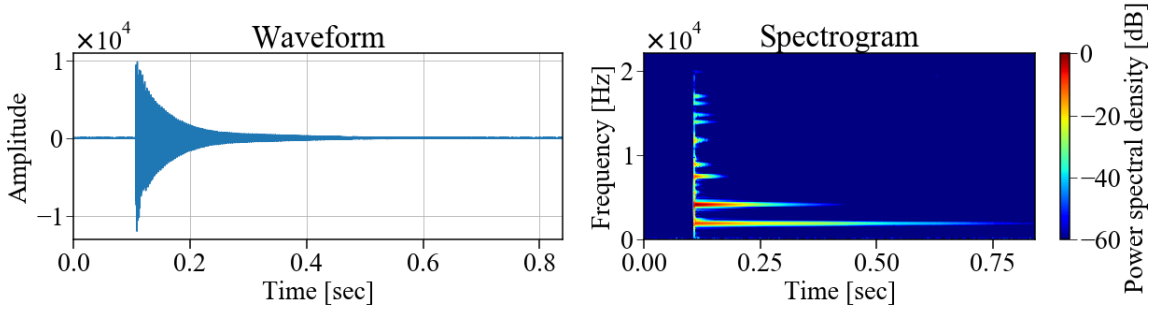


図 1-7 ガラスコップを叩く音の波形（左）とスペクトログラム（右）．

このような形のスペクトログラムは音源物体の振動に由来する音であれば，打撃音に限らず確認できる．その一例として，ガラスの割れる音が挙げられる．図 1-8 にガラスのスペクトログラムを示す．

コンピュータグラフィックスにおいて，映像内の出来事と同期したガラスの割れる音を合成する手法の研究がされており，Wang ら[37]は，ガラスの割れる音を破片の音の集合として合成し，音の合成に成功している．それによると，各破片の音は破壊時の変位および床面等との衝突によって励振され，独自に減衰振動を行うことで音を発する．さらにこの研究では，各破片の減衰振動は厳密な破片形状からではなく，楕円体で近似した形を用いて効率的に計算されている．つまり，スペクトルの詳細なマッチングはされておらず，ガラス由来の音であるかどうかが重要である．

識別の観点では，ガラスの割れる音は，最初に大きな音がした後に複数の突発音が現れながら非常に早く減衰する，という振幅の時間変化に非常に大きな特徴が現れる音である．しかし，スペクトルがフラットで時間変化だけがガラスの割れる音を踏襲するように合成した音の場合，人間は認識が難しくなる[34]．そのため，破片の音がガラスらしいスペクトルを持っていることはガラスの割れる音を特徴付ける重要な要素である可能性がある．そのため，振動由来の突発音における材質の音色による認識はガラスの割れる音の認識でも重要である．

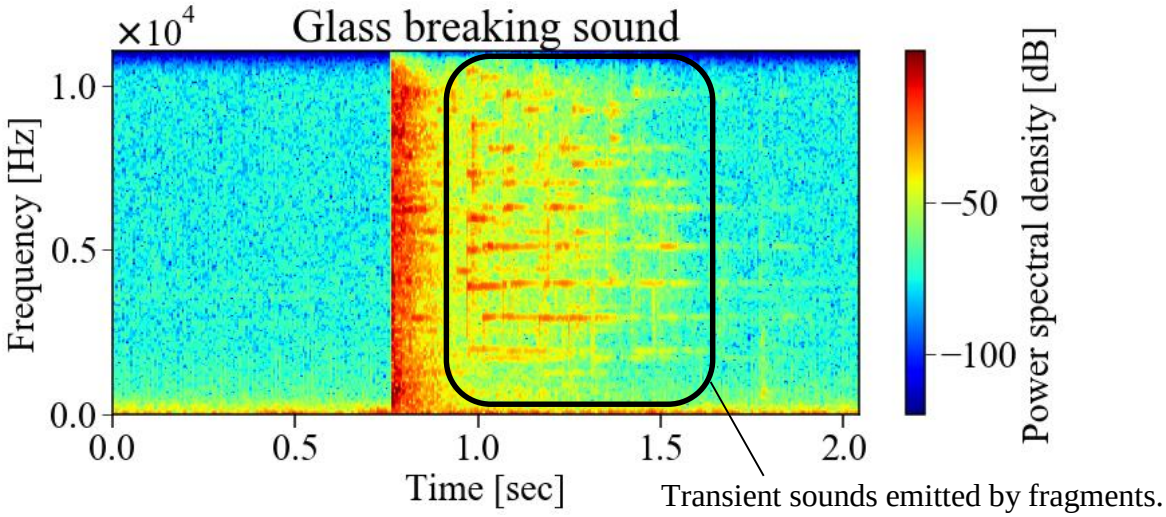


図 1-8 ガラスの割れる音のスペクトログラム．

ガラスの割れる音は異常事態を知らせる音として重要な環境音であるが，他に踏切の警鐘や自転車のベルの音も異常事態を知らせる音として重要な環境音である．また，食器の音は食事中を表す音として，ライフログや見守り介護，ビデオのシーン解析に有用であり，風鈴の音もビデオのシーン解析において重要な音である．他にも金の音や流しに水が落ちる音など，持続成分を持つ突発音が表れる環境音でかつ認識することに意味がある音は多い．

しかし，突発音は足音や物の落下音，ちょっとした衝突音として日常的に頻発する音である．そのような雑音となる突発音と区別して上記の重要な音を認識するには，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．そのため，持続を持つ突発音の音色を表す特徴表現が必要となる．

人が持続を持つ突発音や上記のガラスの割れる音，食器の音，鐘の音など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掛かりなのではないかと考える．そのような材質による持続成分を持つ突発音の音色の違いとして，打撃による突破音の材質の違いが考えられる．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している．打撃による振動は，ガラスの割れる音での破片の振動や鐘の振動などと本質的には大きな違いはなく，それにより発せられる音もその発生原理に大きな違いはないため，打撃音に限定してしまっても問題ない．

先行研究にて，打撃による突発音の特徴量として，音色の違いから材質の違いを認識するための特徴量が提案されている[38]–[40]．これらの研究では形状の影響を受けず材質の違いを表す特徴量として，内部減衰を表す以下の$\tan\phi$を利用している．

$$\tan\phi = \frac{a_i}{\pi f_i} \tag{1-2}$$

ここで，元の信号を N 個の周波数帯に分けてから減衰部分（パワー最大となる時刻とその後最大パワーの $1/e$ まで減衰する時刻の間）のみを取り出した i 番目の周波数帯の信号の時間変化を$x_i(t)\ (i = 1, 2, \dots, N)$としたとき，$\log\big(x_i(t)\big) \approx a_i t + b_i$ と直線近似して得られた傾きをa_iとしている．また，$f_i\ (i = 1, 2, \dots, N)$を各周波数帯の中心周波数としている．

f_iに関する関数$\tan\phi$を二次関数で近似した係数[39]および，以下の式(1-3)で計算される特徴量[38]を用いることで材質の識別が可能とされている．

$$\frac{\sum_{i=1}^{N} \frac{a_i}{\pi f_i} w_i}{\sum_{i=1}^{N} w_i} \tag{1-3}$$

ここで，各周波数のパワーの合計をw_iとした．

しかし，それぞれ音源物体が棒状である場合[39]と板状である場合[38]にて識別可能であるか確認されており，それ以外の形状や，手に持っているなどで他の物と接触している状況では識別性能が著しく低下する．

人の材質の聞き分けと特徴量の関係性についても議論されており，Giordano[38]は式（1-３）の値の違いで識別可能なものの，人の認識ではスペクトルの分布に関する特徴も参照しているようだとしている．Aramaki ら[41]は，４種類の特徴量をガラス，金属，木の３種類の材質に由来する突発音から計算し，それらの特徴量やその中間の値を再現するように音の加工を施すことで，音色の変換，例えばガラス由来の音を金属らしく聞こえる音やその中間の音に加

工すること，に成功している．Koumura ら[42]は人間が突発音を認識する際，空間の反響に関しての事前知識のあるなしで認識精度が変化すること，反響が未知の場合には$\tan\phi$のような時間変化に関する特徴よりもスペクトルの分布に関する特徴へと判断基準の重みが偏ることを示している．

しかし，これらの研究にて用いられている特徴量では，打撃に由来する突発音の材質を高い精度では識別できず（4 章参照），金属音とガラスの音と木の音，のような日常生活において容易く聞き分けてしまう音であっても，その特徴がどこに現れているのかは明らかになってはいない．

1.5 研究目的

本研究では，持続成分を持つ突発的な音として，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．

具体的には，打撃による 4 種類の持続成分を持つ突発音（木板を叩く音，陶器を叩く音，ガラス容器を叩く音，金属板を叩く音），及び持続成分を持たない 2 種類の突発音（プラケースを叩く音と手を叩く音）の計 6 種類での識別を行う．

持続成分を持つ 4 種類の突発音はその材質による識別のため，持続成分を持たない 2 種類の突発音は持続を持たない音を除去して識別を行うことができるかを評価するためである．

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．

持続成分を持つ突発的な音の識別を行う理由は，ガラスの割れる音や踏切の鐘の音，自転車のベルの音，食器の音など，異常音検出やビデオのシーン解析において認識したい様々な音に持続成分を持つ突発音が表れるためである．そして材質の違いの識別を行う理由は，これらの音の音色として，金属音であることやガラスらしい音であることが重要な要素であると考えたためである．また，これらの持続を持つ突発音の生成要因は打撃による音と本質的に大きな違いはないため，打撃音に限定して問題ないと考えている．

特徴抽出を行うのは，識別の根拠を明確にするとともに，必要なデータ数を少なくするためである．

1.6 本論文の構成

２章では, 本研究で用いる RWCP-SSD データセットに収録された音の説明, および学習データとテストデータの分割について述べる.

３章では，本研究で新たに用いるパワー最大時のスペクトル重心と持続成分の顕著さの平均と分散という特徴量の計算方法について述べた後，既存特徴量の計算方法についても簡単に述べる.

続く 4 章では，3 章で述べた特徴量の評価のため，識別実験を行う．その際，mRMR[43]による特徴選択を行なった．結果，既存特徴量の識別精度 73.1 %を超える識別精度 84.8 %が得られている．次に，識別精度の比較対象として，スペクトログラムの主成分分析に基づく識別[44]と畳み込みニューラルネットワークでの転移学習に基づく識別[15]を行なっている．主成分分析に基づく識別の精度は 70.0 %，転移学習に基づく識別の精度は 91.7 %，計算した特徴量での識別の精度は 84.5 %であった．転移学習に基づく識別の精度との差について，誤識別の傾向の分析を行っている. 次に, 反響が異なる場合にどのように出力されるのかについても実験したところ, 識別精度の著しい低下が確認された. 最後に, 学習時に利用していないクラスの音を入力した場合の出力を調べることで，未知の音が入力された場合の挙動を確認している. 持続成分を持たない突発音は手を叩く音とプラケースを叩く音に, そうでない音はそれぞれ対応したクラスへと識別される結果となった.

最後に 5 章では，本研究のまとめを行う.

なお，本論文では技術研究組合，新情報処理開発機構の提供する RWCP-SSD (Real World Computing Partnership – Sound Source Database) を利用した．本章で利用した音はガラスの割れる音以外は全て著者が録音した音であるが，以後の章で利用する音は全て RWCP-SSD データセットに収録されている音である．本章で利用したガラスの割れる音は ESC-50 データセットから利用した.

第2章　突発音データ詳細

2.1 はじめに

本章では，以後の章で利用するデータの詳細説明と雑音除去方法，学習データとテストデータの分割について述べる．

利用する音は，打撃による 4 種類の持続成分を持つ突発音(木板を叩く音，陶器を叩く音，ガラス容器を叩く音，金属板を叩く音)，及び持続成分を持たない 2 種類の突発音(プラケースを叩く音と手を叩く音)の計 6 種類である．

これらの音は RWCP-SSD データベース[7]に含まれる突発音だが，その中にはスペクトルの良く似た音も含まれている．その構成について述べた後に，スペクトルの良く似た音の扱いについても述べる．

RWCP-SSD データベースは無響室の静かな環境で録音されたものだが，全く雑音が含まれないわけではなく，低周波数の定常雑音，ホワイトノイズのような背景雑音，一度の打撃で複数の衝突音が発生してしまうチャタリングの 3 種類の雑音が確認できる．これらの雑音の扱いについて述べる．

2.2 RWCP-SSD の構成

2.2.1 利用する音の詳細

打撃由来の突発音の特徴の分析を行うにあたり，RWCP-SSD データセットに収録されている音を利用した．RWCP-SSD データセットは国立情報学研究所から提供されるデータセットであり，木板や陶器皿を叩く音が含まれる「衝突系音源」，ヤスリで擦る音やスプレーを噴射する音，手を叩く音が含まれる「動作系音源」，鈴音やコインを落とす音，紙を破る音が含まれる「特徴的音源」，合計 105 種類の音約 9700 個のサンプルが収録されている．これらの音は無響室にて録音されている．音源以外に会議室や畳部屋，残響可変室で計測されたインパルス応答（計測には TSP[45]を利用）が全 8 種類収録されているため，反響を再現することも可能となっている．収録されているデータはサンプリング周波数 48kHz の 16bit 整数である．ただし，本研究では広く使われているサンプリング周波数である 44.1kHz にダウンサンプリングしてから利用している．より詳細な録音条件は[7]参照.

続く 4 章では RWCP-SSD から 18 種類の音計 1740 個のサンプルを利用するため，その詳細説明を行う．18 種類の音は，木由来の打撃音，金属由来の打撃音，陶器由来の打撃音，ガラス由来の打撃音，プラスチック由来の打撃音，手を叩く音の 6 種類に分けられる．

木由来の打撃音，金属由来の打撃音，陶器由来の打撃音，ガラス由来の打撃音の計 4 種類は持続成分を持つ音，プラスチック由来の打撃音と手を叩く音の 2 種類は持続成分を持たない音である．持続成分を持つ 4 種類の突発音はその材質による識別のため，持続成分を持たない 2 種類の突発音は持続を持たない音を除去して識別を行うことができるかを評価するため選出している．

1. 木由来の打撃音

RWCP-SSD に収録される以下の 3 種類の音計 300 個を利用．“teak1”，“cherry1”，“magno1”は RWCP-SSD 内で利用されている名称である．その後の 000.wav 等はファイル名であり，通し番号が振られている．図 2-1 に木板 ABC を叩く様子の画像を，図 2-2 に波形とスペクトログラムの一例を示す．

teak1, 000.wav～099.wav：木板 A を手で持ち，木棒で 1 回叩く音．計 100 音．

cherry1, 000.wav～099.wav：木板 B を手で持ち，木棒で 1 回叩く音．計 100 音．

magno1, 000.wav～099.wav：木板 C を手で持ち，木棒で 1 回叩く音．計 100 音．

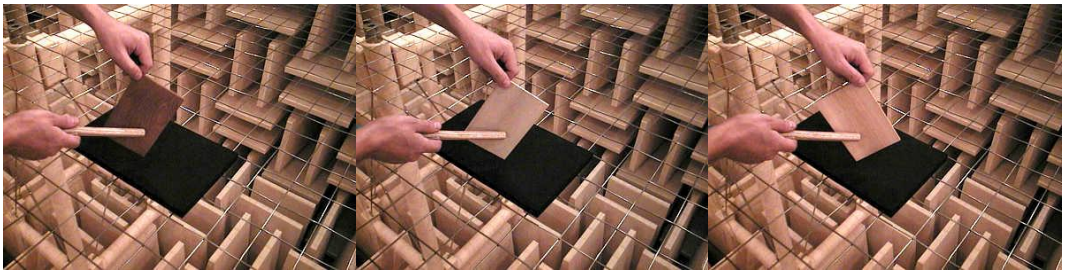


図 2-1 木板を叩く様子．左から順に木板 A，木板 B，木板 C の画像．RWCP-SSD データセット配布 CD より引用．

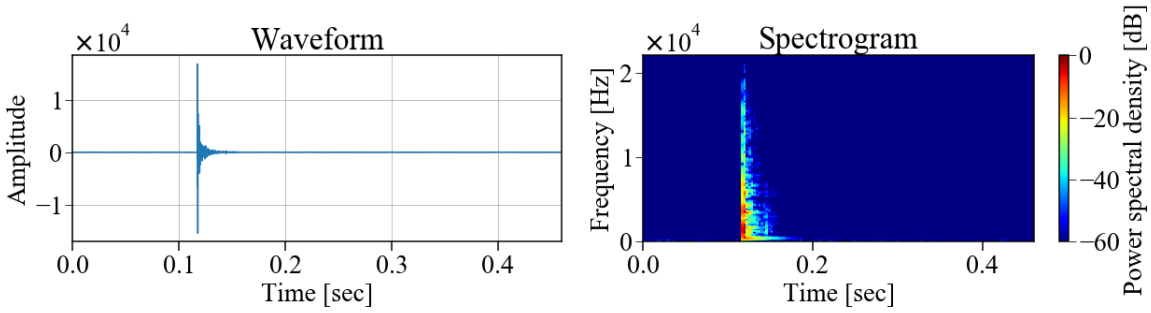


図 2-2 木板を叩く音（cherry1, 000.wav）の波形（左）とスペクトログラム（右）例．スペクトログラムは窓幅 256，ハミング窓による短時間フーリエ変換で計算．

2. 金属由来の打撃音

RWCP-SSD に収録される以下の 4 種類の音計 400 個を利用．図 2-3 に波形とスペクトログラムの一例を示す．

metal05, 000.wav～099.wav：金属板（0.5 mm 厚）を手で持ち，金属棒で 1 回叩く音．計 100 音．

metal10, 000.wav～099.wav：金属板（1.0 mm 厚）を手で持ち，金属棒で 1 回叩く音．計 100 音．

metal15, 000.wav～099.wav：金属板（1.5 mm 厚）を手で持ち，金属棒で 1 回叩く音．計 100 音．

bowl, 000.wav～099.wav：金属製ボウルを手で持ち, 金属棒で 1 回叩く音. 計 100 音.

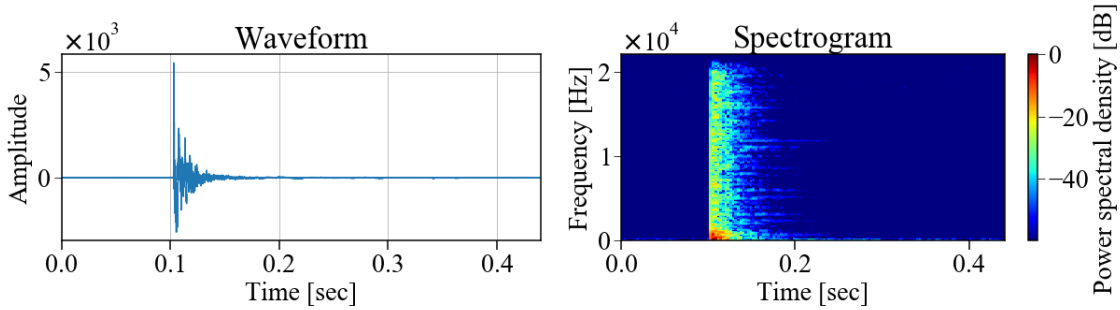


図 2-3 金属板を叩く音（metal05, 000.wav）の波形（左）とスペクトログラム（右）例.

3. 陶器由来の打撃音

RWCP-SSD に収録される以下の 3 種類の音計 300 個を利用. 図 2-4 に陶器 ABCDE の画像を示す. 底の深い皿やコップも含まれており, 金属製ボウルやガラスコップと似た形状のものも含まれていることになる. 図 2-5 に波形とスペクトログラムの一例を示す.

china1, 000.wav～099.wav：陶器 ABCDE を吸音板上，横を木棒かスプーンで 1 回叩く音. 計 100 音.

china2, 000.wav～099.wav：陶器 ABCDE を吸音板上，底を木棒かスプーンで 1 回叩く音. 計 100 音.

china3, 000.wav～099.wav：陶器 ABCDE を手に持ち，横を木棒かスプーンで 1 回叩く音. 計 100 音.

図 2-4 陶器 ABCDE の画像. RWCP-SSD 配布 CD より引用.

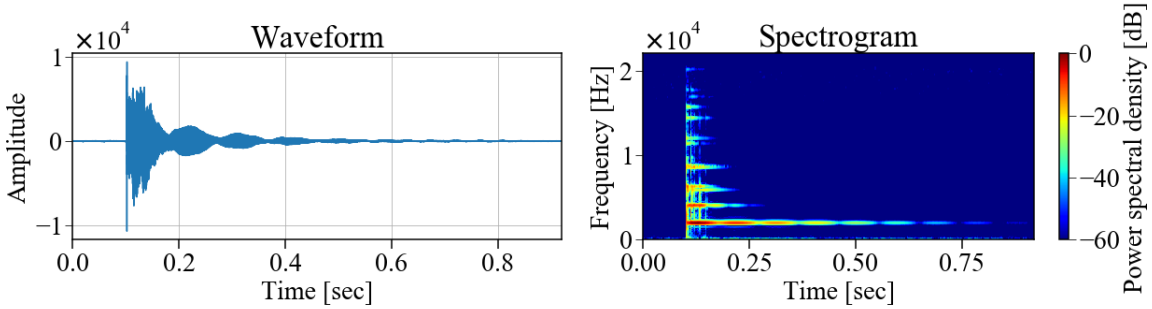


図 2-5 陶器を叩く音（china1, 000.wav）の波形（左）とスペクトログラム（右）例.

4. ガラス由来の打撃音

RWCP-SSD に収録される以下の 4 種類の音計 400 個を利用．図 2-6 にガラスコップ ABCDE とガラス瓶 ABCDE の画像を, 図 2-7 に波形とスペクトログラムの一例を示す.

cup1, 000.wav～099.wav：ガラスコップ ABCDE を吸音板状，木板かスプーンで 1 回叩く音．計 100 音.

cup2, 000.wav～099.wav：ガラスコップ ABCDE を手に持ち，木板かスプーンで 1 回叩く音．計 100 音.

bottle1, 000.wav～099.wav：ガラス瓶 ABCDE を吸音板上，木板かスプーンで横から 1 回叩く音．計 100 音.

bottle2, 000.wav～099.wav：ガラス瓶 ABCDE を吸音板上，木板かスプーンで口を上から 1 回叩く音．計 100 音.

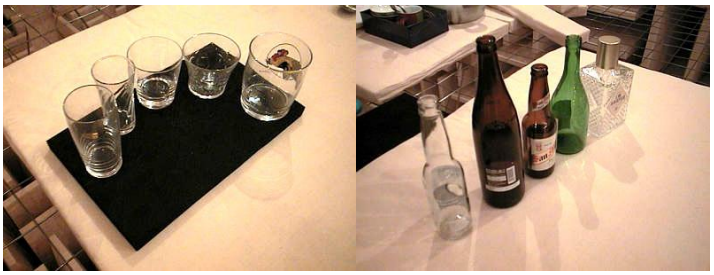


図 2-6 左はガラスコップ ABCDE，右はガラス瓶 ABCDE の画像．RWCP-SSD 配布 CD より引用.

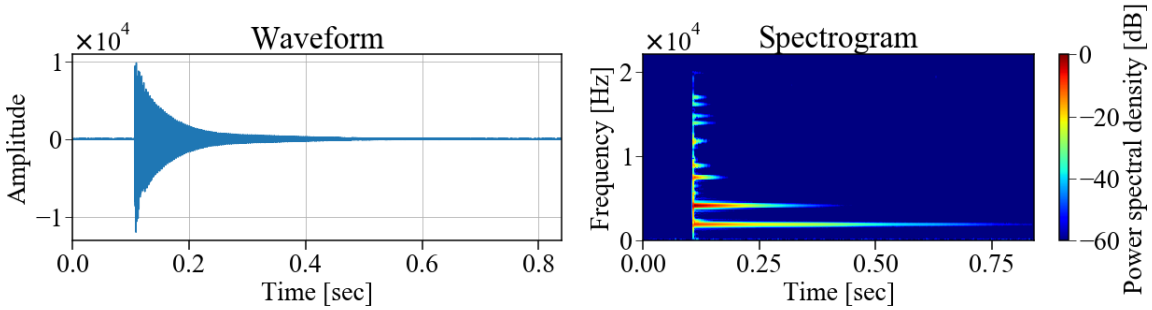


図 2-7 ガラスコップを叩く音(cup1, 000.wav)の波形（左）とスペクトログラム（右）例.

5. プラスチック由来の打撃音

RWCP-SSD に収録される以下の 2 種類の音計 200 個を利用．図 2-8 に波形とスペクトログラムの一例を示す.

case1, 000.wav～099.wav：プラケース ABC を手で持ち，木棒で 1 回叩く音．計 100 音.

case2, 000.wav～049.wav：プラケース ABC を手で持ち，勢い良く閉じる音．計 50

音.

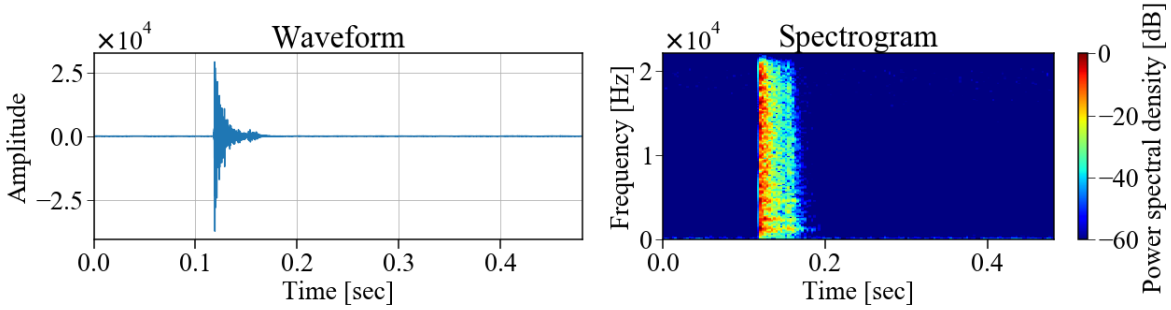


図 2-8 プラケースを叩く音（case1, 000.wav）の波形（左）とスペクトログラム（右）例.

6. 手を叩く突発音

RWCP-SSD に収録される以下の 2 種類の音計 200 個を利用．手の叩き方によるスペクトルの変化が確認されている[46]が，どの程度の種類の叩き方での手を叩く音が収録されているかは不明である．図 2-9 に波形とスペクトログラムの一例を示す.

clap1, 000.wav～099.wav： 1 回手を叩く音．計 100 音.

clap2, 000.wav～089.wav： 1 回手を叩く音．計 90 音.

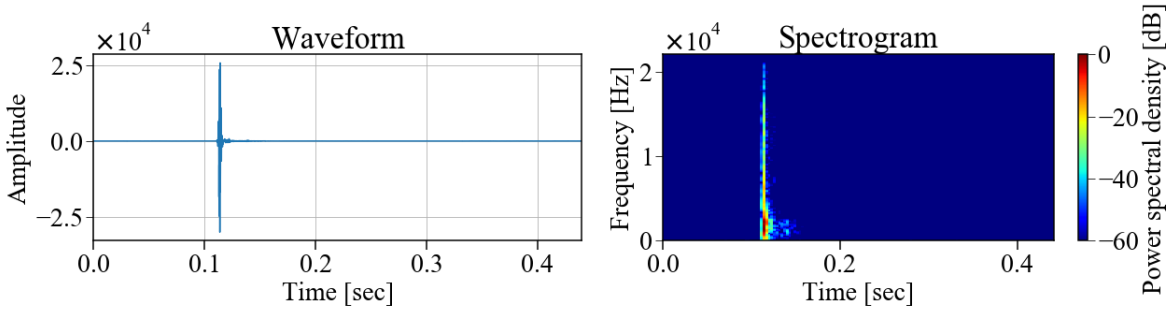


図 2-9 手を叩く音（clap1, 000.wav）の波形（左）とスペクトログラム（右）例.

以上の音のうち，木由来の打撃音，金属由来の打撃音，陶器由来の打撃音，ガラス由来の打撃音の 4 種類は，減衰振動由来の持続成分を含む音である．一方，手を叩く音は非常に減衰の早い突発音であり，プラスチック由来の打撃音は音の持続時間は手を叩く音よりも長いが，持続成分ではない要素が多分に現れている突発音である．プラスチック由来の打撃音と手を叩く音は減衰振動由来の横線状の成分を含まない音の代表として選出している.

2.2.2 非常に良く似たスペクトルを持つ音

打撃によって引き起こされる音の要因としては，音源物体の振動だけでなく，急激な加速度変化や間隙にある空気の放出，物体の変形も考えられる[47]．しかし，平板に球体を衝突させた場合の衝撃音は平板の振動特性に強く依存することが確認されており[48]，本研究で利用している木由来の打撃音と金属由来の打撃音，陶器由来の打撃音，ガラス由来の打撃音の 4 種類の音は音源物体の振動による音であると思われる．

連続体の粘弾性振動で考えると，その振動モードは形状や終端条件，弾性係数，及び内部減衰によって決まり，励起されるモードは初期条件によって決まる．そのため，例えば陶器由来の打撃音である china1 は合計 100 音収録されているものの，音源物体が 5 種類であるためスペクトルが酷似している音の組が存在する可能性がある．

図 2-10 に china1 のうち 000.wav～003.wav のスペクトログラムを，図 2-11 に china1 のうち 000.wav，010.wav，020.wav，030.wav の計 4 個の音のスペクトログラムを示す．図 2-10 の各音のスペクトログラムを見比べると，横線状に現れている持続成分の周波数値が非常によく似ていること，各持続成分のパワーもよく似ていることがわかる．これは 000.wav～009.wav に限らず，010.wav～019.wav，020.wav～029.wav のように，10 個ずつ刻みで非常によく似たスペクトルの音が収録されている．

一方，図 2-11 は 10 個飛ばしで音のスペクトログラムを表示したものである．各音のスペクトログラムを見比べると 5 つある行同士では持続成分の現れる周波数値が異なり，各行内の二つの音では各持続成分のパワーがわずかに異なることがわかる．図 2-11 では，1 行目は 000.wav と 010.wav，2 行目は 020.wav と 030.wav となるように表示しており，各行はそれぞれ陶器 A と B に対応し，周波数値の異なる音となっている．各列での音の違いは恐らく叩く位置や力の変更に起因する．つまり，china1 として収録されている音は，非常によく似た音が 1 組 10 個で 10 種類収録されており，20 個毎に音が大きく変化する．

以上は china1 のみで示しているが，手を叩く音を除く他の音（5 個の音源を使っていない木板や金属板，プラケースに関しても）でも同様に 10 個毎に似た音が収録され，20 個毎に音が大きく変化するようになっている．木板の場合の例も図 2-12 と図 2-13 に示す．

実際に環境音認識を応用する場面において，スペクトルが酷似することは稀であるため，音の再現性が高すぎる識別は意味を持たない．そのため，スペクトルが非常によく似ているかどうかは重要な問題となる．それについては，2.3 にて詳しく述べる．

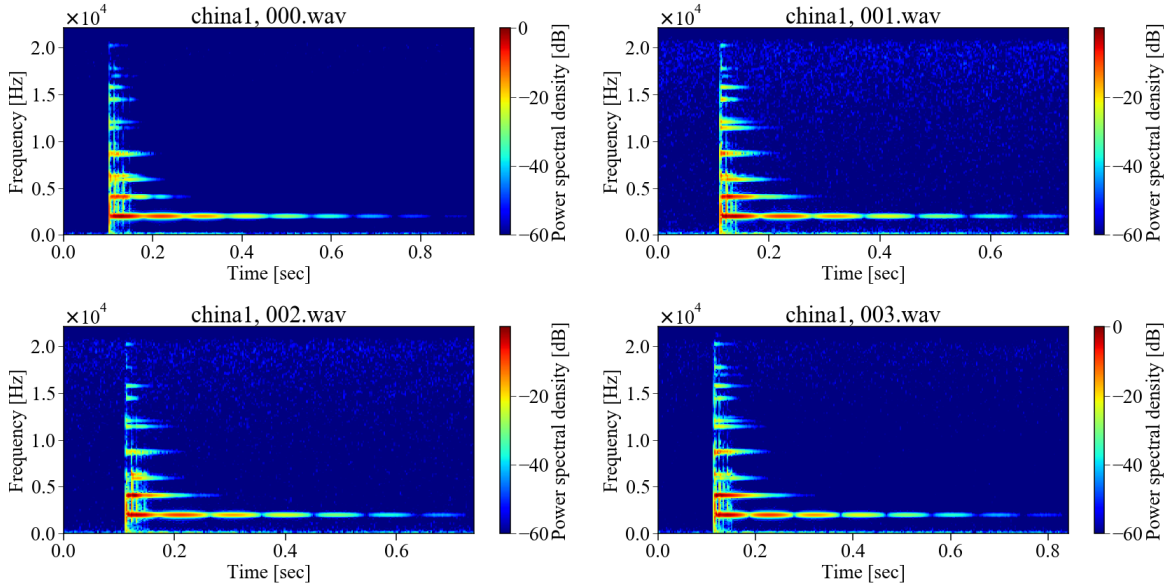


図 2-10 china1, 000.wav～003.wav の計 4 個の音のスペクトログラム．スペクトログラムは窓幅 256, ハミング窓関数での短時間フーリエ変換により計算.

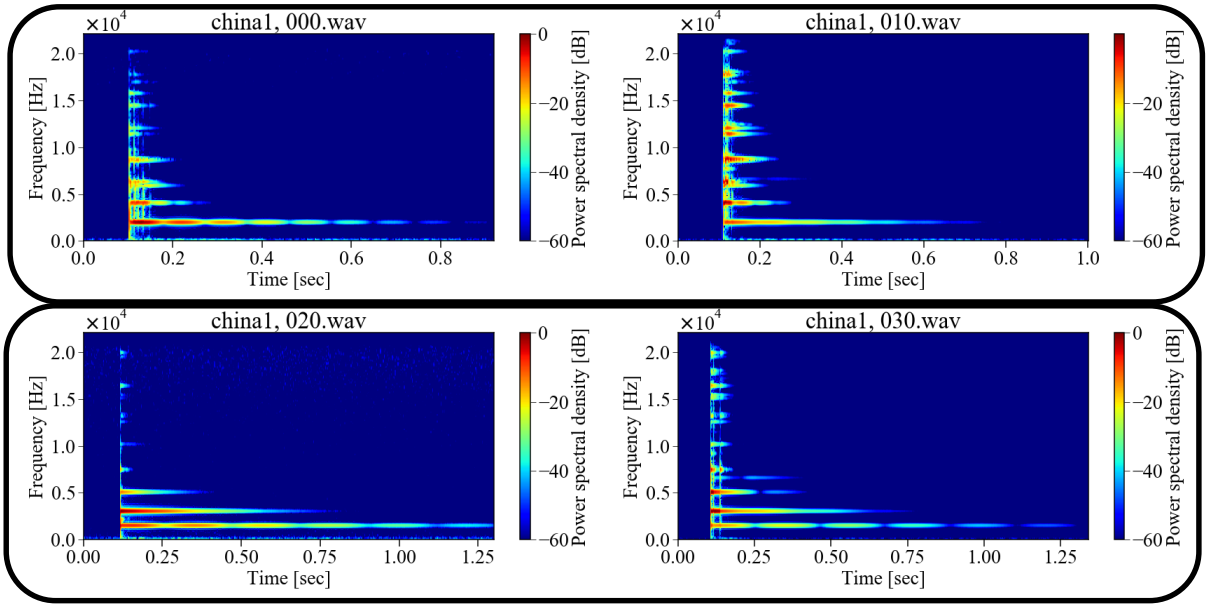


図 2-11 china1, 000.wav, 010.wav, 020.wav, 030.wav の計 4 個の音のスペクトログラム．スペクトログラムは窓幅 256, ハミング窓関数による短時間フーリエ変換で計算．黒枠で囲われた音同士では持続成分の周波数が一致している.

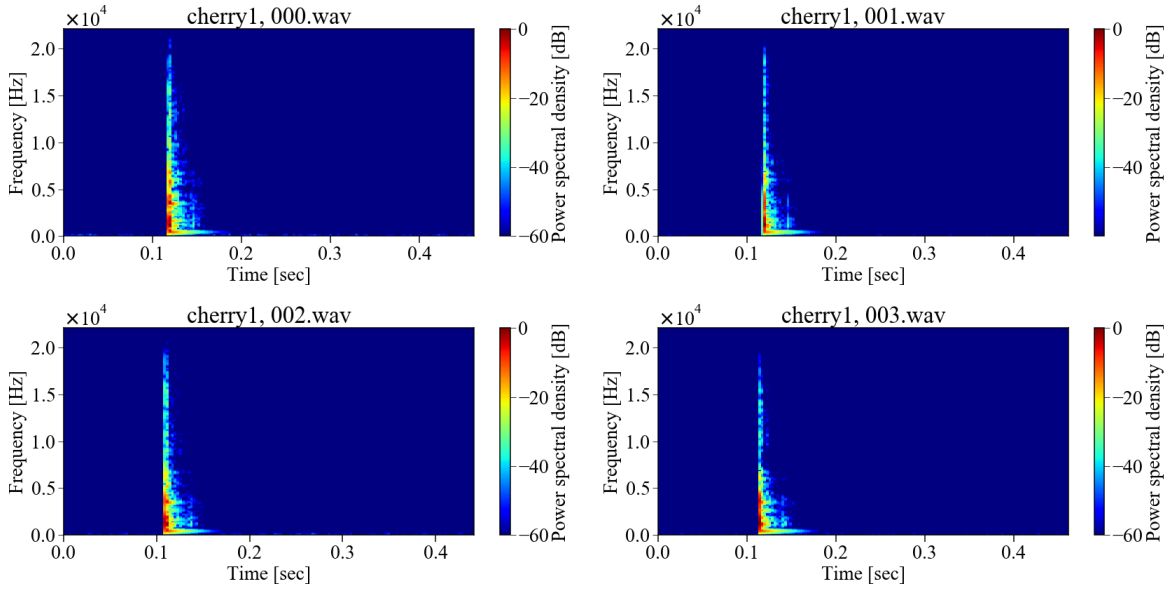


図 2-12 cherry1, 000.wav～003.wav の計 3 個の音のスペクトログラム．スペクトログラムは窓幅 256, ハミング窓関数での短時間フーリエ変換により計算.

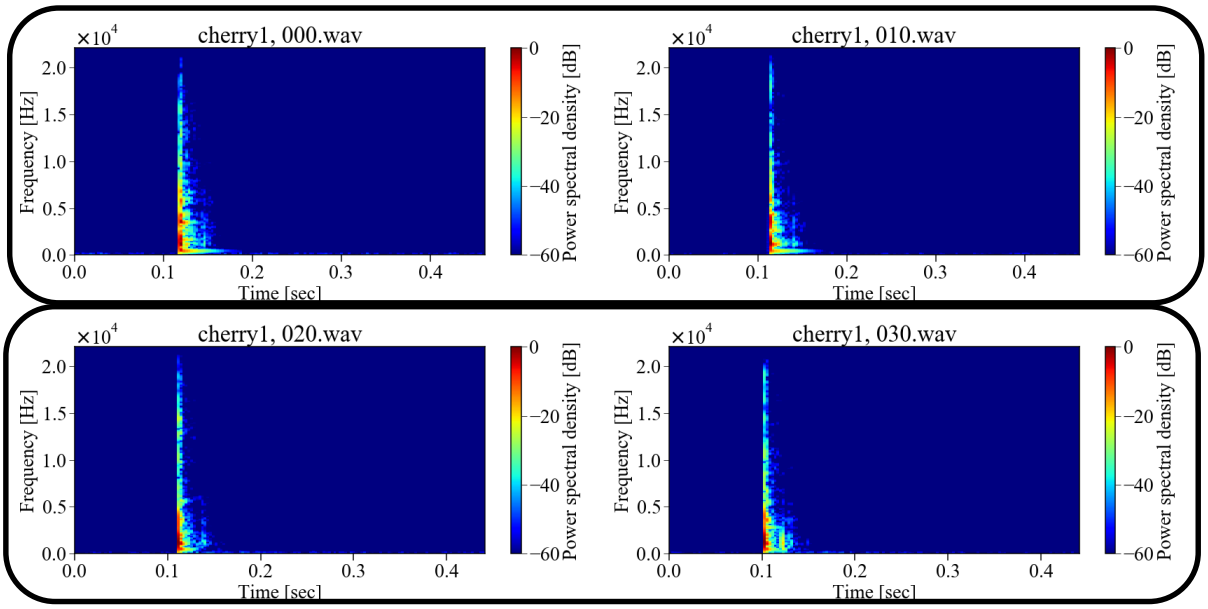


図 2-13 cherry1, 000.wav, 010.wav, 020.wav, 030.wav の計 4 個の音のスペクトログラム．スペクトログラムは窓幅 256, ハミング窓関数による短時間フーリエ変換で計算．黒枠で囲われた音同士ではスペクトログラムが類似している.

2.2.3 雑音の除去

RWCP-SSD は無響室の静かな環境で録音されているが，スペクトログラム上に諸々のノイズが現れるため除去が必要である．

2.2.3.1 ハイパスフィルタによる低周波数に現れる定常雑音の除去

図 2-14 は手を叩く音と金属製ボウルを叩く音のスペクトログラムである．スペクトログラムは定 Q 変換[49]を用いて計算している．定 Q 変換は短時間フーリエ変換の派生として生まれたものであり，周波数毎に計算に利用するフレーム長を変更することで低い周波数では高い周波数分解能を，高い周波数では高い時間分解能を得られる．周波数が倍になると周波数分解能が半分になるため，周波数軸が対数的に量子化される．また，click は縦線とはならず，図 2-14 手を叩く音のスペクトログラムのように裾野が現れる．図 2-14 を見ると，手を叩く音において，音の発生である約 0.1 秒以前から録音停止まで一貫して現れている周波数成分が約 240Hz に確認できる．この約 240Hz での定常的な雑音は図 2-14 の金属製ボウルを叩く音においても確認できる．約 240Hz 以外にも手を叩く音と金属製ボウルを叩く音の両方で共通して現れる定常的な雑音が少なくとも約 200Hz と約 150Hz で確認できる．

本研究では，図 2-15 に示す周波数特性を持つハイパスフィルタ（FIR フィルタ，512 点）を利用してこの雑音を除去した．利用した音によっては最低周波数が 400 Hz 前後であるため，一部音では音源物体に由来する横線状の持続成分も抑えられるが，雑音の除去を優先している．

2.2.3.2 全体にわたって現れる背景雑音の除去

図 2-14 の手を叩く音のスペクトログラムには，上述の定常的な雑音以外にもホワイトノイズのようなごま塩状の背景雑音も現れている．図 2-16 はガラスコップを叩く音 2 個に対して上述のハイパスフィルタを適用した後に，短時間フーリエ変換と定 Q 変換でスペクトログラムを計算し，パワーを最大以下 50 dB までを表示したものである．図 2-16 左側の二つのスペクトログラムにおいて，本来であれば無音であるはずの領域にノイズが現れている様子が確認できる．このノイズはパワーを最大以下 50 dB ではなく 50 よりも小さい値にしていけばいずれ他の無音部分と同様に最低値となるが，そうするとパワーの弱い持続成分までも消されてしまう．パワーを最大以下 40 dB，45 dB，50 dB，55 dB，60 dB と変化させた際にノイズが現れる音の数は 50 dB 以下で一つのみ（cup1, 013.wav）であったため，本研究では 50 dB を採用し，以後の計算では除去が必要な場合にはパワー最大以下 50 dB までを利用することでノイズの除去を行なった．雑音が残る 1 個の音（cup1, 013.wav）は利用する音から除外している．

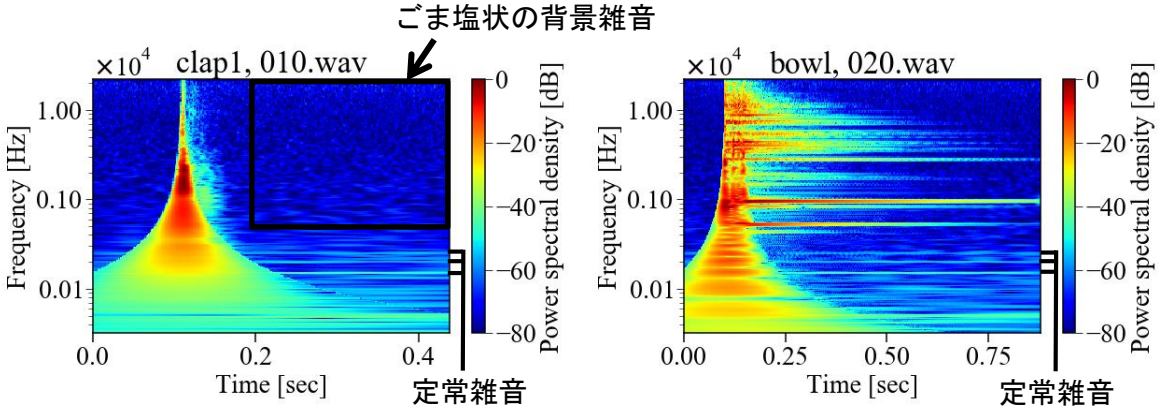


図 2-14 手を叩く音のスペクトログラム（左）と金属製ボウルを叩く音のスペクトログラム（右）．スペクトログラムは定 Q 変換で計算.

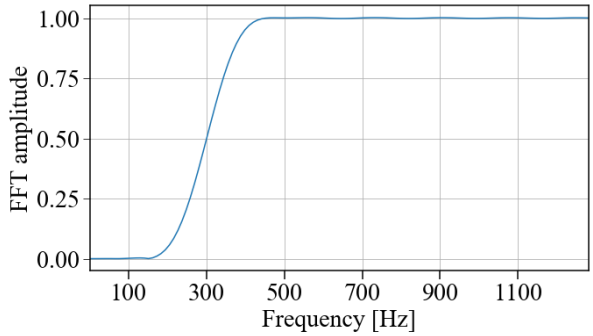


図 2-15 ノイズ除去に利用したハイパスフィルタの周波数特性．1300 Hz 以下を表示.

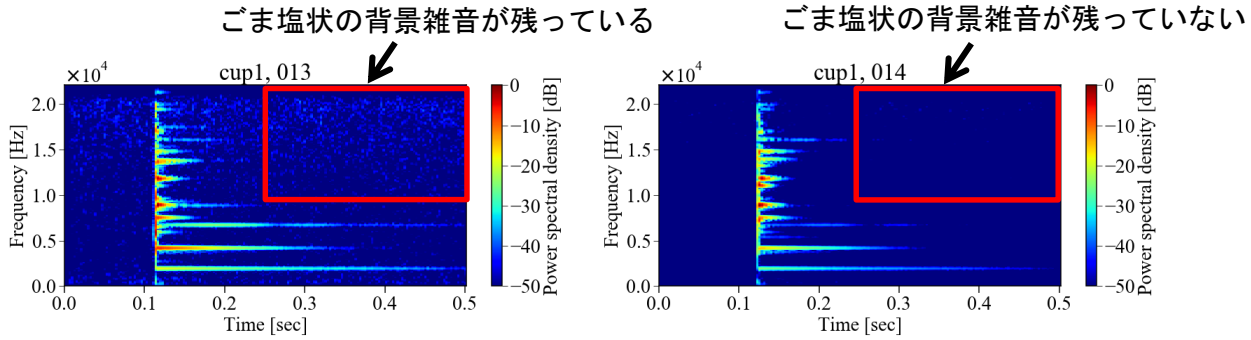


図 2-16 ガラスコップを叩く音（cup1, 013.wav）の短時間フーリエ変換によるスペクトログラム（左）とガラスコップを叩く音(cup1, 014.wav)の短時間フーリエ変換によるスペクトログラム（右）.

2.2.3.3 チャタリングが現れる音の分離

本論文では，一度叩くだけで複数の突発音が発生することをチャタリングと呼び表す．先述の約 250Hz 以下の定常雑音とごま塩状の雑音は録音したい音とは別の要因によって現れている雑音だが，それとは別に図 2-17 に示す音のようにチャタリングが発生していると，録音したい音源から発せられた音ではあるが，特徴量の計算方法によっては計算結果に影響を与えるノイズとなってしまう可能性がある．図 2-17 の音では不連続な立ち上がりが 3 個確認できる．そこで，チャタリングが存在する音としない音の分離を行い，4 章にて利用している．

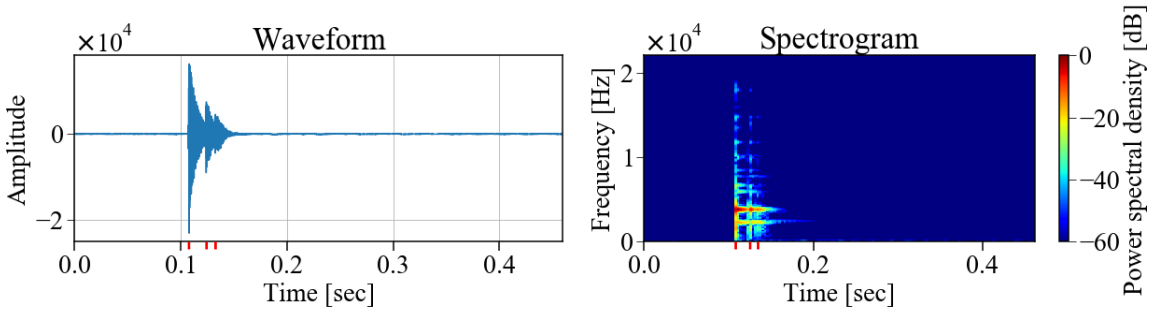


図 2-17 陶器を叩く音（china2, 060.wav）の波形（左）とスペクトログラム（右）．時間軸上赤い印を付けている時刻でチャタリングが 3 回現れている．

チャタリングの有無の分離は何らかの計算した値をもとに客観的に行ったのではなく，目視により行った．また，木板を叩く音ではチャタリングを全く含まない音は収録されていなかったため，図 2-18 のようにチャタリングがはっきりと現れる音はチャタリングを含む音として，図 2-19 のようにチャタリングがはっきりとは現れていない音はチャタリングを含まない音として分類した．

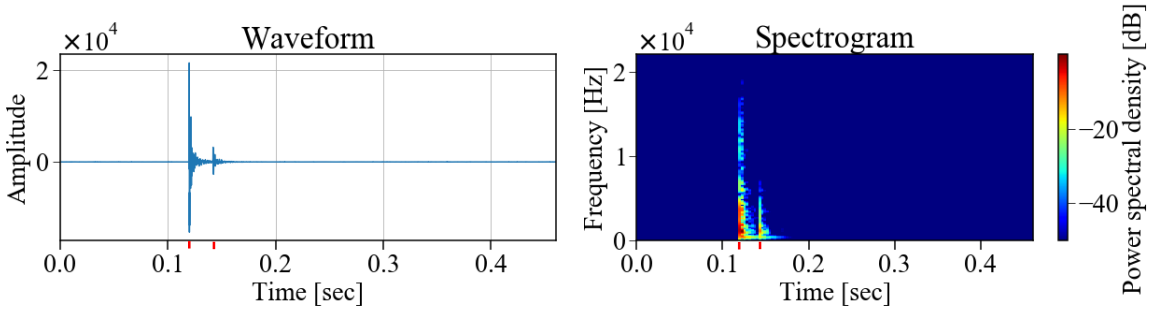


図 2-18 木板を叩く音（cherry1, 008.wav）の波形（左）とスペクトログラム（右）．時間軸上赤い印を付けている時刻でチャタリングが計 2 回現れている．

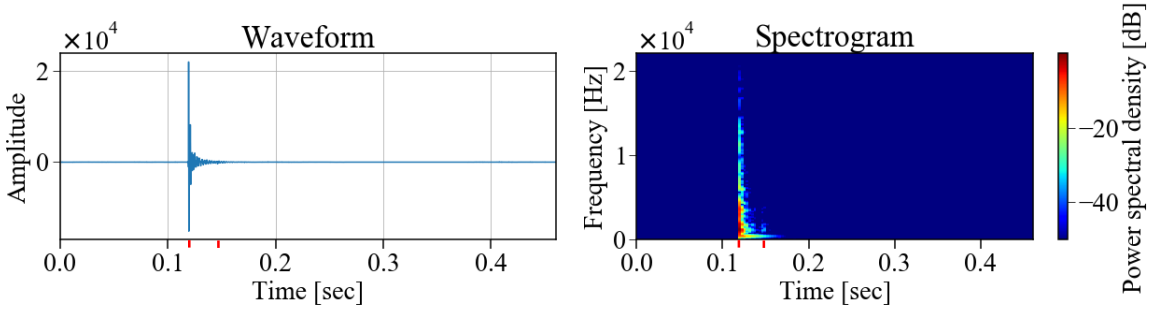


図 2-19 木板を叩く音（cherry1, 009.wav）の波形（左）とスペクトログラム（右）．時間軸上赤い印を付けている時刻でチャタリングが計 2 回現れている．

2.3 学習データとテストデータの分割

2.2.2 のとおり， RWCP-SSD には非常に良く似たスペクトルを持つ音が含まれる． RWCP-SSD を利用した研究では，例えば china1 は china1 に識別し，china2 は china2 に識別する，というように陶器由来の音同士でも複数のラベルに分割し，データを学習データとテストデータに分割する際にランダム割り振りを行なっている研究が多く，このような問題設定においては 98%以上の識別精度が得られている[50]–[53].

一方で, 上記のような特定度の非常に高い問題設定ではなく, 比較的特定度の低い問題設定で RWCP-SSD を利用している研究は少ない．Wang ら[44]は china1 と china2 と china3 のように陶器を叩く音であるのに複数の異なるラベルに分かれているのは不自然として，纏めて一つのラベルを付け，105 種類ある RWCP-SSD の音は 62 種類まで纏められている．これにより音の特定度は少し下がり，識別精度も 87.18%と低くなっている．しかし，学習データとテストデータを全データからランダムに分割しているため，スペクトルの非常によく似た音が 10 個ずつ含まれるというデータセットの仕様により学習データとテストデータの間に高い確率で非常に良く似たスペクトルを持つ音の組が生まれため，依然として特定度の高い識別となっている．

そこで, 以下の 3 通りで分割方法を変化させることで, 学習データと試験データの間での音の再現性，つまり音の特定度を操作した実験を行なった．

1. 十の位で 10 分割

   2.2.1 節で述べたとおり，RWCP-SSD に収録されている音には通し番号が振られている．通し番号をもとに 10 の位が等しいものを一つのグループとして 10 分割交差検証を行う．例えば，010.wav, 011.wav, 012.wav, 013.wav, 014.wav, 015.wav, 016.wav, 017.wav, 018.wav, 019.wav の十個が一つのグループになる．このとき，各グループにはスペクトルの非常によく似た音のみが含まれるようになり, 手を叩く音以外では他のグループとの間でスペクトルが非常に良く似ている組みは存在しない．

2. 一の位で 10 分割

   通し番号の 1 の位が等しいものを一つのグループとして 10 分割交差検証を行う．例えば，000.wav, 010.wav, 020.wav, 030.wav, 040.wav, 050.wav, 060.wav, 070.wav, 080.wav, 090.wav の十個が一つのグループになる．このとき，スペクトルの非常に良く似た 10 個の音は 10 個全ての分割に散らばることになる．

3. 前から順に 5 分割

   通し番号をもとに，000.wav~019.wav，020.wav～039.wav，040.wav～059.wav，060.wav～079.wav，080.wav～099.wav という風に連番で 5 分割する．このとき，スペクトルが

非常に良く似た音 10 個同士だけでなく 2.2.2 節で述べた図 2-11 の 000.wav と 010.wav のようにスペクトルに多少違いはあるけれど似ている音同士も全て一つのグループに属することとなり，各分割間での音の再現性は 3 種類の分割の中で最も低くなる．

RWCP-SSD データセットを利用した特定度の低い識別を行なっている Wang らの手法[44]での識別精度を上記 3 種類の分割方法で評価した．識別のパラメータは文献内と同様に小規模なデータセットを用意して設定した．結果は以下の表 1 のようになった．RWCP-SSD に含まれる全ての音を利用した場合の識別精度と，2.2.1 節で示した突発音のみを利用した場合の識別精度を示している．ラベル付けは Wang らと同様，陶器を叩く音である china1 と china2 と china3 は全て同じ china という音として識別する．識別結果を表 2-1 に示す．

1 の位で 10 分割した場合の精度が最も高く，十の位で 10 分割した場合が次に高い．前から順に 5 分割した場合でもある程度識別出来ているが，他と比べ精度が大きく下がっている．この順番は音の特定度の高さに合致している．

このように，分類に求められる音の特定度が異なる場合，同じ手法と音でも結果が異なる場合があることに注意が必要である．

表 2-1　RWCP に収録される音を複数の分割方法で分割した際の識別精度

分割方法	全音での識別精度	突発音のみでの識別精度
十の位で 10 分割	80.2 %	86.9 %
1 の位で 10 分割	88.1 %	93.5 %
前から順に 5 分割	73.4 %	70.0 %

2.4 本章のまとめ

本章では，以後の章で利用するデータの詳細説明と雑音除去方法，学習データとテストデータの分割について述べた．

利用するデータは RWCP-SSD データベースに含まれる突発音だが，その中にはスペクトルの非常に似た音も含まれている．詳しくは，スペクトルの非常に似た音が 10 音ずつ，それら 10 音ずつの組の 2 組で構成されるスペクトルの似た 20 音ずつの類似音が現れる．

そして，スペクトルの似た音同士が学習データとテストデータの双方に含まれる場合では識別結果が変化してしまうため，その分割方法についても述べた．

また，RWCP-SSD データベースは無響室の静かな環境で録音されたものだが，全く雑音が含まれないわけではなく，低周波数の定常雑音，ホワイトノイズのような背景雑音，一度の打撃で複数の衝突音が発生してしまうチャタリングの 3 種類の雑音が確認できる．これらの雑音について，低周波数での定常雑音はハイパスフィルタをかけることで，ホワイトノイズのような背景雑音はパワー最大以下 50 dB までを扱うことで除去できる．チャタリングについては，チャタリングのある音とチャタリングの無い音の 2 グループに分割することで，以後の章でチャタリングによる特徴量の変化を確認できるようにしている．

第3章　特徴量の計算方法

3.1 はじめに

本章では，本研究で扱う特徴量の計算方法を示す．このうち，スペクトル重心と持続成分の顕著さ（平均と分散）は新規に提案したもの，tan ϕ，波形の減衰，立ち上がり時間，減衰時間，Spectral bandwidth，ラフネス，Spectral rolloff，スペクトルの滑らかさ，Zero crossing rate，波形のなめらかさ，チャタリングの数，チャタリングの時間間隔，持続成分の明確度合いの計 13 種類は先行研究にて提案されている特徴量である[38]-[40].

図 3-1 のように，部分的に周波数を上げる加工を行い試聴したところ，聴き比べれば異なる音に聞こえるものの違う種類の音には聞こえず，ガラスを叩く音として聞こえた．そのため，調波構造のようなスペクトルの構造は存在しないと推測し，スペクトル重心と持続成分の顕著さは，スペクトルの概形を捉えるための計算を行っている．

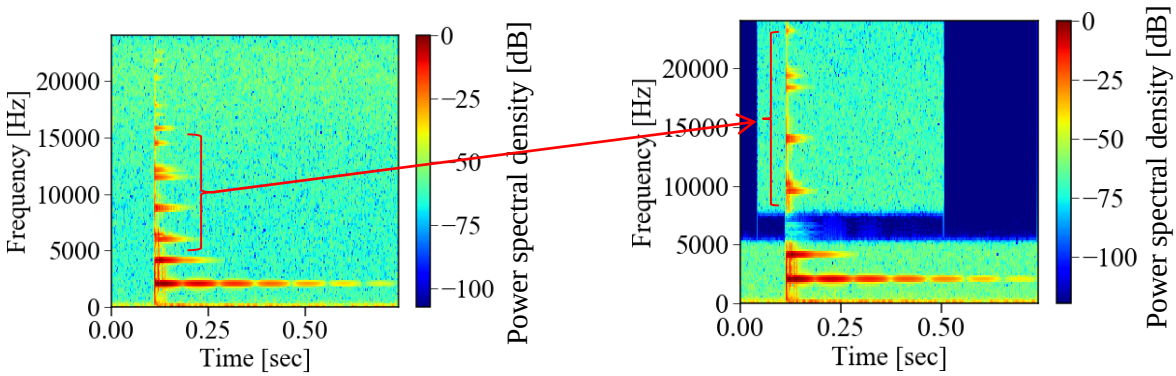


図 3-1 ガラスを叩く音（左）と加工後の音（右）のスペクトログラム.

本章では適宜音の詳細を表示して特徴量の分析を行っているが，cherry1, china1, metal05, cup1, bottle1, case1, clap1 のみを利用し，他の突発音については 1 章でのスペクトルが似ているかどうかの調査以外では確認しないようにしている．

短時間フーリエ変換について，短時間フーリエ変換ではスペクトルの時間変化を捉えるため，図 3-2 のように信号全体を複数のフレームに分けてからそれぞれのフレームに関して離散フーリエ変換を行う．この分割について，フレームの長さを窓幅，フレーム同士の重なり幅をオーバーラップと呼ぶ．また，フレームの端の不連続さが計算結果に与える影響を抑えるために窓関数をかける．

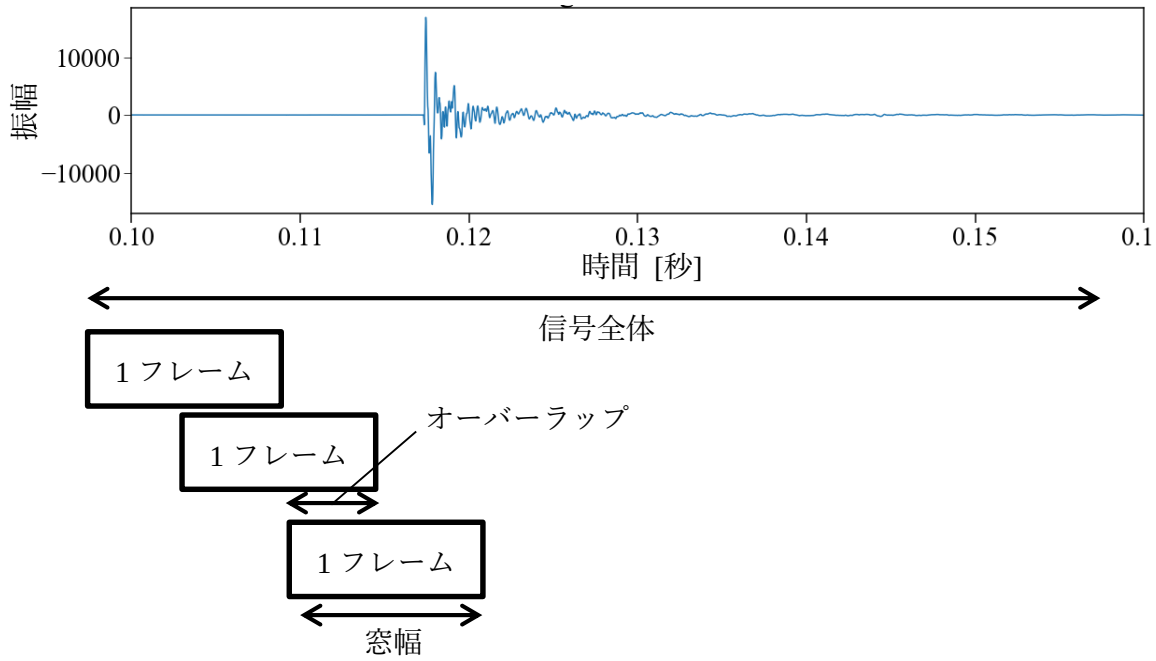


図 3-2 短時間フーリエ変換での信号の複数のフレームへの分割の模式図.

3.2 持続成分の顕著さ

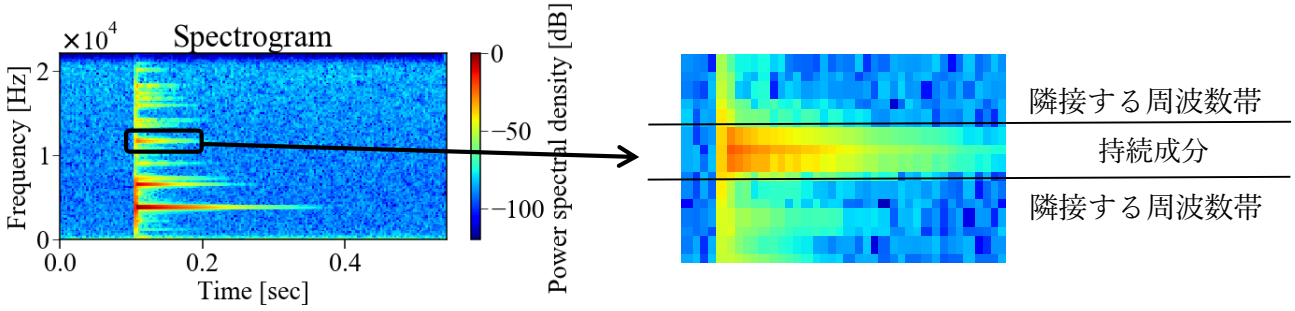


図 3-3 ガラス瓶を叩く音（bottle1, 000.wav）のスペクトログラム（左）と拡大図（右）.

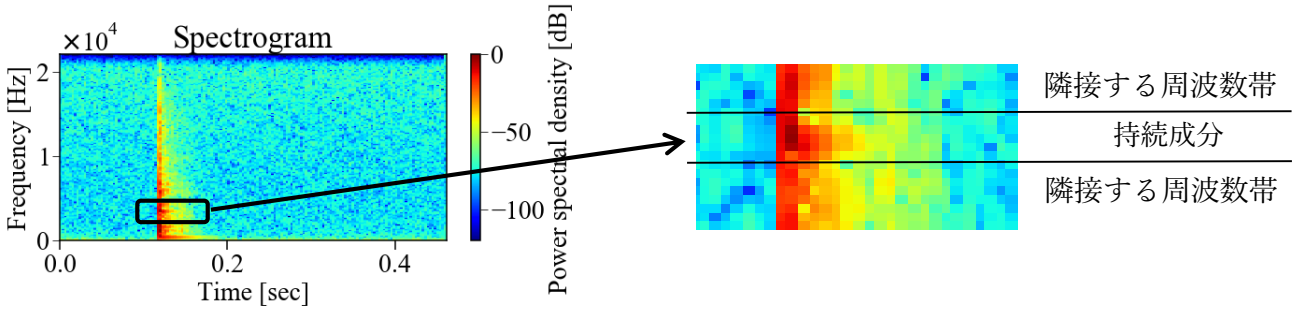


図 3-4 木板を叩く音（cherry1, 000.wav）のスペクトログラム（左）と拡大図（右）.

図 3-3 にガラス瓶を叩く音のスペクトログラムを示す．スペクトログラム中に，複数の持続成分が現れていることが確認できる（例えば黒枠内）． 図 3-4 は木板を叩く音のスペクトログラムである．ガラスコップを叩く音のスペクトログラムに比べ，持続成分が非常に醜くなっているが，この音でも複数の周波数にて，パワーの強い部分が時間方向に連なっていることが確認できる（例えば黒枠内）．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おり，持続成分は音色に影響を与える．

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付ける要素であり，材質に関係する要素である．

複数の音の持続成分の観察により，次の要素が音を見分ける特徴となるのではないかと考えた．ガラス瓶を叩く音（図 3-3）では持続成分である横線は隣接する周波数帯とのパワーの差が大きく，持続成分がはっきりと現れている．一方で，木板を叩く音（図 3-4）では隣接する周波数帯とのパワーの差が小さくなっており，持続成分と隣接周波数帯との境界ははっきりしていない．

これは減衰の違いによるものだと思われるが，この違いを特徴量とするため，以下の手順で持続成分の顕著さを計算する．

1. 短時間フーリエ変換（窓幅：256 サンプル，オーバーラップ：128 サンプル，窓関数：ハミング窓）によりスペクトログラムを計算し，その後デシベルに変換する．その際，スペクトログラム中のパワー最大となる所を 120 dB とし，0 dB 以下は 0 dB に丸める．

   短時間フーリエ変換で得られたスペクトログラムを$S(\omega,t)$とすると，この変換は式(3-1)と式(3-2)で表される．ここで，ωとtはそれぞれ周波数と時間を表す．

$$S_{\mathrm{dB}} = 10\log_{10}\left(\frac{S(\omega,t)}{\max\limits_{\omega,t} S(\omega,t)}\right) + 120 \tag{3-1}$$

$$S'_{\mathrm{dB}}(\omega,t) = \max(0, S_{\mathrm{dB}}) \tag{3-2}$$

2. 図 3-5 に示すフィルタを畳み込む（式(3-3)）．フィルタの横幅は 17 (約 52 ミリ秒)とした．フィルタは平均が 0 であるため，ホワイトノイズのような局所的でないオフセットの影響は抑えられている．このフィルタは横線状の部分を強調するためのものである．

$$P(\omega,t) = \sum_{i=-2}^{2}\sum_{j=-8}^{8} S'_{\mathrm{dB}}(\omega-i, t-j)h(i,j) \tag{3-3}$$

$$h(x,y) = \begin{cases} 0 & if\ \ y \geq 3 \\ -0.15 & if\ \ y = 2 \\ -0.35 & if\ \ y = 1 \\ 1 & if\ \ y = 0 \\ -0.35 & if\ \ y = -1 \\ -0.15 & if\ \ y = -2 \\ 0 & if\ \ y \leq -3 \end{cases} \tag{3-4}$$

ここで，$h(x,y)$は図 3-5 のフィルタを表す．

図 3-5 横線強調フィルタ．各行の値は上から順に 0，-0.15，-0.35，1，-0.35，-0.15，0.

3. フィルタを畳み込んだ結果（$P(\omega,t)$）が正の値を取っている部分が横線状になっている部分に相当する．しかし，ホワイトノイズに対して計算した場合でも正の値を持つ部分は現れるため，背景雑音に対応する部分は除去が必要である．そのため，式(3-5)の方法により, 対応する時間, 周波数でのパワーがスペクトログラム全体での最大以下 50dB であり，かつ$P(\omega,t)$が正の値になっている部分の抽出を行った．

$$P'(\omega,t)=\begin{cases}\max\left(0,P(\omega,t)\right) & if\ \ S'_{\mathrm{dB}}(\omega,\mathrm{t})>70\\ 0 & otherwise\end{cases} \tag{3-5}$$

4. スペクトログラムでのパワーが大きい持続成分だけでなく小さい持続成分も利用するため，対応する時間，周波数でのスペクトログラムのパワーで補正を行う．

$$P''(\omega,t)=\frac{P'(\omega,t)}{S'_{\mathrm{dB}}(\omega,t)} \tag{3-6}$$

以上の手順で，各周波数各フレームでの持続成分の顕著さを計算した．一例として，ガラス瓶を叩く音での結果を図 3-6 に示す．スペクトログラム中で横線を形成している持続成分が抽出されていることが確認できる．

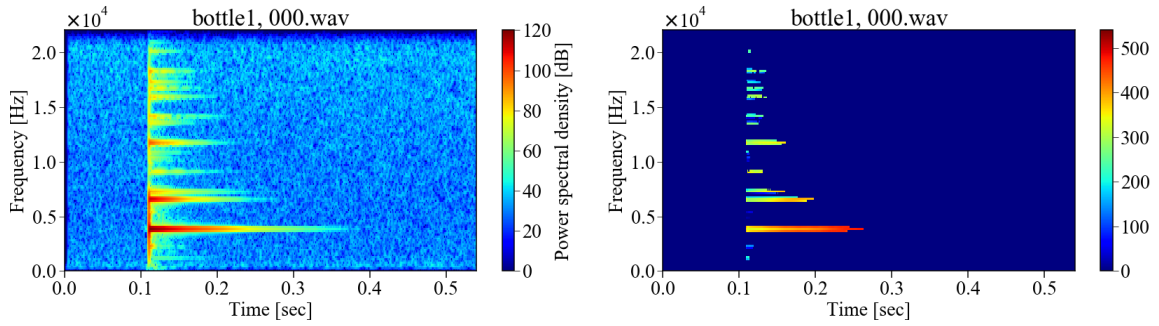


図 3-6 ガラス瓶を叩く音（bottle1, 000.wav）のスペクトログラム（左）と$P''(\omega,t)$（右）．

次に，低次元の特徴量とするため，パワーが最大となるフレームにおいて，平均と分散を計算して特徴量とした．ここで平均と分散を利用しているのは，スペクトルに調波構造のような特定の構造が現れず，持続成分が具体的にどの周波数に現れているかは重要ではないだろうと考えたためである．平均と分散を計算する際，正の値のみを利用して計算している．

5. 音のパワーが最大となるフレームに対応する時刻を$t_{\max}$と置く．$t_{\max}$は短時間フーリエ変換（窓幅：256 サンプル，オーバーラップ：255 サンプル，窓関数：ハミング窓）でのパワーが最大となるフレームに対応する．パワーが最大となるフレームでの分割に揃

うように，1番目の手順での短時間フーリエ変換の分割方法を調整している．

6. 式（3-6）で得られた$P''(\omega, t)$から$t = t_{\max}$での正の値のみを集め，P_+とする．

$$P_+ = \{P''(\omega_i, t_{\max}) \,|\, P''(\omega_i, t_{\max}) \neq 0, i = 1,2, \ldots, N\} \tag{3-7}$$

ここで，Nは周波数便の数を表す．

7. P_+の平均μ_{P_+}と分散σ_{P_+}を計算する．

図 3-7～図 3-12 に各種突発音のスペクトログラムと対応する$P'(\omega_i, t_{\max})$を示す．持続成分が表れている周波数で$P'(\omega_i, t_{\max})$が大きくなっていること，パワーの小さい持続成分でもピークを形成していることが確認できる．また，持続成分が表れない手を叩く音や，持続成分が表れているのかはっきりしないプラケースを叩く音では全体的に値が小さく，持続成分がはっきりと表れているガラスを叩く音や，陶器を叩く音では全体的に値が大きくなっている．そのため，平均と分散を特徴量として用いる．

ただし，$P'(\omega_i, t_{\max})$でピークを形成している部分が全て持続成分となっているわけではなく，特に最も低い周波数でのピークは，恐らく雑音除去のためのバンドパスフィルタに由来するものである．そして個々のピークの値が物理量として明確な意味を持っているわけではない．持続成分がはっきりと現れていない木板を叩く音では，持続成分というよりも局所的にパワーが大きくなっている部分にて，他の周波数帯よりも大きな値を取っている．

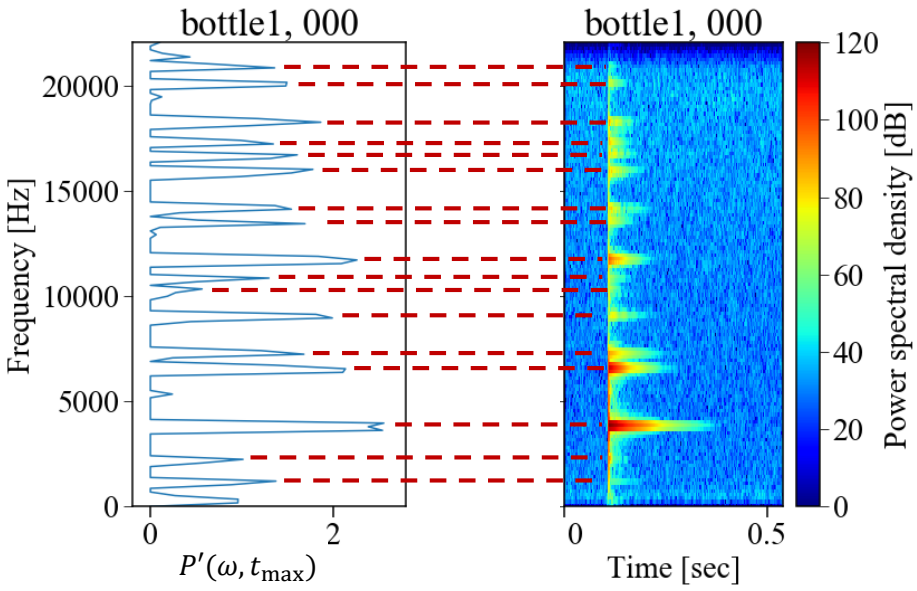


図 3-7 ガラス瓶を叩く音（bottle1, 000.wav）の立ち上がりでの持続成分の顕著さ$P''(\omega, t_{\max})$（左）とスペクトログラム（右）．$\mu_{P_+} = 0.29$，$\sigma_{P_+} = 0.076$である．図中破線は$P'(\omega, t_{\max})$のピークと持続成分の対応を示している．

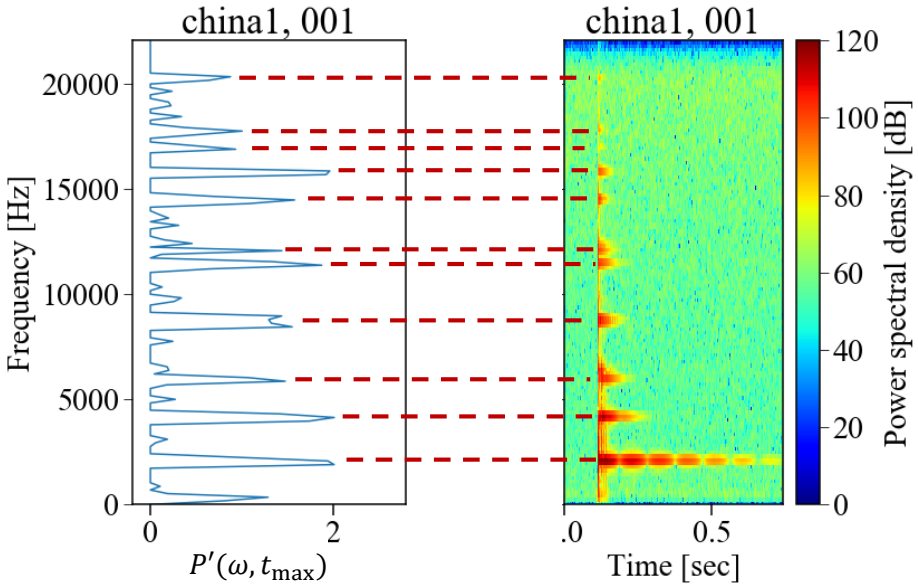


図 3-8 陶器を叩く音（china1, 001.wav）の立ち上がりでの持続成分の顕著さ$P'(\omega, t_{\max})$（左）とスペクトログラム（右）．$\mu_{P_+} = 0.33$，$\sigma_{P_+} = 0.049$である．図中破線は$P'(\omega, t_{\max})$のピークと持続成分の対応を示している．

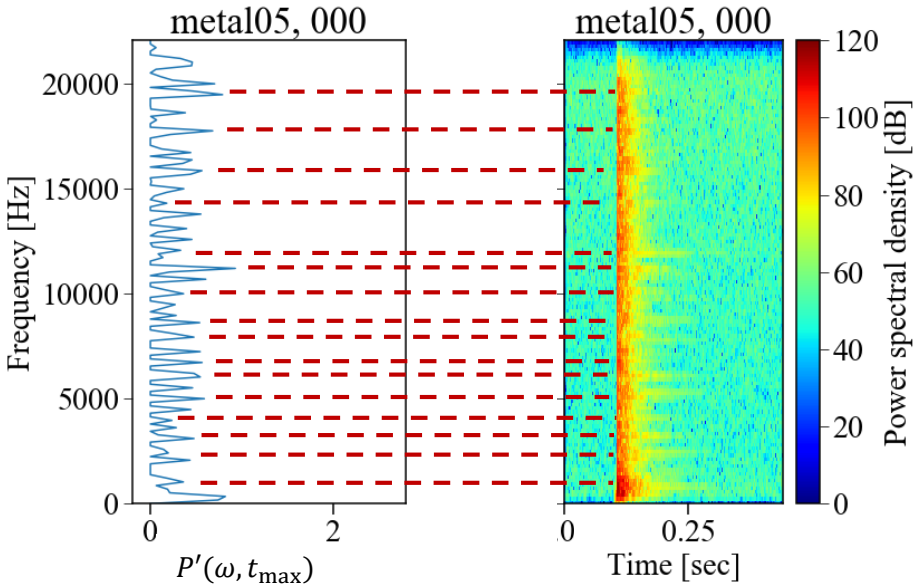


図 3-9 金属板を叩く音（metal05, 000.wav）の立ち上がりでの持続成分の顕著さ$P'(\omega, t_{\max})$（左）とスペクトログラム（右）．$\mu_{P_+} = 0.80$，$\sigma_{P_+} = 0.42$である．図中破線は$P'(\omega, t_{\max})$のピークと持続成分の対応を示している．

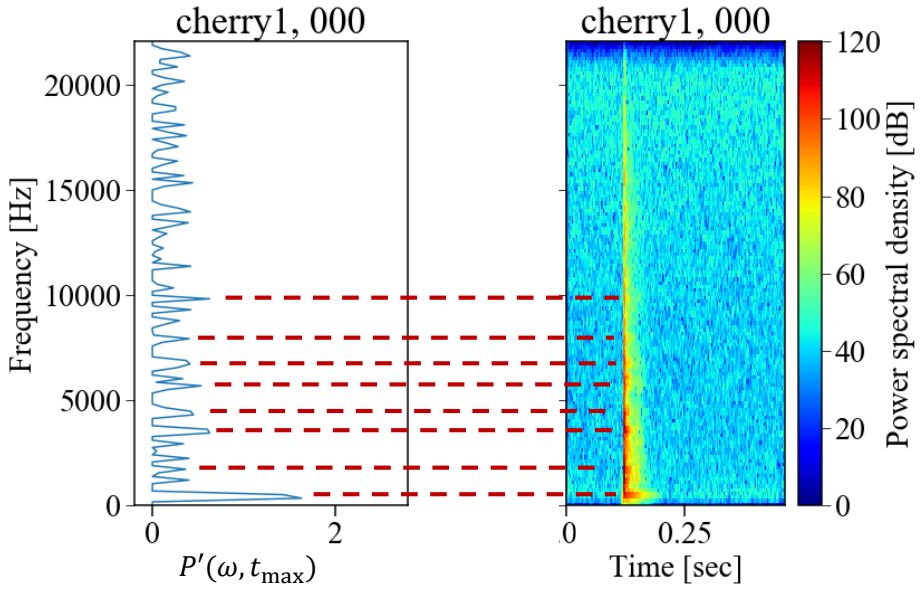


図 3-10 木板を叩く音（cherry1, 000.wav）の立ち上がりでの持続成分の顕著さ$P'(\omega, t_{\max})$（左）とスペクトログラム（右）．$\mu_{P_+} = 1.1$，$\sigma_{P_+} = 0.53$である．図中破線は$P'(\omega, t_{\max})$のピークと持続成分及びパワーのピークとの対応を示している．

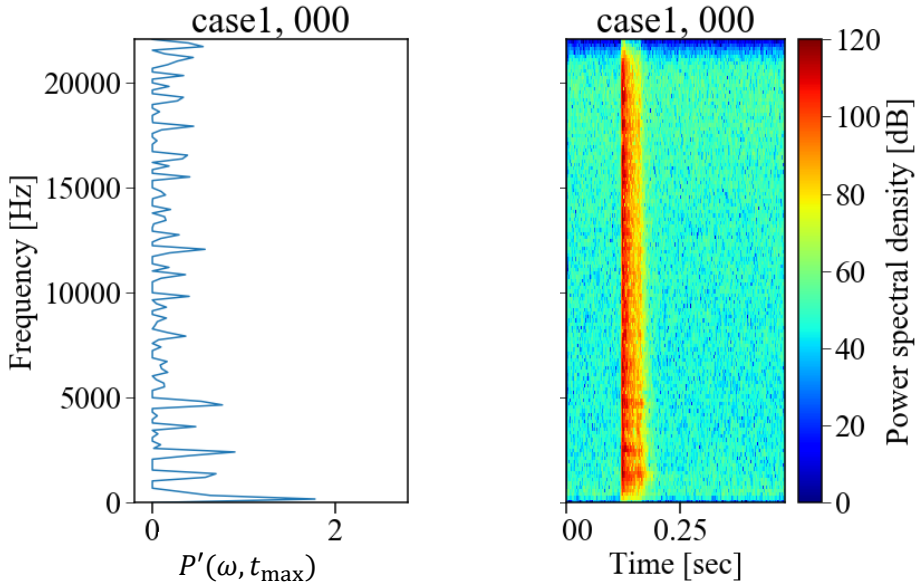


図 3-11 プラケースを叩く音（case1, 000.wav）の立ち上がりでの持続成分の顕著さ$P'(\omega, t_{\max})$（左）とスペクトログラム（右）．$\mu_{P_+} = 0.26$，$\sigma_{P_+} = 0.079$である．

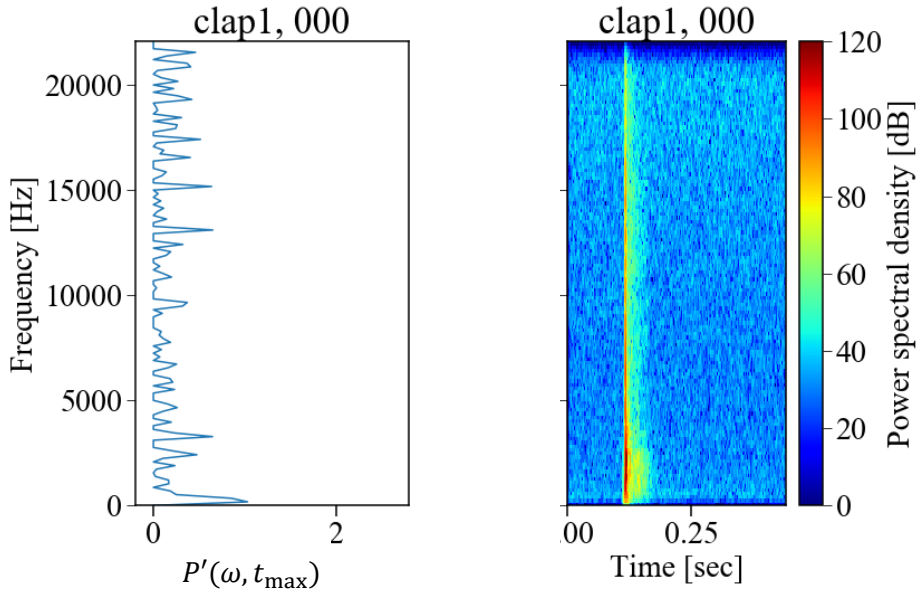


図 3-12 手を叩く音（clap1, 000.wav）の立ち上がりでの持続成分の顕著さ$P'(\omega, t_{\max})$（左）とスペクトログラム（右）．$\mu_{P_+} = 0.23$，$\sigma_{P_+} = 0.039$である．

3.3 スペクトル重心

スペクトル重心は良く知られた音響特徴量であり，人が聴いた感覚としての音色空間との相関も調べられている[54]．そして，基本周波数や調波構造のように，特定の周波数に持続成分が現れているか否かを捉えるものではなく，スペクトル全体の概形を捉えるものである．本節ではスペクトル重心を，変化が非常に早い突発音に適用し特徴量として計算する．スペクトル重心(spectral centroid)は以下の式(3-8)で計算される．

$$\frac{\sum_{i=1}^{N} \omega_i |s(\omega_i)|}{\sum_{i=1}^{N} |s(\omega_i)|} \tag{3-8}$$

ここで，$\omega_i \quad (i = 1,2,\dots,N)$は周波数，$s(\omega)$は周波数$\omega$に対応するスペクトルである．スペクトルの計算は，離散フーリエ変換で計算した場合と定 Q 変換で計算した場合の 2 種類を利用している．離散フーリエ変換は式（3-9）で，定 Q 変換は式（3-10）で表される．それぞれ線形周波数でのスペクトル重心と対数周波数でのスペクトル重心と呼ぶ.

$$s(\omega) = \sum_{k=1}^{N} w(k)x(k) \exp\left(-j\frac{2\pi\omega}{N}\right) \tag{3-9}$$

$$s(\omega) = \frac{1}{N(\omega)} \sum_{k=1}^{N(\omega)} w(k,\omega)\, x(k) \exp\left(-j\frac{2\pi k Q}{N(\omega)}\right) \tag{3-10}$$

$$N(\omega) = Q\frac{f_s}{\omega} \tag{3-11}$$

ここで，N は窓幅，w は窓関数を表す．定 Q 変換では周波数に応じて窓幅を変更するため，N と w は周波数に依る．また，Q は予め決めた定数，f_sはサンプリング周波数を表す.

スペクトルにホワイトノイズのような背景雑音が含まれると値が変化するため，本研究ではスペクトログラムの計算の後パワー最大以下 50 dB を利用した．これは，式(3-1)における 120 を 50 に置き換えたのちに式(3-2)に代入することで計算できる.

窓幅 256 サンプル，オーバーラップ 255 サンプルとして短時間フーリエ変換で計算されたスペクトログラムを基に，ガラスコップを叩く音のスペクトログラムとスペクトル重心の時間変化を計算した結果を図 3-13 に示す. この図のように, 突発音のスペクトル重心は時々刻々と変化し，最終的には最も長く続く持続成分の周波数値となる.

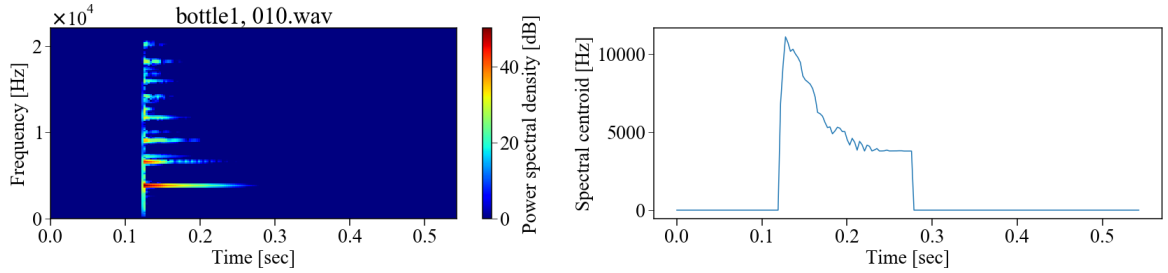


図 3-13 ガラス瓶を叩く音（bottle1, 010.wav）のスペクトログラム（左）と線形周波数でのスペクトル重心の時間変化（右）.

次に，変化するスペクトル重心から何を特徴量として抽出するかについて述べる.

時間変化の様子について，ガラス瓶を叩く音 3 個（bottle2, 030.wav, 031.wav, 032.wav）及び金属製ボウルを叩く音 3 個（bowl, 080.wav, 081.wav, 082.wav）をまとめて表示したものをそれぞれ図 3-14 と図 3-15 に示す．これらはスペクトルの非常によく似た音同士である．また，ガラス瓶を叩く音 5 個（bowl, 000.wav, 010.wav, 0w0.wav, 030.wav, 040.wav）及び金属製ボウルを叩く音 5 個（bowl, 000.wav, 010.wav, 020.wav, 030.wav, 040.wav）をまとめて表示したものをそれぞれ図 3-16 と図 3-17 に示す．これらは，音の種類は同じだがスペクトルの異なる音同士である.

図 3-14 から，線形周波数でのスペクトル重心の値はパワー最大となる時刻から 0.1 秒以上の間，近い値を取っていることが確認できる．図 3-15 では，一つの音（080.wav）のみパワー最大となる時間の後すぐに値がずれていっているが，残り二つでは 0.1 秒を過ぎても近い値を取っている．これらは，スペクトルの非常によく似た音同士である．一方，図 3-16 と図 3-17 から，同じ種類の音ではあるけれどスペクトルが似通っていない音では 0.1 秒後ともなると，ある程度異なる値を取っている．しかし，パワー最大となる時刻付近では近い値を取っている．また，これらの傾向は対数周波数でのスペクトル重心でも確認できる.

以上のことから，パワー最大となる時点でのスペクトル重心を特徴量として利用する.

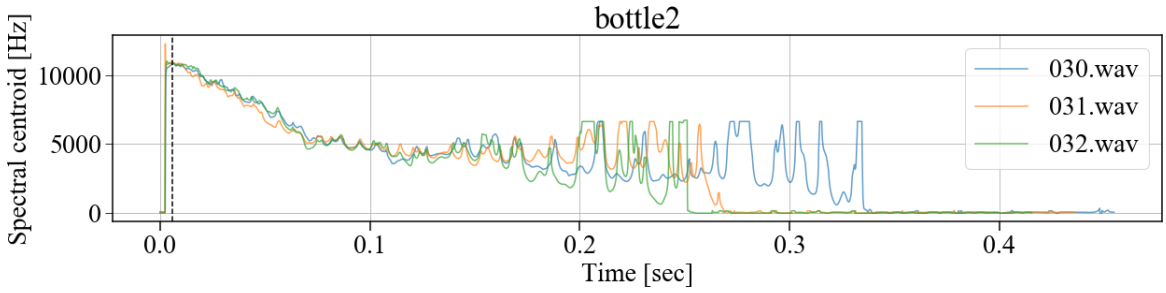


図 3-14 スペクトルの非常によく似た音同士（bottle2, 030.wav, 031.wav, 032.wav）での線形周波数でのスペクトル重心の時間変化．パワーが最大となる時刻（図中黒点線）が揃うように頭出しを行っている.

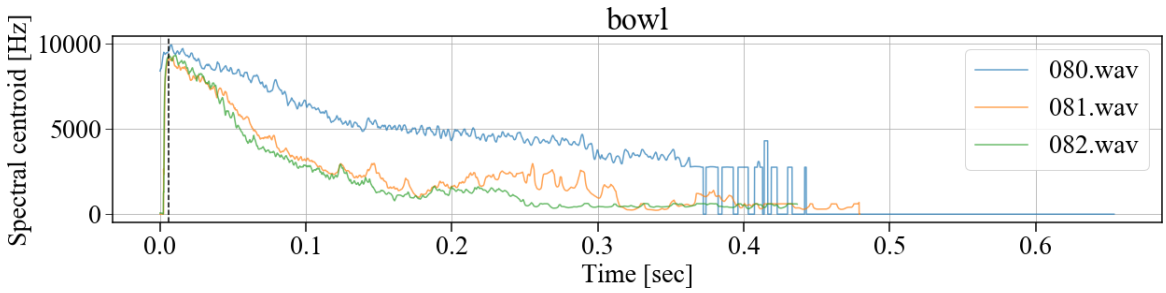


図 3-15 スペクトルの非常によく似た音同士（bowl, 080.wav, 081.wav, 082.wav）での線形周波数でのスペクトル重心の時間変化．パワーが最大となる時刻（図中黒点線）が揃うように頭出しを行っている．

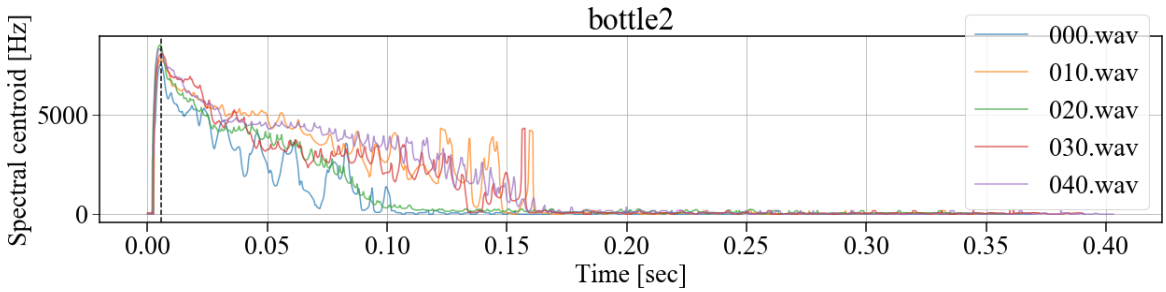


図 3-16 同じ種類の音同士（bottle2, 000.wav, 010.wav, 020.wav, 030.wav, 040.wav）での線形周波数でのスペクトル重心の時間変化．パワーが最大となる時刻（図中黒点線）が揃うように頭出しを行っている．

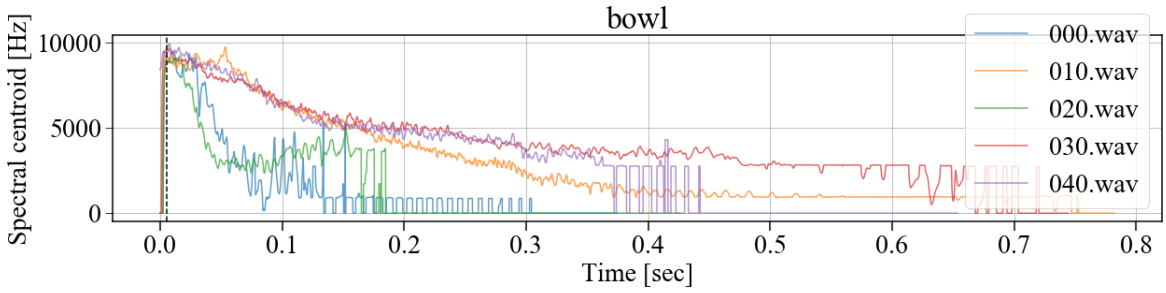


図 3-17 同じ種類の音同士（bowl, 000.wav, 010.wav, 020.wav, 030.wav, 040.wav）での線形周波数でのスペクトル重心の時間変化．パワーが最大となる時刻（図中黒点線）が揃うように頭出しを行っている．

時々刻々と変化するスペクトル重心から，パワー最大となるフレームでの値を参照して特徴量とする場合，信号から切り出すフレームのずれに注意する必要がある．音によっては，パワー最大時の線形周波数でのスペクトル重心は 6.4 kHz だが，フレームを 1 サンプルずつ 128 サンプルまでずらしていった際に得られる 128 個の線形周波数でのスペクトル重心は，平均が 7.6 kHz，標準偏差が 1.1 kHz，というように大きなばらつきを持ってしまう．そのため，フレームのずれには注意が必要である．本研究では，フレーム長さ 256 サンプル，フレームの重複 255 サンプルでの短時間フーリエ変換で計算したスペクトログラムからパワーが最大となるフレームを用いて線形周波数でのスペクトル重心を，計算されたパワーが最大となるフレームをもとに，長さを伸長し定 Q 変換を行うことで，対数周波数でのスペクトル重心を計算した．

ここまでに，スペクトログラム全体を利用したスペクトル重心の計算を説明したが，スペクトログラム中で顕著に表れているのは持続成分であり，持続成分は上記のスペクトル重心の値に大きく寄与していると思われる．そのため，スペクトログラム全体から計算せず，持続部分のみでスペクトル重心を計算することも考えられる．そこで，以下の手順により持続成分のみで線形周波数でのスペクトル重心を計算した．

8. フィルタの出力のうち正の値を持っている部分において横線が表れていることになるため，式(3-12)により持続成分以外のパワーを 0 とする．

$$X'_{\mathrm{persist}}(\omega,t)=\begin{cases}X'(\omega,t) & if\ \ P(\omega,t)>0\\ 0 & otherwise\end{cases} \tag{3-12}$$

9. 得られた$X'_{\mathrm{persist}}(\omega,t)$から ω のみを変数とした$X'_{\mathrm{persist}}(\omega)\big|_{t=t}$を式(3-8)の$X$に代入することでスペクトル重心を計算．

以上の処理で計算した持続成分のみを利用した線形周波数でのスペクトル重心の時間変化と，スペクトログラム全体を利用して計算したスペクトル重心の時間変化を図 3-18 に示す．約 0.11 秒の黒点線がパワー最大となる時刻を表すが，この時点でのスペクトル重心の値がほぼ一致していることが確認できる．図 3-18 に示す以外の音で試した場合でも，同様に値がほぼ一致した．本研究では背景雑音の無い音を利用するため，スペクトログラム全体を利用して計算するが，持続成分のみから計算することもできる．

定Q変換で計算したスペクトログラムでは，持続成分を抽出する処理ができていないため，対数周波数でのスペクトル重心についても同様に持続成分のみで計算した結果がスペクトログラム全体から計算した結果と一致するかどうかは不明である．

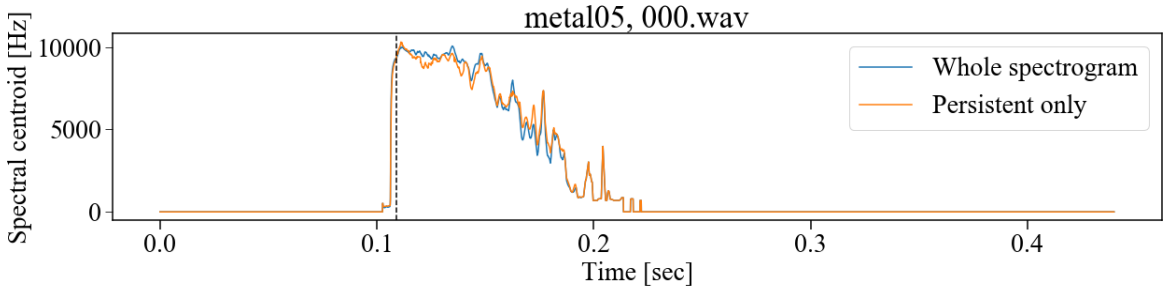


図 3-18 スペクトログラム全体を利用して計算した線形周波数でのスペクトル重心の時間変化と，持続成分のみを利用して計算した線形周波数でのスペクトル重心の時間変化．どちらも音は同じ音（metal05, 000.wav）を利用．図中黒点線はパワーが最大となる時刻を表している．

3.4 その他特徴量

打撃由来の突発音に関する先行研究[38],[40],[41]で利用されている特徴量を計算する．これは，既存特徴量での識別精度との比較及び，既存特徴量のうち提案する特徴量と組み合わせて利用することで精度が向上すものを選び出すためである．

3.4.1 $\tan\phi$

以下の式（3-13）で計算できる値を特徴量として計算する[38].

$$\frac{\sum_{i=1}^{N}\frac{a_i}{\pi f_i}w_i}{\sum_{i=1}^{N}w_i} \tag{3-13}$$

ここで，音の信号を N 個の周波数帯に分けてから減衰部分（パワー最大となる時刻とその後最大パワーの $1/e$ まで減衰する時刻の間）のみを取り出した i 番目の周波数帯の信号の時間変化を$x_i(t)$ $(i=1,2,\dots,N)$としたとき，$\log\big(x_i(t)\big)\approx a_i t+b_i$ と直線近似して得られた傾きをa_iとしている．また，f_i $(i=1,2,\dots,N)$を各周波数帯の中心周波数としている．

3.4.2 波形の減衰

波形の減衰部分の傾きをスペクトル重心で割った値を特徴量として計算する[41]．スペクトル重心は式(3-8)で計算できる．波形の減衰部分の傾きは，まず，元の音圧波形に対して，ヒルベルト変換を利用した波形包絡を計算する．その後，カットオフ周波数 50 Hz のバタワースフィルタを用いて低い周波数のみを残す．これにより大まかな振幅の時間変化が得られるため，それをもとに減衰部分を線型近似することで傾きを得る．図 3-19 に示すのは，上から順に波形，ヒルベルト変換により得られた波形包絡，バタワースフィルタにより得られた大まかな減衰の様子である．一番下のグラフが振幅の減衰の大まかな様子を捉えていることが確認できる．この図のうち，二つの赤破線で囲まれている領域を減衰部として，線型近似を行う．ただし，対数で変換したものに対して線型近似を行う．

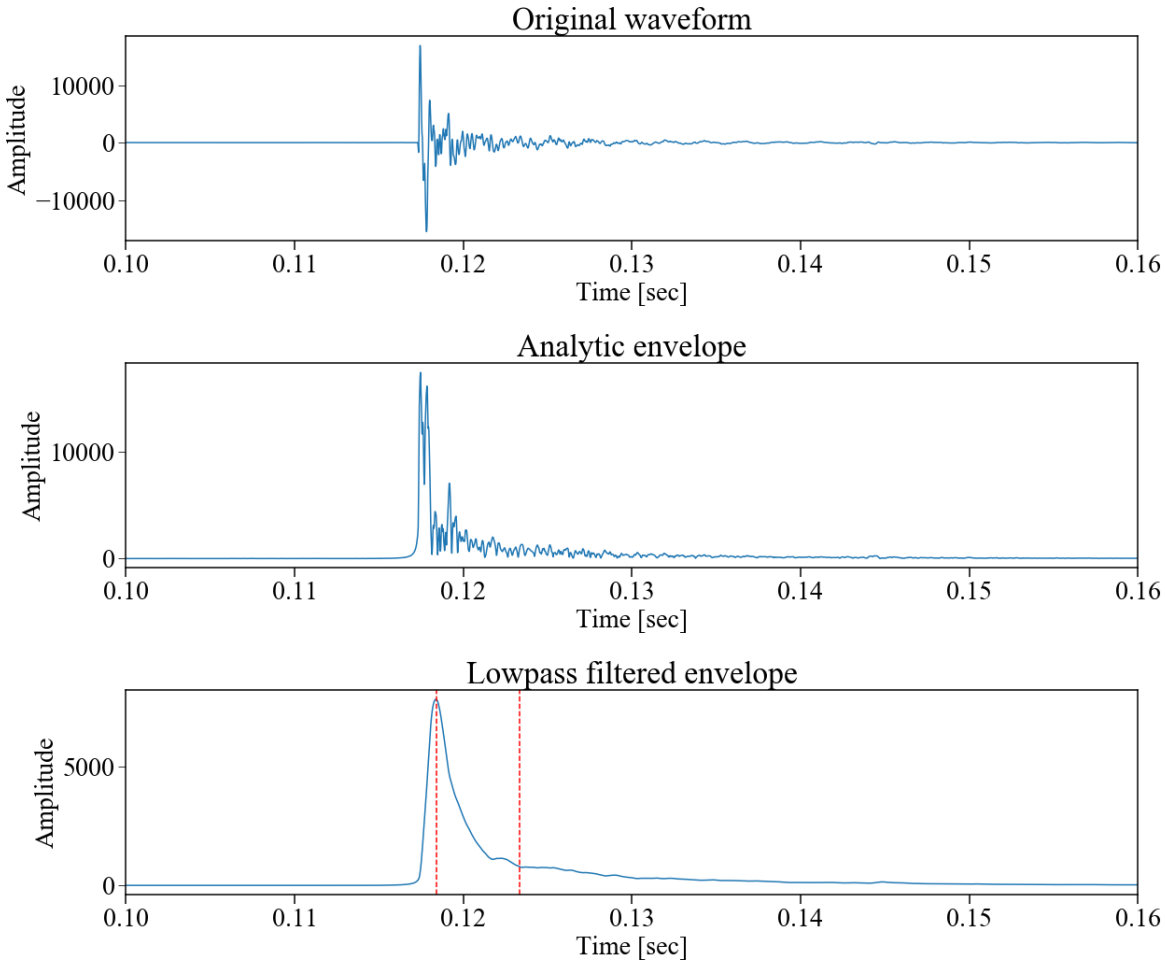


図 3-19 音圧波形（上段），ヒルベルト変換で得られた波形包絡（中段），バタワースフィルタを中段の信号にかけることで得られた波形振幅のおおまかな時間変化の様子を捉えたもの（下段）．下段図中の赤破線は左にあるものが信号が最大値となる時刻を表し，右側にあるものが最大以後最初に最大値の 10 %となった時刻を表す.

3.4.3 立ち上がりの早さ，減衰の速さ

図 3-20 のように振幅を正規化して絶対値をとった波形が 0.2 を超えた時点から 0.8 を超えた時点までの時間を立ち上がりの早さとして，その後 0.8 を下回った時点から，0.2 を下回った時点までの時間を減衰の早さとして計算する[40].

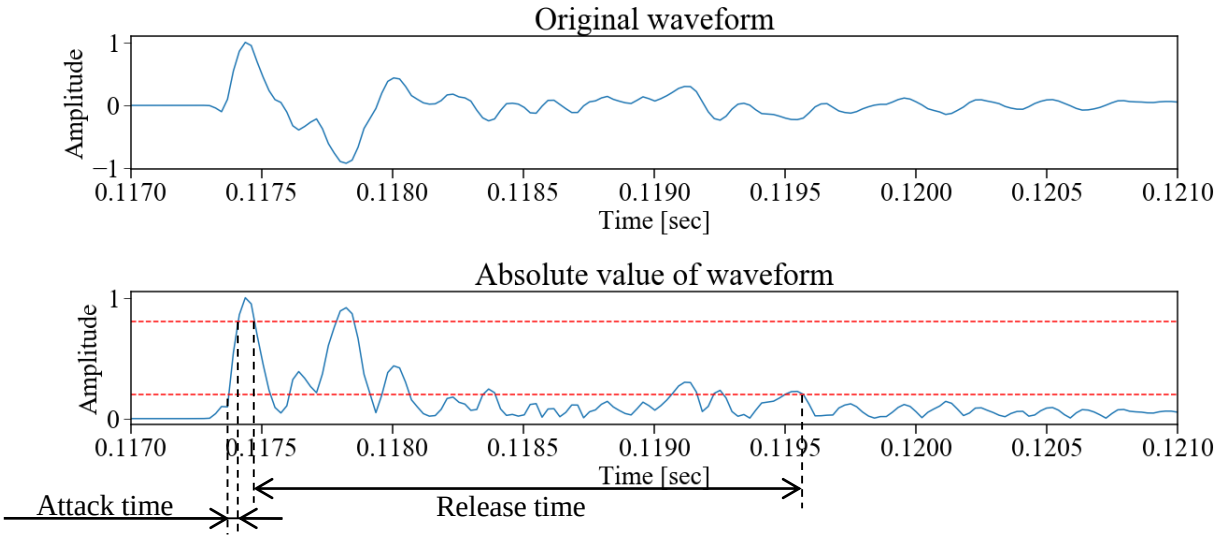


図 3-20 音圧波形（上段）と波形の絶対値（下段）.

3.4.4 Spectral bandwidth

以下の式（3-14）で計算できる値を特徴量として計算する[41].

$$\frac{1}{2\pi}\sqrt{\frac{\sum_k |w(k)|\left(\omega(k)-2\pi\times(spectral\ \ centroid)\right)^2}{\sum_k |w(k)|}} \tag{3-14}$$

ここで，$w(k)$はスペクトルを，$\omega(k)$は周波数を表す．また，スペクトル重心は式(3-8)で計算できる．

3.4.5 ラフネス

スペクトルに表れる極値に対して，全ての組で以下の式（3-15）で計算できる値を計算し，総和を撮ったものを特徴量として計算する[41].

$$\tau_{mn} = 0.5(A_m A_n)^{0.1} \times \left(\frac{2\min(A_m, A_n)}{A_m + A_n}\right)^{3.11} \times \left(e^{-3.5v|\omega_m-\omega_n|} - e^{-5.75v|\omega_m-\omega_n|}\right) \tag{3-15}$$

ここで，A_m, A_nはそれぞれスペクトルの振幅を，ω_m, ω_nはそれぞれ周波数を，vは

$$v = \frac{0.24}{0.0207 \times \min(\omega_m, \omega_n) + 2\pi \times 18.96} \tag{3-16}$$

である.

3.4.6 Spectral rolloff

スペクトルの振幅を低い周波数から累積していき，総和の x %を初めて超える周波数値を spectral rolloff と呼ぶ．x=0.85 の場合の値を特徴量として計算する[40].

3.4.7 スペクトルのなめらかさ

振幅スペクトルに対して，ローパスフィルタをかけ，元の振幅スペクトルとの相関係数を特徴量として計算する[40].

3.4.8 Zero crossing rate と波形のなめらかさ

波形が正の値から負の値に変化する頻度，つまり波形が 0 を交差する頻度，を zero crossing rate (ZCR)と呼ぶ．音圧波形に対する音の立ち上がりから約 5 ms での ZCR と，音圧波形に対してカットオフ周波数 4 kHz のローパスフィルタをかけた場合の音の立ち上がりから約 5 ms での ZCR を特徴量として計算する[40]．後者は波形の滑らかさを指す．

3.4.9 チャタリングの数と時間間隔

以下の図 3-21 では，時間軸上の赤い印の時刻にて突発音が 3 回発生している．このように一度の打撃で複数の突発音が発生することをチャタリングと呼び，スペクトログラム上で確認できるチャタリングについて，その本数と時間間隔を特徴量として計算する[40]．ただし，チャタリングが複数現れる場合は時間間隔の平均を計算する．

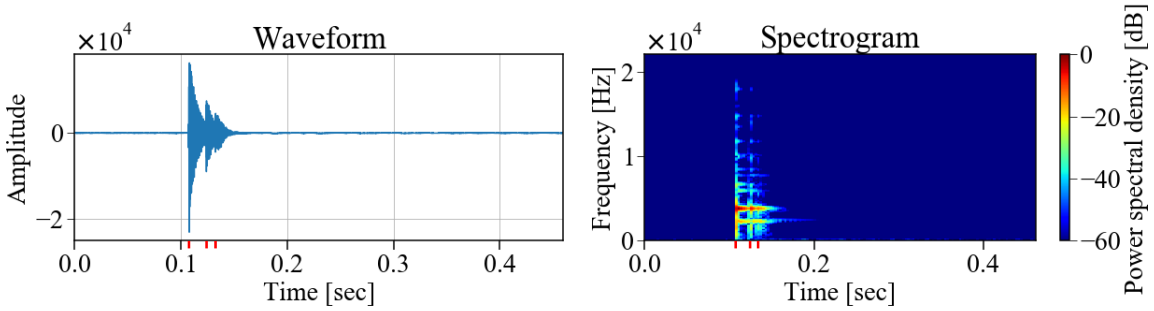


図 3-21 陶器を叩く音（china2, 060.wav）の波形（左）とスペクトログラム（右）．時間軸上赤い印を付けている時刻でチャタリングが 3 回現れている．

3.5 本章のまとめ

本章では，本研究で扱う特徴量の計算方法を示した．このうち，スペクトル重心と持続成分の顕著さ（平均と分散）は新規に提案したもの，tanϕ，波形の減衰，立ち上がり時間，減衰時間，Spectral bandwidth，ラフネス，Spectral rolloff，スペクトルの滑らかさ，Zero crossing rate，波形のなめらかさ，チャタリングの数，チャタリングの時間間隔，持続成分の明確度合いの計 13 種類は先行研究[38],[40],[41]にて提案されている特徴量である．新規に提案したものは，観察により考案したものである．

打撃由来の突発音において調波構造のようなスペクトルの詳細部分に現れる特徴は存在しないのではないかという仮定のもと，スペクトルの大まかな分布であるスペクトル重心を計算し，持続成分の顕著さについても，周波数値を落とすように平均と分散を計算している．

スペクトル重心の計算では，音の立ち上がりの時点でのスペクトル重心を特徴量としているが，頭出しが非常に重要である．また，持続成分のみのスペクトルの値をもとにスペクトル重心を計算した場合でもスペクトログラム全体をもとにスペクトル重心を計算した場合とほぼ同じ値が得られるが，本研究では背景雑音は考慮しておらず，持続成分のみから計算する理由がないため，スペクトログラム全体をもとに計算している．

持続成分の顕著さについて，スペクトログラム中で持続成分が現れている周波数において，持続成分の顕著さにピークが現れることが確認できるが，持続成分の顕著さにピークが現れているからといってスペクトログラム中に持続成分が現れているとは限らず，その点で計算方法に問題があるが，ガラスや陶器の音では値が大きく，木板や金属板では値が小さいという傾向は捉えているため，本研究では計算方法を変えず進めている．

次章ではこれらの特徴量の評価を行う．

第4章 識別による特徴量の評価

4.1 はじめに

3 章にて，持続成分の顕著さの平均と分散及びスペクトル重心を，持続成分を持つ突発音の新たな特徴量として示し，計算方法を提示した．

本章では，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．

まず初めに，既存特徴量との比較として，既存特徴量のみでの識別精度とそこに新たに提案した特徴量を加えた場合での識別精度を計算した．この際，複数の特徴量の組を適切に利用することで識別性能が変化する可能性があるため，mRMR による特徴選択を行った．

次に，人為的に設計し計算した特徴量による識別精度と，既存手法での識別精度の比較を行った．

その後，反響による影響の評価を行い，学習時に含まれない種類の音（例えば小太鼓の音）を入力した際の挙動を調べた．

利用しているデータと分割方法は以下のとおりである．

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．

識別の際，チャタリングが含まれる音と含まれない音を分けてしまうと，各種類の音での音の数に極端な偏りが生まれてしまう．そのため，識別においてはチャタリングが含まれるか否かによるデータの分割をしていない．

交差検証には 2 章でも示している以下の 3 種類の分割方法を利用している．ただし，各分割方法において，1 つ目のグループをパラメータの調整用に利用している．

1. 十の位で 10 分割

   2.2.1 節で述べたとおり，RWCP-SSD に収録されている音には通し番号が振られている．通し番号をもとに 10 の位が等しいものを一つのグループとして 10 分割交差検証を行う．例えば，010.wav, 011.wav, 012.wav, 013.wav, 014.wav, 015.wav, 016.wav, 017.wav, 018.wav, 019.wav の十個が一つのグループになる．このとき，各グループにはスペクトルの非常によく似た音のみが含まれるようになり，手を叩く音以外では他のグループとの間でスペクトルが非常に良く似ている組は存在しない．

2. 一の位で 10 分割

   通し番号の 1 の位が等しいものを一つのグループとして 10 分割交差検証を行う．例えば，000.wav, 010.wav, 020.wav, 030.wav, 040.wav, 050.wav, 060.wav, 070.wav, 080.wav, 090.wav の十個が一つのグループになる．このとき，スペクトルの非常に良く似た 10 個の音は 10 個全ての分割に散らばることになる．

3. 前から順に 5 分割

   通し番号をもとに，000.wav~019.wav，020.wav～039.wav，040.wav～059.wav，060.wav～079.wav，080.wav～099.wav という風に連番で 5 分割する．このとき，スペクトルが非常に良く似た音 10 個同士だけでなく 2.2.2 節で述べた 000.wav と 010.wav のようにスペクトルに多少違いはあるけれど似ている音同士も全て一つのグループに属することとなり，各分割間での音の再現性は最も低い．

一の位で 10 分割したものは，学習データと試験データで音の再現性が最も高くなり，前から順に 5 分割したものは，音の再現性が最も低くなる．実際の応用を考えると，前から順に 5 分割した場合での性能が重要だが，比較として他の二つも行っている．

4.2 mRMR による特徴選択

特徴量を用いた識別精度の比較をするにあたり，特徴選択を利用した．

異なるクラスで取りうる特徴量の値が全く被っていない場合，完全な識別が可能となる．しかし，計算により得られた単独の特徴量で分離ができるとは限らない．例えば，図 4-1 のような状況において，縦軸や横軸に垂直な識別境界ではほとんど識別ができないが，点線で示された斜めの線を引くことができればほとんど識別できる．つまり，複数の特徴量を組み合わせて識別することが必要である．

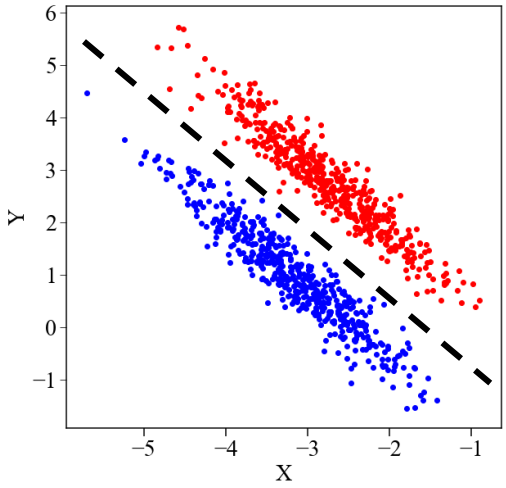


図 4-1 複数の特徴量を用いることで分離が可能となる分布の例．横軸がある特徴量 X を，縦軸がある特徴量 Y を，各点が一つのデータを点の色がクラスを表している．破線を識別境界とすれば二つのクラスのデータは完全に分離される．

特徴量を闇雲に増やしていけば，何らかの特徴量の組で分離が可能となるかもしれないが，識別の判断根拠を分かりやすくするには，必要最小限の特徴量での識別であることが重要である．特徴量の組み合わせについて考える際，すべての組み合わせについて試し，最も少ない特徴量で十分に高い識別精度を発揮した組み合わせを採用する方法もあるが，本研究では 17 個の特徴量を計算しているため，$2^{17} - 1 = 131{,}071$通り存在する．そのすべてで識別精度を求めるとすると，各 10 秒で終わるとしても 15 日かかることになる．そこで，本研究では mRMR[43]を利用した特徴選択を行った．

ある特徴量の値の変化がクラスの違いを反映しておらず，クラスの分離に全く寄与しない場合，この特徴量は必要ない．また，ある二つの特徴量の間に図 4-2 のように強い相関が表れている場合，識別のためには冗長であり，どちらか片方の特徴量で十分である．クラスの分離に寄与し，図 4-2 のように冗長性が高くない特徴量の組を探し出す手法として mRMR[43]と呼ばれる指標が提案されている．

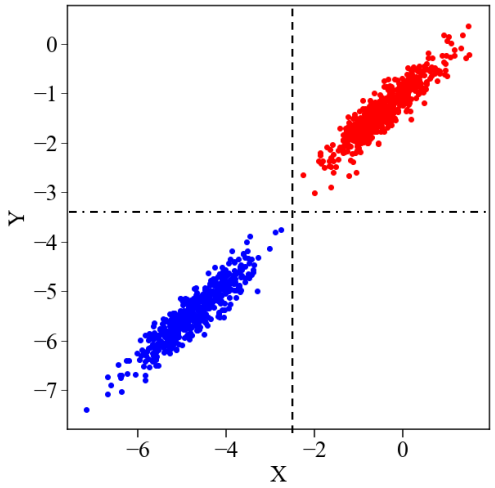


図 4-2 特徴量 X と特徴量 Y に相関がある場合の一例．破線もしくは一点鎖線のみで分離できるため，特徴量 X と特徴量 Y のどちらか一方で識別には十分である．

特徴選択は，M次元の特徴量からなるデータ$X = \{x_i, i = 1,2, \dots, M\}$が多数与えられた時，$M$次元空間$R^M$から，クラスの分離を適切に表すことができる m 個の特徴量で張られる部分空間R^mを探し出すことに相当する．mRMR では，$m-1$個の特徴量の組S_{m-1}が選ばれているとき，以下の式（4-1）で表される指標に基づいて最良の特徴量を 1 個追加してm個の特徴量の組 S_m を得る．これを$m = 1$から順に繰り返すことで，特徴量に順番をつける．

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right] \tag{4-1}$$

ここで，cはクラスを，$I(x;y)$は相互情報量を表し，

$$I(x;y) = \int\int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{4-2}$$

である．式（4.1）の第一項は特徴量が分離に寄与するかどうか，第二項はすでに選ばれている特徴量と冗長な関係にないかどうかを意味する．

この方法を用いて前章で示した特徴量を並べると，次のようになる．これは，2 章で示した RWCP-SSD に収録されている音を利用して計算している．また，計算にあたって量子化が必要なため，平均 0，標準偏差 1 となるように標準化を行った後，0 を基準に 1/3 刻みで量子化している．

1. 持続成分の顕著さ平均
2. スペクトルの滑らかさ
3. $\tan\varphi$
4. 持続成分の顕著さ分散
5. 線形周波数でのスペクトル重心
6. 波形の減衰
7. Spectral bandwidth
8. 減衰時間
9. 対数周波数でのスペクトル重心
10. 波形の滑らかさ
11. Spectral rolloff
12. Zero crossing rate
13. 立ち上がり時間
14. ラフネス
15. チャタリングの数
16. 持続成分の明確度合い
17. チャタリング間隔の平均

得られた順番で一つずつ特徴量を増やしていったときの識別精度の変化を示したものが，図 4-3〜図 4-5 である．識別精度はデータを 10 個ないしは 5 個のグループに分割し，パラメータ調整用の一つ以外のグループを利用した交差検証で計算している. 分割方法は 4.1 節で示したとおりであり，図 4-3〜図 4-5 がそれぞれ対応している．図は各グループが試験データとなった際の識別精度をそれぞれ計算し, 平均と分散を図中のエラーバーとして表示している. 丸で表されているのは学習データでの識別精度, 四角で表されているのは試験データでの精度である. クラスごとのデータ数に違いがあるため，識別精度はクラスごとにデータ数の逆数で重み付けをして計算している．また，識別には決定木を利用している．決定木の不純度の計算にはジニ係数を利用し，木の深さと枝の剪定のための各ノードに含まれるサンプル数の下限はパラメータ調整用のデータを利用した 5 分割交差検証により決めた．決定木は図 4-1 のように斜めに識別境界を引くことを苦手としており，クラスごとのデータ数が揃っていない場合不順度の計算に問題があるため，過学習を起こしやすいが，識別過程がはっきりと確認できる点が環境音認識における本研究の目的である判別の根拠を明確にすることに合致しているため，決定木を利用している．

図 4-3 を見ると，学習データでの精度と試験データでの精度の増減は同じような形となっており，学習データでの精度が上がればその分だけ試験データでの精度も上がっている．そのため，過学習は起きていない．また，3 つ目の特徴量までを利用することで試験データでの精度が 80%以上となり，その後特徴量を追加していっても大きな精度の向上は確認できない.

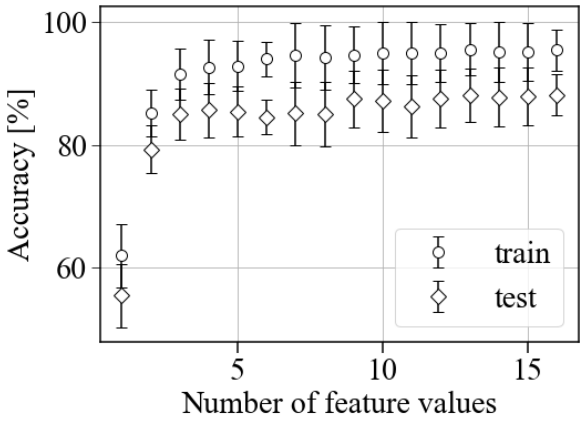


図 4-3 十の位で 10 分割した場合での，一つずつ特徴量を増やしていった際の各精度を示している．丸と四角はそれぞれ学習データでの精度と試験データでの精度を表している．

一方図 4-4 では，学習データでの精度と試験データでの精度の差が図 4-3 に比べて小さくなっている．また，特徴量を増やしていくことで，増分は減ってはいくものの，試験データでの精度が増え続けている．学習データとテストデータでの精度の差及び特徴量を増やすほど精度が上がるかどうかという，図 4-3 と図 4-4 の違いには，音の特定度が違うという課題の違いに由来する．

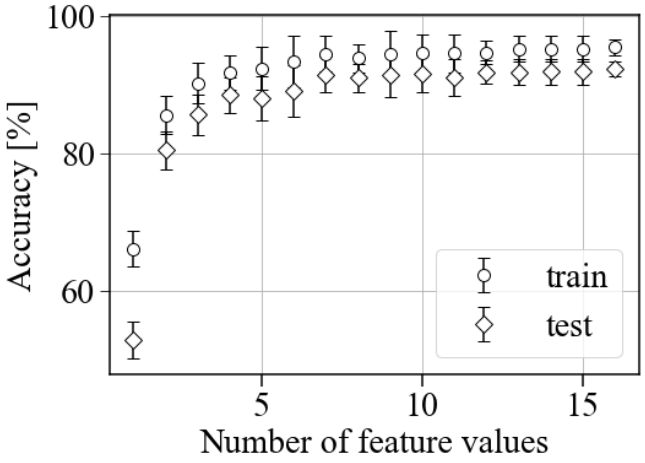


図 4-4 一の位で 10 分割した場合での，一つずつ特徴量を増やしていった際の各精度を示している．丸と四角はそれぞれ学習データでの精度と試験データでの精度を表している．

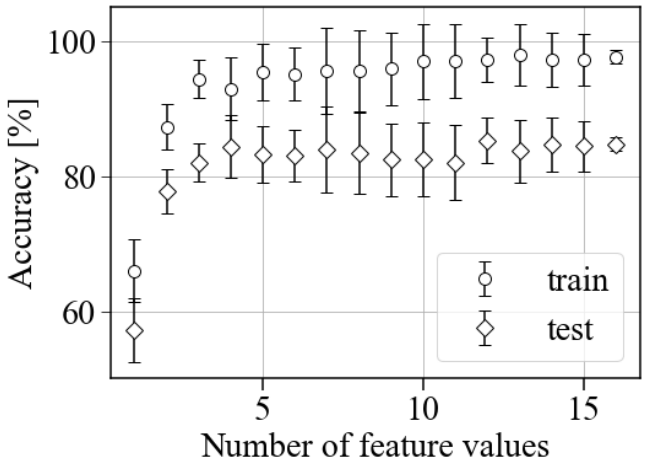


図 4-5 前から順に 5 分割した場合での，一つずつ特徴量を増やしていった際の各精度を示している．丸と四角はそれぞれ学習データでの精度と試験データでの精度を表している

図 4-5 に示しているのは，前から順に 5 分割した場合での識別精度である．これは音の特定度が低く，最も実際の利用に近い状況での分割方法である．テストデータと学習データでの識別精度の乖離は図 4-3 や図 4-4 よりも大きい．学習データでの識別精度の平均は 4 個目の特徴量でいったん下がるものの，その後順調に上がり続けている．一方，試験データでの識別精度は 5 個目の特徴量追加後に 7 個目を追加するとき以外では下がり続けており，わずかに過学習の疑いがある．テストデータでの識別精度は 4 個目以後ほとんど上がらず，7 個目の特徴量と 12 個目の特徴量でわずかに上がるが，4 個目までの特徴量のときの精度と大差ない．そのため，4 個目の特徴量までで識別には十分である．ただし，これは今回のデータでの場合であり，わずかに過学習の疑いがあるため，データ数を増やした場合には変化する可能性がある．

図 4-3〜図 4-5 の結果をまとめると，音の特定度が高い課題（図 4-4）では，特徴量が多いほど，学習データに適合し，それに伴い試験データでも精度が向上している．一方，音の特定度が低い課題（図 4-5）では，特徴量を増やしても試験データでの識別精度はあまり上がらず，持続成分の顕著さの平均と分散，tanϕ，スペクトルの滑らかさの計 4 個の特徴量で高い精度が得られている．

次に，本研究で提案したスペクトル重心と持続成分の顕著さの二つを除いた場合で同様の実験を行った．スペクトル重心に関しては既存研究[40]でも計算されているため，計算方法が異なるものの線形周波数でのスペクトル重心が特徴量に含まれている．得られた特徴量の順番は以下の通りである．

1. tan φ
2. スペクトルの滑らかさ
3. 波形の滑らかさ
4. Spectral bandwidth
5. 減衰時間
6. 波形の減衰
7. スペクトル重心
8. 立ち上がり時間
9. Spectral rolloff
10. ラフネス
11. Zero crossing rate
12. チャタリングの数
13. 明確度合い
14. チャタリング間隔の平均

前から 5 分割した分割検証（図 4-5 と同じ分割）にて，上記の順番で一つずつ特徴量を増やしていったときの識別精度の変化を示したものが，図 4-6 である．
識別精度は 80 %に満たず，先ほどの図 4-5 の状況よりも低くなっている．また，特徴量の順番に関してtanϕとスペクトルの滑らかさ，spectral bandwidth が早い段階で現れていることは，先ほどのスペクトル重心と持続成分の顕著さを含めた場合での特徴選択の結果と同様である．

以上の結果から，本来は識別に有用だが持続成分の顕著さの平均や分散と冗長な関係を持っているために抑制されていた特徴量は存在しない．そして，持続成分の顕著さの平均と分散があると識別精度が向上する．そのため，持続成分の顕著さは，識別に有効でありながら，既存の特徴量と相関が低い関係性がある新たな特徴量であることが分かる．

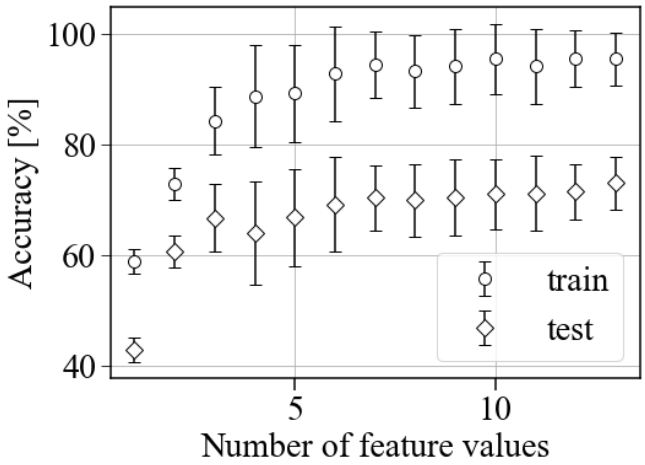


図 4-6 前から順に 5 分割した場合での，一つずつ特徴量を増やしていった際の各精度を示している．丸と四角はそれぞれ学習データでの精度と試験データでの精度を表している．

4.3 ６種類の突発音の識別での他手法との比較

本節では，特徴量による識別の精度とそれ以外の既存手法での精度との比較を行う．識別には 4.2 節で用いた音と同じ音を利用し，分割方法は前から順に 5 分割を利用する．この分割方法は音の特定度が低く，最も実際の利用に近い状況での分割方法である．パラメータ調整用のデータには５分割された 1 つ目のグループを利用する．

特徴量を利用する識別には，持続成分の顕著さ平均，スペクトルの滑らかさ，tanϕ，持続成分の顕著さ分散，の計 4 つの特徴量を利用し，決定木による識別を行う．不順度はジニ係数を利用し，木の深さと枝の選定のためのノードの下限サンプル数は，パラメータ調整用のデータでの 5 分割交差検証により決定した．

比較として他手法での識別も行なう．利用した手法は，メルスケールのスペクトログラムに対する主成分分析に基づく手法[44]と，メルスケールのスペクトログラムを入力とした畳み込みニューラルネットワークにて転移学習による識別を行なっている手法[15]である．前者は，RWCP-SSD データセットを利用した識別評価を行なっている手法であり，２章で示した識別と同じである．後者の畳み込みニューラルネットワークによる手法は，転移学習を利用している点で今回のデータ数の少ない識別に適しているため採用している．この転移学習による手法は ESC-50 データセットにて，人間の識別よりも高い精度（83.5％）を達成しており，畳み込み層による特徴抽出の後に使われる識別のための層を，線形サポートベクトルマシンに変えてしまっても同程度の精度（82.8％）が得られている．データ数が少ない状況でも良く機能すると考えたため，本研究ではサポートベクトルマシンを利用して識別を行なった．パラメータ C はパラメータ調整用のデータでの５分割交差検証により決定した．畳み込み層での特徴抽出で得られる特徴量は 2048 次元である．

本研究で計算した 4 種類の特徴量と決定木での識別精度は 84.5±4.7 %，メルスケールのスペクトログラムの主成分分析に基づく識別の精度は 70.0±7.2 %，転移学習により得られた特徴量とサポートベクトルマシンでの識別の精度は 91.7±4.8 %であった．ただし，クラスごとのデータ数に違いがあるため，精度はクラスごとにデータ数の逆数で重み付けをして計算している．このうち，特徴量と決定木での識別の混同行列を表 4-1 に，転移学習による特徴量抽出と線形サポートベクトルマシンでの識別の混同行列を表 4-2 に示す．

本研究で計算した 4 種類の特徴量による識別と，転移学習により得られた特徴量に基づく識別との混同行列の違いで特筆すべきは，表 4-1 と表 4-2 にて赤文字になっている，木板の音を陶器の音へと誤認する事，及びガラスと陶器の混同である．

木板を叩く音のうち陶器の音へと誤識別した音は例えば，図 4-7 に示す音であり，これは木板を叩く音としては持続成分がはっきりと現れる音である．

次に，陶器の音をガラスの音へと誤識別する数は，両手法とも高くなっている．これらの音は実際に聞いてみても非常によく似ており，識別が難しい音である．本研究で計算した 4 種類の特徴量での識別では，転移学習により得られた特徴量とサポートベクトルマシンでの識別に比べ，ガラ

スと陶器の間の誤識別の数が倍になっている．そのため，本来識別が不可能な音ではないけれども何らかの特徴を落としてしまったために誤識別している可能性がある．

これら以外の誤識別では，プラケースを叩く音が手を叩く音に誤識別している数が多い．

表 4-1　本研究で計算した 4 種類の特徴量と決定木での識別の混同行列

		予測結果					
		金属	硝子	プラケース	手	陶器	木
正解のクラス	金属	307	1	1	0	11	0
	硝子	2	257	0	0	60	0
	プラケース	2	0	95	13	0	0
	手	0	0	2	141	3	4
	陶器	8	65	0	0	134	33
	木	2	0	0	1	19	218

表 4-2　転移学習により得られた特徴量とサポートベクトルマシンでの識別の混同行列

		予測結果					
		金属	硝子	プラケース	手	陶器	木
正解のクラス	金属	290	23	0	3	0	4
	硝子	0	307	0	0	12	0
	プラケース	0	0	105	5	0	0
	手	0	0	0	148	0	2
	陶器	8	56	1	0	165	10
	木	0	0	0	3	0	237

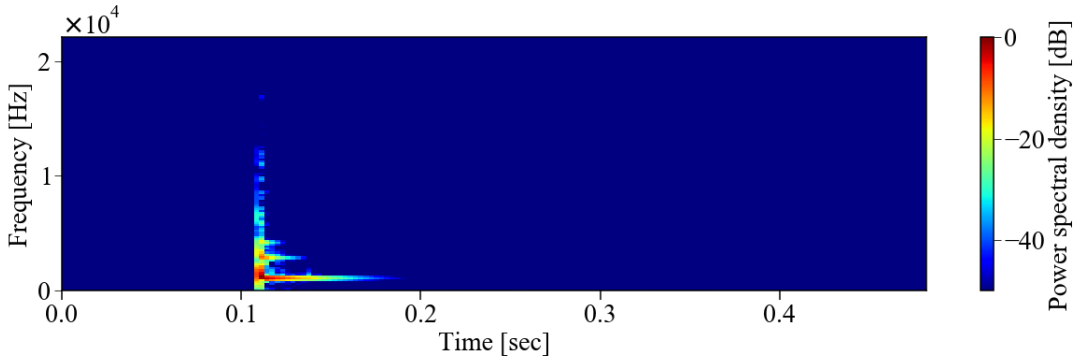


図 4-7 木板を叩く音（cherry1, 010.wav）のスペクトログラム．この音は陶器を叩く音に誤識別された．

4.4 決定木の判断基準

決定木の中身について，本研究で利用している全ての打撃由来の突発音を深さ 3 の決定木により分類した結果を図 4-8 に示す．より詳細な 5 層の決定木については付録に掲載している．

まず持続成分の顕著さの分散によって，持続成分が顕著に表れるガラスと陶器の音と，持続成分の顕著に表れないプラケース，金属，手を叩く音，木板の音とが分離されている．しかし，一部の木板はガラスや陶器の音と同じく持続成分の顕著さの分散が高くなっており，ガラスや陶器と同じ側に分離されている．この木板の音はその後にスペクトルの滑らかさにより陶器の音とある程度分離されているが，前節での木板を叩く音と陶器を叩く音との誤識別の原因となっている．持続成分の顕著さの分散で分離されたのちに，左側の木では持続成分の顕著さの平均による分離が横に並んでいるが，これは，図 4-1 のように斜めに識別境界を引く必要があることを表している．

持続成分の顕著さの分散で分離された左の木から，スペクトルの滑らかさはプラケースを叩く音と金属板を叩く音の組，手を叩く音と木板を叩く音の組，に分離している．プラケースを叩く音と金属板を叩く音はスペクトログラムにおいて，高い周波数まである程度の時間 noise patch のようにしてパワーが強く表れている音であるため，持続成分以外の周波数構成に関する特徴となっている可能性がある．これは右下の木板と陶器の場合も同様であり，陶器を叩く音では高い周波数まで持続成分だけが表れているが，木板ではそうではない．

持続成分の分散で分離された左側の木において，プラケースを叩く音と金属板を叩く音の組，手を叩く音と木板を叩く音の組はどちらも持続成分が少しでも表れているか否かの違いがスペクトログラムから確認できるため，持続成分の顕著さの平均で持続成分の有無を表すことができている．

持続成分の分散で分離された右側の木では，ガラスと陶器の分離が行われており，これは tanΦ に基づいている．tanΦは各周波数帯での持続成分の減衰早さを振幅スペクトルでの重みづけにより平均化したものであり，音の大部分が持続成分により構成されているガラスや陶器では減衰早さを細かく計算できており，わずかに減衰の早い陶器の音を分離できているのではないかと考える．

以上の事から，当初の狙い通り，持続成分の顕著さは持続成分がはっきりと表れているかどうかによる音の分類に機能している．

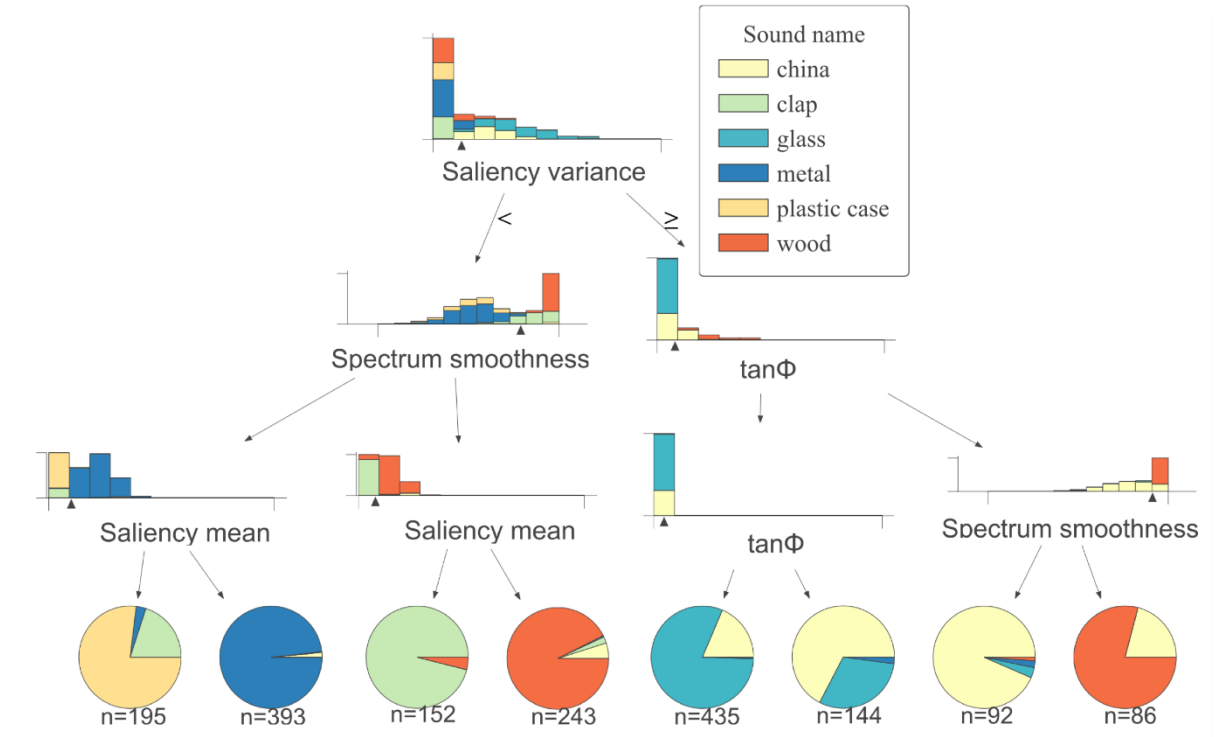


図 4-8 決定木の識別の様子．各ノードのヒストグラムはその下に示された特徴量についてのヒストグラムを表し，その大小によって，x 軸上のマーカーよりも大きな値の音は右に，小さな値の音は左に振り分けている．最終段は各終端ノードまでたどり着いた音の数と種類の割合を表している．Plastic case はプラスチックケースを叩く音，metal は金属を叩く音，clap は手をたたこと，wood は木板を叩く音，glass はガラス容器を叩く音，china は陶器を叩く音をそれぞれ意味する．

4.5 特徴量の分布

本節では，各特徴量についてスペクトルのよく似た音同士でのばらつきを計算している．これは，計算方法により生まれる誤差の評価にあたる．値のばらつきが小さいほど，同じ音での取りうる値の範囲が小さくなるため，他の音と分布が重なる可能性が下がるため，値のばらつきは小さいほうが良い．また，同時にチャタリングによる値の変化についても述べる．ばらつきの計算方法は以下のとおりである．ただし，この計算ではチャタリングを含まない音のみを利用している．

1. スペクトルのよく似た音同士 20 個ずつ（000.wav～019.wav，020.wav～039.wav 等）のグループに分割する．この際，チャタリングを含まない音が 4 個以下となるグループは取り除く．
2. 各グループでの特徴量の分散をそれぞれ計算する．音の種類 l，K 番目のグループに属する音の特徴量を x_{lki}，その平均を $\overline{x_{lk}}$，音の数を N_{lk} として，各グループでの特徴量の分散 s_{lk} は，

$$s_{lk} = \frac{1}{N_{lk}} \sum_{i=1}^{N_{lk}} (x_{lki} - \overline{x_{lk}})^2 \tag{4-3}$$

となる．

3. 音の種類ごとに分散の平均を取り，その後に平方根を取る．音の種類は木板を叩く音，ガラスを叩く音，等である．ラベル l に属するグループの数を N_l として，

$$\sqrt{\frac{1}{N_l} \sum_{k=1}^{N_l} s_{lk}} \tag{4-4}$$

を計算する．

箱ひげ図を用いて表示を行っているが，箱ひげ図については以下の図 4-9 の通りである．

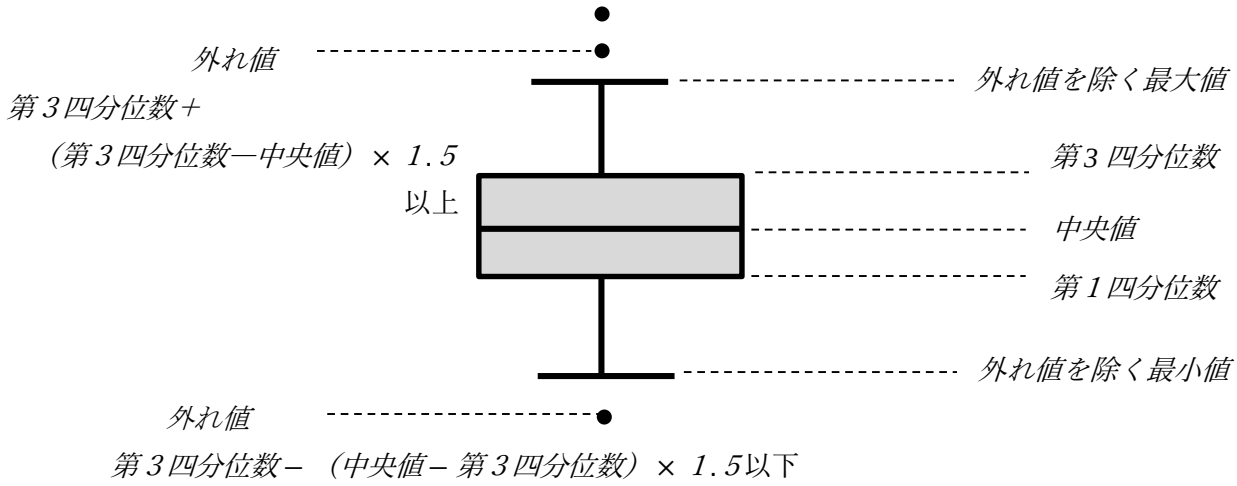


図 4-9 箱ひげ図の説明.

4.5.1 持続成分の顕著さの平均

図 4-10 に音の種類ごとの持続成分の明確度合いの平均（μ_{P_+}）の分布を，表 4-3 に式(4-4)から計算した各クラスの音におけるスペクトルのよく似た音同士での値のばらつきを示す．

表 4-3 のばらつきと 図 4-10 の分布を照らし合わせると，各種類の音のばらつきが，チャタリングの含まれない場合での第 1 四分位数から中央値までの幅に近い値になっていることが確認できる．各種類の音について，スペクトルのよく似た音は 10 組存在することを考えると，スペクトルのよく似た音同士の値の違いが，各種類の音全体での値のばらつきに占める割合は小さくない．

チャタリングについて，横線を強調するためのフィルタが約 5 2 ミリ秒の信号を参照しているため，その範囲にチャタリングがあるならば値が影響を受けるはずである．図 4-10 からは，チャタリングがある場合には値が小さくなる傾向が確認できる．ただし，よく似た音同士でない場合に値の分布が異なる可能性がある事，よく似た音同士での値のばらつきが小さくないこと，種類ごとで考えると決して音の数は多くないことから，チャタリングの影響が確かにあるかどうかの検定は行っていない．

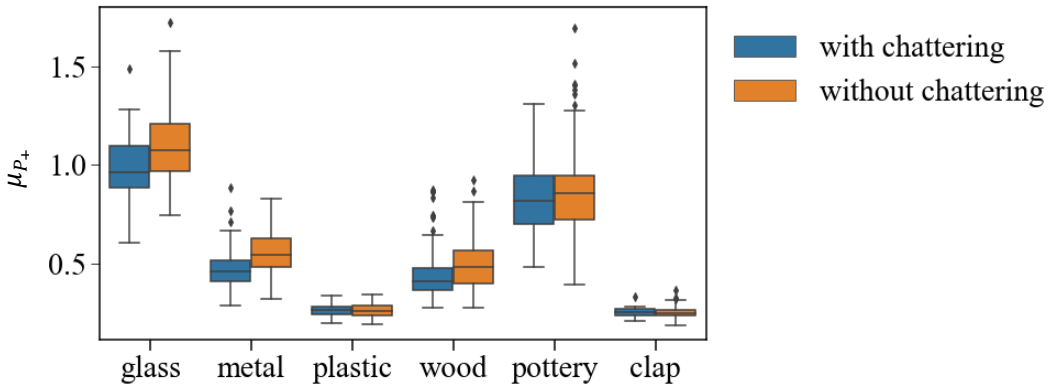


図 4-10 クラス毎の持続成分の顕著さ平均（μ_{P_+}）の分布．青はチャタリングの含まれる音を，橙はチャタリングの含まれない音を表す．Glass はガラス容器を叩く音，metal は金属を叩く音，plastic はプラケースを叩く音，wood は木板を叩く音，pottery は陶器を叩く音，clap は手を叩く音を指す．

表 4-3 式(4-4)から計算した，よく似た音での持続成分の顕著さ平均（μ_{P_+}）のばらつき．チャタリングの含まれない音のみで計算．

音の種類	ガラス	金属	プラケース	木板	陶器	手を叩く音
ばらつき	0.091	0.054	0.027	0.072	0.11	0.024

4.5.2 持続成分の顕著さの分散

図 4-11 に各クラスの音での持続成分の明確度合いの分散（σ_{P_+}）の分布を示し，式(4-4)から計算した各クラスの音における，スペクトルのよく似た音同士での値のばらつきを表 4-3 に示す．平均と同様，スペクトルの似た音同士のばらつきは各種類の音全体での第 1 四分位数から中央値までの幅に近い値となっており，スペクトルのよく似た音同士の値のばらつきが，各種類の音全体での値のばらつきに占める割合は大きい．また，チャタリングによって値が小さくなる傾向があるように見えるものの，平均の場合と同様の理由から検定は行っていない．

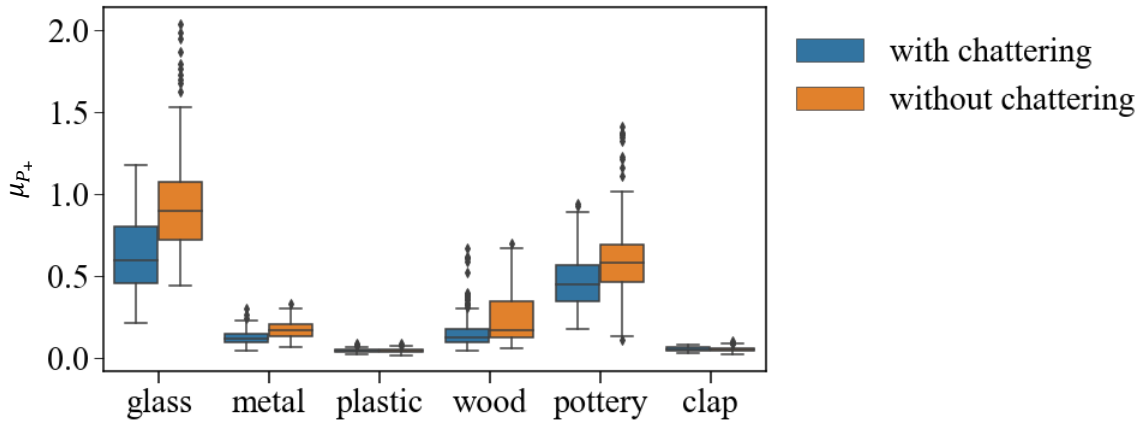


図 4-11 クラス毎の持続成分の顕著さの分散（σ_{P_+}）の分布．Glass はガラス容器を叩く音，metal は金属を叩く音，plastic はプラケースを叩く音，wood は木板を叩く音，pottery は陶器を叩く音，clap は手を叩く音を指す．

表 4-4 式(4-4)から計算した，よく似た音での持続成分の顕著さ分散（σ_{P_+}）のばらつき．チャタリングの含まれない音のみで計算．

音の種類	ガラス	金属	プラケース	木板	陶器	手を叩く音
ばらつき	0.16	0.030	0.011	0.049	0.12	0.013

4.5.3 tanϕの分散

図 4-12 に各クラスの音での tanϕの分布を示し，式(4-4)から計算した各クラスの音におけるスペクトルのよく似た音同士での値のばらつきを表 4-5 に示す．スペクトルの似た音同士のばらつきは，中央値が偏っているプラケースを叩く音以外では，各種類の音全体での第 1 四分位数から中央値までの幅よりも小さい値となっており，スペクトルのよく似た音同士の値のばらつきが各種類の音全体での値のばらつきに占める割合は持続成分の顕著さよりは小さい．また，チャタリングによって値が小さくなる傾向があるように見えるものの，検定は行っていない．

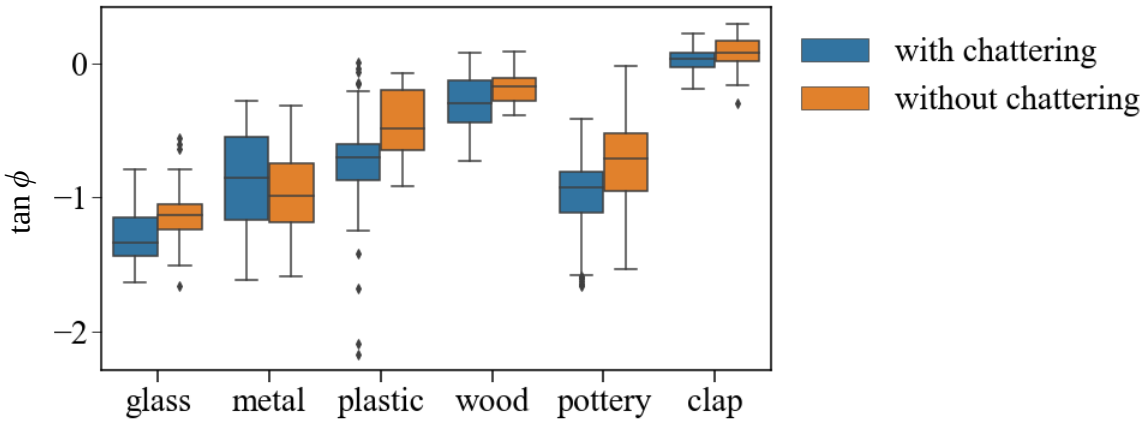


図 4-12 クラスごとの tanϕの分布．Glass はガラス容器を叩く音，metal は金属を叩く音，plastic はプラケースを叩く音，wood は木板を叩く音，pottery は陶器を叩く音，clap は手を叩く音を指す．

表 4-5 式(4-4)から計算した，よく似た音での tanϕのばらつき．チャタリングの含まれない音のみで計算．

音の種類	ガラス	金属	プラケース	木板	陶器	手を叩く音
tanphi	0.087	0.065	0.096	0.042	0.15	0.075

4.5.4 スペクトルの滑らかさ

図 4-13 に各クラスの音でのスペクトルの滑らかさの分布を示し，式(4-4)から計算した各クラスの音におけるスペクトルのよく似た音同士での値のばらつきを表 4-6 に示す．スペクトルの似た音同士のばらつきは，プラケースを叩く音と手を叩く音以外では，各種類の音全体での第 1 四分位数から中央値までの幅よりも小さい値となっており，スペクトルのよく似た音同士の値のばらつきが各種類の音全体での値のばらつきに占める割合は他の特徴量と比べて比較的小さい．また，チャタリングによって値が変化する傾向も無い．

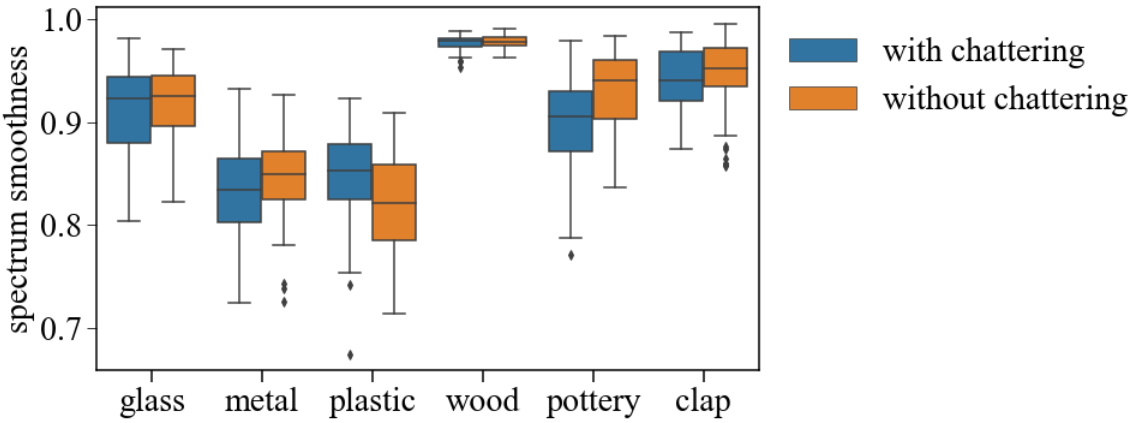


図 4-13 クラスごとのスペクトルの滑らかさの分布．Glass はガラス容器を叩く音，metal は金属を叩く音，plastic はプラケースを叩く音，wood は木板を叩く音，pottery は陶器を叩く音，clap は手を叩く音を指す．

表 4-6　式(4-4)から計算した，よく似た音でのスペクトルの滑らかさのばらつき．チャタリングの含まれない音のみで計算．

音の種類	ガラス	金属	プラケース	木板	陶器	手を叩く音
スペクトルの滑らかさ	0.025	0.023	0.034	0.0045	0.021	0.022

4.5.5 まとめ

持続成分の顕著さの平均と分散では，スペクトルのよく似た音同士での特徴量のばらつきが音全体での特徴量の範囲に対して 1/4 程度であるため，決して小さいとは言えない．

しかし 4.3 節の識別において，学習データと試験データでの音の再現性が高いほど識別精度が良くなっている．そのため，スペクトルのよく似た音同士での特徴量のばらつきは現状の識別では，識別精度への影響は小さい．同様の理由から，チャタリングによる識別への影響も小さい．

4.6 学習データと全く同じ音に反響を畳み込んだ音の識別

ここまでは RWCP-SSD に収録されている無響室録音のデータを用いた識別を行なったが，本節では反響がある場合に識別がどのように変化するのかを示す．

利用するデータは，２章で示している RWCP-SSD に収録されている 6 種類の音と，それらに対して RWCP-SSD に収録されている会議室での反響（残響時間 0.38 秒）を畳み込んだ音である．無響室録音での音すべてを用いて決定木を学習させたのちに，反響を畳み込んだ音の識別を行なった．決定木のパラメータは，不順度がジニ係数，木の深さ 5，各ノードの最小サンプル数 5 である．

結果の混同行列を表 4-7 に示す．多くの音は金属と木に分類されている．そのため，反響の有無による特徴量の変化は大きな課題である．反響により値の分布が最も大きく変化する特徴量は $\tan\phi$ であるが，4.4 での可視化結果から識別に影響を与えているのは，持続成分の顕著さの分散である．その値の分布を図 4-14 に示す．ガラス以外の殆どの音が 0.2～0.3 付近の値をとるように変化しており，もともとそこにあるのは金属を叩く音と木板を叩く音である．この反響による特徴量の変化が識別結果を左右しており，反響の影響が問題となる状況では反響に頑健な抽出方法に改良する必要がある．

表 4-7　学習データと全く同じ音に反響を畳み込んだ音の識別結果

		予測結果					
		陶器	プラケース	硝子	金属	手	木
正解のクラス	陶器	3	9	128	123	0	27
	プラケース	0	24	0	125	0	1
	硝子	7	0	334	50	0	8
	金属	0	2	5	393	0	0
	手	0	2	0	129	0	59
	木	0	2	0	32	0	266

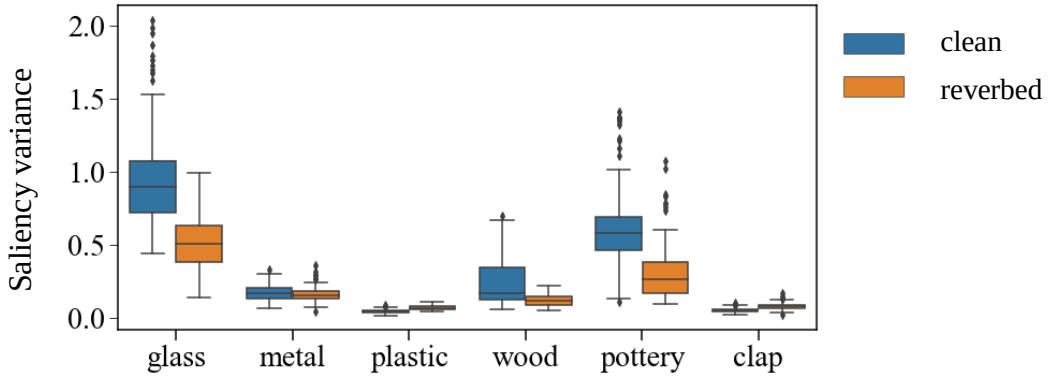


図 4-14 各種類の音での無響室録音での値の分布（青）と会議室での反響を畳み込んだ場合の値の分布（橙）．この分布はチャタリングが含まれない音のみを用いている．

4.7 未知の突発音の識別

実際にライフログや監視等に応用する場合は，多種多様な音が入力され，その中から，持続のある突発音を識別することになる．そのような，学習時に利用した音とは異なる種類の音が入力された際に，どのような挙動をするのかは重要である．

しかし，4.6 節にて反響がある場合では識別ができないことが確認できている．そこで，RWCP-SSD データセットに含まれる他の音を利用した識別を行なった．

具体的には，持続成分の顕著さ平均，スペクトルの滑らかさ，tanϕ，持続成分の顕著さ分散，の計 4 個の特徴量と決定木を用いた識別が，以下に示す音を何に識別するかを調べた．

同時に，木板を叩く音と金属製の物体を叩く音での識別結果も調べているが，これらの音は学習データに含まれていたものとは木板場合は寸法が，金属製物体の場合は形状が大きく異なるため，同種の音は学習データに含まれているけれども似た音は全く含まれていない．

- 緩衝材を潰す音．非常に減衰が早く，持続成分が現れない音である．
- クラッカーの音．非常に減衰が早く，持続成分が現れない音である．
- カスタネットの音．カスタネットは木製ではなくプラスチック製のものが使われている．
- 太鼓の音．減衰は早いが，持続成分が現れる．
- 金属製の貯金箱を叩く音．金属由来の音である．
- コーヒー缶を叩く音．金属由来の音である．
- スプレーを噴射する音．突発音ではなく，持続成分が現れない音である．
- 木板を叩く音．先の実験では利用していない音を利用した．

識別の結果を以下の表 4-8 に示す．また，特徴量の分布の例を図 4-15～図 4-17 に示す．

表 4-8　2 章で示した音以外の音に対する識別結果.

		予測結果					
		金属	硝子	プラケース	手	陶器	木
正解のクラス	緩衝材	0	0	40	34	0	26
	クラッカー	0	0	16	0	0	2
	カスタネット	0	0	10	2	15	73
	太鼓	11	0	17	0	64	8
	貯金箱	88	0	12	0	0	0
	コーヒー缶	81	1	0	0	18	0
	スプレー	9	0	81	0	0	10
	木	0	0	0	1	53	345

図 4-15 に学習に利用した 6 種類の音と緩衝材を潰す音の特徴量の分布を，図 4-16 に学習に利用した 6 種類の音と金属製貯金箱を叩く音の特徴量の分布を示す．図は可視化のために主成分分析を行い，第 1 主成分と第 2 主成分を表示している．

緩衝材を潰す音は，プラケースを叩く音，手を叩く音，木板を叩く音の 3 種類に識別されているが，図 4-15 からこれは特徴量の値がそれらの音での特徴量の分布に重なっているためであることが確認できる．

金属製貯金箱を叩く音は，金属板を叩く音，プラケースを叩く音に識別されており，主には金属板を叩く音に識別されている．この識別についても，図 4-16 から特徴量の値がそれらの音での特徴量の分布に重なっているためであることが確認できる．

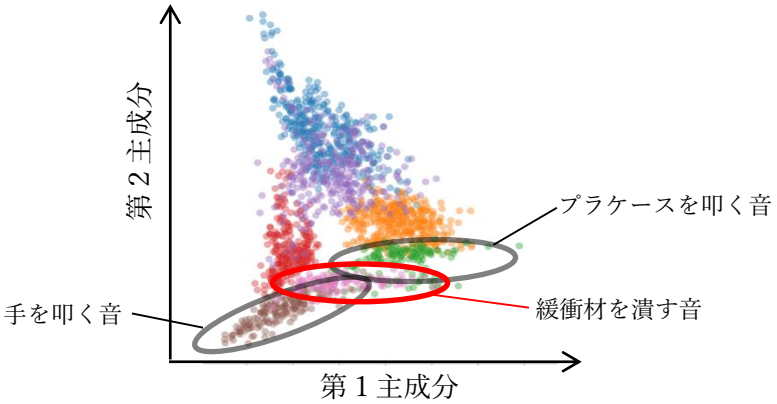


図 4-15 学習に利用した各音と，緩衝材を潰す音の特徴量の分布．可視化のために主成分分析を行い，第 1 主成分と第二主成分を表示している．音は，青色がガラス，橙が金属，緑がプラケース，赤が木板，紫が陶器，茶色が手を叩く音であり，緩衝材を潰す音はピンクで示している．

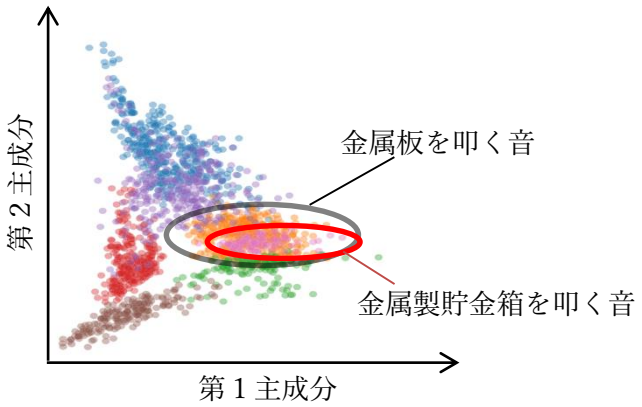


図 4-16 学習に利用した各音と，金属製貯金箱を叩く音の特徴量の分布．可視化のために主成分分析を行い，第 1 主成分と第二主成分を表示している．音は，青色がガラス，橙が金属，緑がプラケース，赤が木板，紫が陶器，茶色が手を叩く音であり，金属製貯金箱を叩く音はピンクで示している．

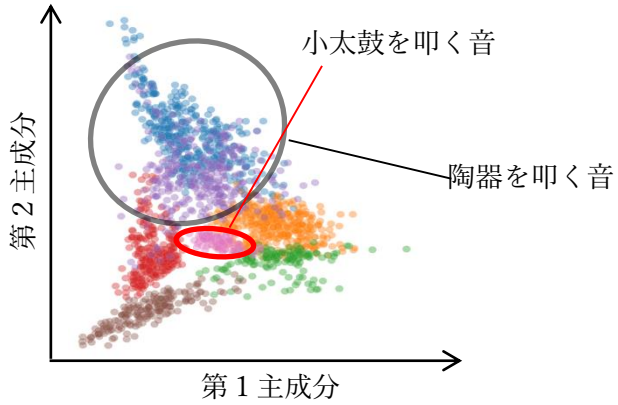


図 4-17　学習に利用した各音と，小太鼓を叩く音の特徴量の分布．可視化のために主成分分析を行い，第１主成分と第二主成分を表示している．音は，青色がガラス，橙が金属，緑がプラケース，赤が木板，紫が陶器，茶色が手を叩く音であり，金属製貯金箱を叩く音はピンクで示している．

一方，小太鼓を叩く音は，主に陶器を叩く音に分類されているが，分布はどの音ともあまり重なっていない．また，決定木での識別もサンプル数が 6 しかないノードに識別された．そのため，音が存在しない値を取ったために分類規則上たまたま陶器を叩く音へと識別されたことになる．

これらの結果をまとめると，持続成分を持たない緩衝材を潰す音は同じく持続成分を持たないプラケースを叩く音と手を叩く音に識別される．同様に持続成分を持たないクラッカーの音もプラケースを叩く音へと識別される．また，スプレーの音は主にプラケースを叩く音に識別されているが，この音は鳴っている時間自体は長いものの，持続成分を持たない音である．そのため，持続成分を持たない音でかつ手を叩く音よりも長い音であるプラケースを叩く音に識別されたと考える．

一方で，緩衝材を潰す音は一部木板を叩く音へと識別されており，それも特徴量が重なっているため識別モデルの問題ではなく，特徴量の問題である．

これらのことから，持続成分のある音とない音を区別することにある程度成功しているものの，持続成分があるけれどもはっきりとは表れておらず，数も少ない木板を叩く音の識別は完全ではない．

コーヒー缶を叩く音と金属製貯金箱を叩く音は金属板を叩く音に，木板を叩く音は木板を叩く音と陶器を叩く音に識別されている．これは，木板を叩く音を陶器を叩く音へと混同することも含めて，4.3 での識別と合致するものである．また，金属製貯金箱を叩く音のみ特徴量の分布を可視化しているが，金属製貯金箱を叩く音の特徴量は，金属板を叩く音の特徴量の分布に重なっている．そのため，この結果はたまたまではなく，4.3 での結果は形状の多少異なる音源物体を利用した音に対しても期待できるものである．

最後に，小太鼓の音は陶器の音に，カスタネットの音は木板の音に識別されているが，これらは

特徴量の値が学習に用いられたどの音の特徴量の分布にも重なっておらず，それによりたまたま陶器と木板へと識別されたことが確認された．これらの音は持続成分を多少なりとも持っており，そのためにプラケースを叩く音や手を叩く音とは異なる特徴量を取ったが，学習データに含まれる持続成分を持つ 4 種類の音のどれとも違う特徴的な持続成分となっている．それらの音が異なる分布となっていることから，持続成分を持っている音として的確に特徴量に反映されていることが確認された．

ただし，小太鼓の音やカスタネットの音を別の音として識別するには異常検出のような枠組みが必要である．

4.8 本章のまとめ

既存特徴量にパワー最大時のスペクトル重心と持続成分の顕著さの平均及び分散を加えた場合と，既存特徴量のみの場合とで mRMR による特徴選択と決定木による識別を行なった．その結果から，持続成分の顕著さの平均及び分散は，既存特徴量では捉えられていない識別に有効な特徴を抽出していることが確認された．

また，音の特定度の異なる 3 種類の分割方法で交差検証を行なった結果，音の特定度が高い場合ほど学習データと試験データでの精度の乖離が少なくなっていることが確認された．このことは，スペクトルのよく似た音同士での値の差が，同種類ではあるけれどよく似た音ではない音との差よりも小さいことを意味する．各特徴量は，スペクトルのよく似た音から計算した場合でも同種類での最大値から最小値の幅の 1/4 のばらつきを持つことが確認できており，持続成分の顕著さに関しては計算方法の都合上，チャタリングの有無の影響を受けて値が変化する音もあるはずだが，よく似た音同士の値の差は，同種類ではあるけれどよく似た音ではない音との値の差よりも小さく，今回の識別には影響を与えていない．

決定木の構造の分析と未知の音への識別を行ったところ，持続成分の顕著さの平均及び分散，$\tan\phi$，スペクトルの滑らかさの 4 つの特徴量を利用した識別において，持続成分の有無が識別結果を左右する大きな要因であることが確認され，持続成分がはっきり表れているか否かは打撃による突発音の材質の識別に有効であることが確認された．さらに，小太鼓の音やカスタネットの音のような持続成分を持っていながら，学習に利用した木板を叩く音，金属板を叩く音，陶器を叩く音，ガラスを叩く音のどれとも異なる特徴を持つ音においても，適切に特徴量が計算されていることが確認された．

持続成分の顕著さの平均及び分散，$\tan\phi$，スペクトルの滑らかさの 4 つの特徴量を利用した識別に残された課題は，木板の音と陶器の音の混同，ガラスと陶器の混同，反響による影響を特徴量が受けてしまうこと，小太鼓やカスタネットの音のように全く違う特徴量を持つ音が存在する場合には異常検出のような枠組みが必要であることである．

第5章　まとめと展望

5.1 結論

環境音認識では，大量のデータを集めることが決して容易ではないこと，環境音の特徴表現について不明点が多いことが問題である．そのため，本研究ではデータ数が少なく済み，識別に使われている音の特徴が理解しやすい特徴量抽出と決定木による識別を行った．

特徴表現が不明な環境音として，持続成分を持つ突発音が挙げられる．持続成分を持つ突発音は，食器の音やガラスの割れる音，踏切の打鐘式警報機の音，自転車のベルの音，鐘の音や流しの水滴の音，コインを落とす音や鍵束の音など，様々な場面で発生する音まで多数存在する重要な音である．

そこで本研究では，持続成分を持つ突発音の識別を行った．人の聞き分けの類推から，そのような音の識別で重要な側面は金属音，ガラスらしい音，のような音源の材質に関するものであると考え，材質の識別という方法で特徴量を評価した．ここで，打撃による突発音に限定しているが，他の持続成分を持つ突発音と打撃による突発音の発生原理は本質的に大きく違わないため，この限定は一般性を損なわない．

本研究では新たに，打撃に由来する持続成分を持つ突発音の特徴量として，持続成分の顕著さ提案しその計算方法を示した．この特徴量は，隣接する周波数帯のパワーとの差分を取ることで，持続成分が顕著に表れていると大きな値を取る特徴量である．

持続成分の顕著さを既存特徴量に加えることで，既存特徴量を用いた識別精度(73.1 %)よりも高い精度(84.8 %)が得られている．そのため，提案した特徴量は既存特徴量では捉えられていない音の特徴を捉えることができている．

さらに特徴選択を行った結果，持続成分の顕著さの平均と分散，$\tan\phi$,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(84.5 %)が得られることが分かった．つまり，これら 4 次元の特徴量以外の既存特徴量を追加しても音の識別に必要な情報は得られず，4 次元の特徴量で十分ということが確認された．

決定木の構造の分析と未知の音への識別を行ったところ，持続成分の顕著さの平均及び分散，$\tan\phi$，スペクトルの滑らかさの 4 つの特徴量を利用した識別において，持続成分の有無が識別結果を左右する大きな要因であることが確認され，持続成分がはっきり表れているか否かは打撃による突発音の材質の識別に有効であることが確認された．さらに，小太鼓の音やカスタネットの音のような持続成分を持っていながら，学習に利用した木板を叩く音，金属板を叩く音，陶器を叩く音，ガラスを叩く音のどれとも異なる特徴の持続成分を持つ音においても，特徴量は的確に音を表していることが確認された．

持続成分の顕著さの平均と分散，$\tan\phi$，スペクトルの滑らかさの計 4 次元の特徴量は，どれもスペクトル全体の統計量となっており，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識別に必要としないことも意味している．つまり，具体的なスペクトルは重要ではなく，例えばスペクトログラムを入力とした識別器を用いる場合でも，周波数分解能の高さはさほど重要ではないことになる．

大規模データセットを用いて深層学習を行った手法での識別結果と比べ，精度は 6%程度劣るものの，陶器とガラスの混同，木板と陶器の混同以外の音では同程度の識別精度が得られた．このことは学習データを少なくしても識別可能である可能性を示してはいるが，比較対象として利用した学習による識別は特に打撃音を識別する目的で提案されたものではないこと，識別精度が超えているわけではないことから，大量のデータが必要であるという環境音認識における問題点の改善までは至っていない．一方で，特徴量と決定木での識別を用いているため 4.4 節で示したように識別の判断基準の明瞭さは高くなっており，識別の明瞭さの改善には繋がっている．

最後に，持続成分の顕著さの平均及び分散，$\tan\phi$，スペクトルの滑らかさの 4 つの特徴量を利用した持続成分の存在する突発音の識別に残された課題は，木板の音と陶器の音の混同，ガラスの音と陶器の音の混同，反響による影響を特徴量が受けてしまうこと，小太鼓やカスタネットの音のように学習データに存在しない特徴量を持つ音が存在する場合には異常検出のような枠組みが必要であることである．

5.2 今後の展望

単発の突発音の識別に関する今後の展望として，反響による影響の除去，ガラスの音と陶器の音の混同及び木板の音と陶器の音の混同の抑制 2 点が挙げられる．

反響による影響について，反響が異なる場合に特徴量が変化してしまうことは，実用上の問題だけでなく，反響を揃えなければならないため，音の収集を難しくしてしまう．そのため，反響の影響を受けないように計算方法を改良する必要がある．

混同，特にガラスの音と陶器の音の混同について，例えば次の方法で計算される特徴量は，ガラスと陶器の識別が可能となる可能性を示している．ガラス瓶を叩く音について，ノイズを除去するためのハイパスフィルタを適用せずに定 Q 変換で計算したスペクトログラムをもとに対数周波数でのスペクトル重心を計算した結果は図 5-1 のようになるが，図 5-1 に示すパワー最大となる時刻（黒破線）とスペクトル重心が最大となる時刻（緑破線）との時間差を計算し，持続成分の顕著さの平均及び分散とスペクトル重心の最大値の 4 種類の特徴量に合わせ，主成分分析を行うと，図 5-2 のようなガラスと陶器が分離された分布が得られる．この時間差というものが何を指しているのかがはっきりとしないため，本研究では特徴量として利用していないが，ガラスと陶器の識別が可能となる可能性を示している．

次に，単発の突発音ではなく突発音の集合から成る環境音の識別に発展させることが考えられる．本研究では単発の持続成分を持つ突発音の識別を行ったが，実際の応用で認識したい音には，ガラスの割れる音のような複数の突発音の集合からなる音も存在する．このような突発音の集合から成る環境音に対する個々の突発音の音色を利用した識別への応用が期待される．

最後に，本研究では指数的に減衰していく突発音のみが入力される状況に限定しているが，実際にはそれ以外の音も入力されるため，何らかの一般化が必要である．

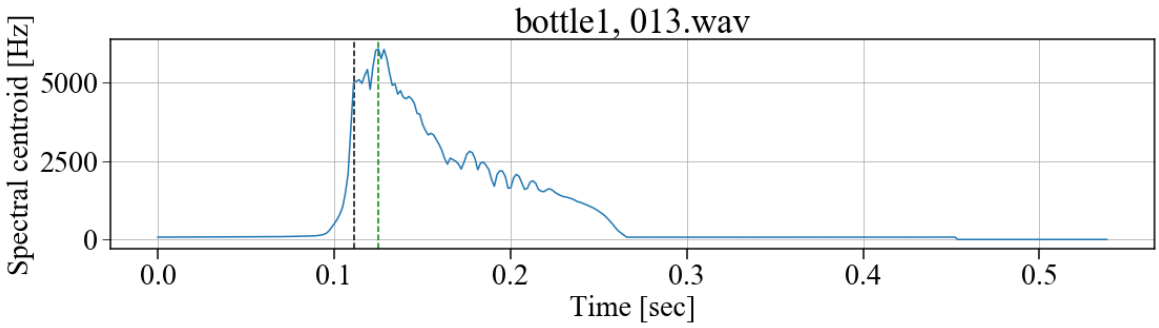


図 5-1 ガラス瓶を叩く音（bottle1, 013.wav）の対数周波数でのスペクトル重心の時間変化．黒い破線がパワー最大時を示し，緑の破線がスペクトル重心の最大時を示す．

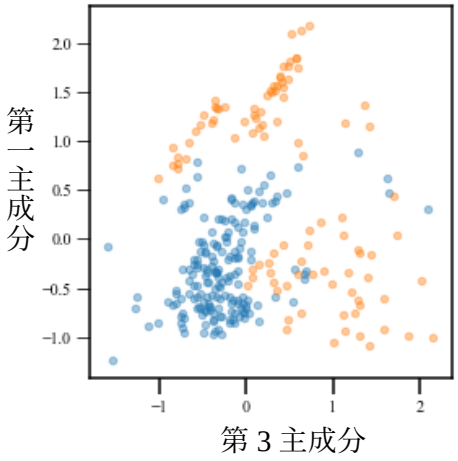


図 5-2 パワー最大とスペクトル重心最大との時間差，持続成分の顕著さの平均及び分散，スペクトル重心の最大値の 4 つの特徴量を 2 章で示したガラス由来の打撃音と陶器由来のだけ基音で計算し，主成分分析した結果を第 1 主成分と第 3 主成分を示した図．青い点がガラスの音，橙の点が陶器の音を表す．

参考文献

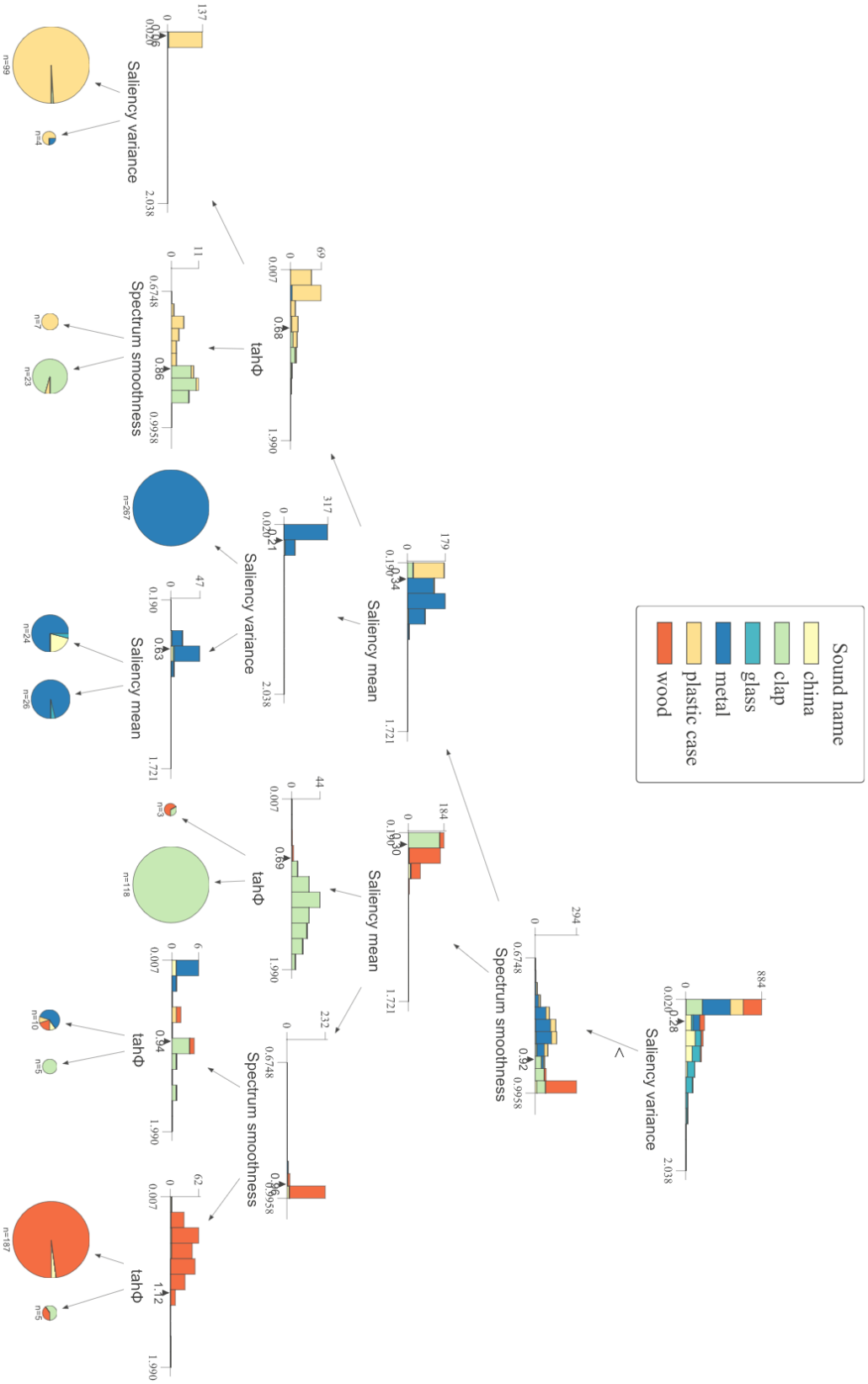
[1] 河本満, 浅野太, 車谷浩一, “マイクロフォンアレイを用いた音環境の見守りによる非日常音と危険状態の検出システム,” *情報処理学会研究報告. UBI, ユビキタスコンピューティングシステム*, vol. 19, pp. 19–26, 2008.
[2] 柘植康彦, 大西昇, “聴覚障害者のための警告音識別,” *電子情報通信学会技術研究報告. IE, 画像工学*, vol. 98, no. 575, pp. 1–6, 1999.
[3] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," *2016 IEEE International Conference on Multimedia and Expo*, Toronto, Ont., pp. 885-888, 2006.
[4] 稲邑哲也 *他*, “飲料缶・ボトル類を目と手と耳で分別廃棄するヒューマノイド行動の実現,” *日本ロボット学会誌 = J. Robot. Soc. Japan*, vol. 25, no. 6, pp. 813–821, 2007.
[5] 小泉 宣夫, 基礎 音響・オーディオ学. 第6刷, コロナ社, 2005.
[6] 大石 康智, “あらゆる音の検出・識別を目指して : 音響イベント検出研究の現在と未来,” 日本音響学会研究発表会講演論文集, 2014, pp. 1521–1524.
[7] 比屋根一雄 *他*, “RWCP実環境音声・音響データベース,” 人工知能学会全国大会論文集, 2002, vol. JSAI02, p. 190.
[8] K. J. Piczak, “ESC: Dataset for environmental sound classification,” *Proceedings of the 2015 ACM Multimedia Conference*, 2015, pp. 1015–1018.
[9] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.
[10] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, “Large-scale weakly labeled semi-supervised sound event detection in domestic environments,” *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, 2018, pp. 19–23.
[11] A. Mesaros, T. Heittola, and T. Virtanen, “TUT Database for Acoustic Scene Classification and Sound Event Detection,” *24th European Signal Processing Conference 2016*, 2016.
[12] C. V Cotton, D. P. W. Ellis, and A. C. Loui, “Soundtrack classification by transient events,” *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 473–476.
[13] O. Kalinli, S. Sundaram, and S. Narayanan, “Saliency-driven unstructured acoustic scene classification using latent perceptual indexing,” *2009 IEEE International Workshop on Multimedia Signal Processing*, 2009, pp. 1–6.
[14] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang, and T. Sainath, “Deep Learning for Audio Signal Processing,” *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 206–219, 2019.
[15] A. Kumar, M. Khadkevich, and C. Fügen, “Knowledge Transfer from Weakly Labeled Audio Using Convolutional Neural Network for Sound Events and Scenes,” *2018 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 326–330, 2017.
[16] 篠田浩一, 音声認識. 講談社, 初版, 2017.
[17] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, 2017.
[18] Z. Boqing, W. Changjian, L. Feng, L. Jin, L. Zengquan, and P. Yu-xing, “Learning Environmental Sounds with Multi-scale Convolutional Neural Network,” *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2018.
[19] H. B. Sailor, D. M. Agrawal, and H. A. Patil, “Unsupervised filterbank learning using Convolutional Restricted Boltzmann Machine for environmental sound classification,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-August, pp. 3107–3111, 2017.
[20] 飯國洋二, 基礎から学ぶ信号処理. 初版, 培風館, 2004.
[21] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, “Exploiting spectro-temporal locality in deep learning based acoustic event detection,” *Eurasip J. Audio, Speech, Music Process.*, vol. 2015, no. 1, Dec. 2015.
[22] Y. Wang and F. Metze, “A first attempt at polyphonic sound event detection using connectionist temporal classification,” *2017 IEEE International Conference on Acoustics, Speech and Signal*

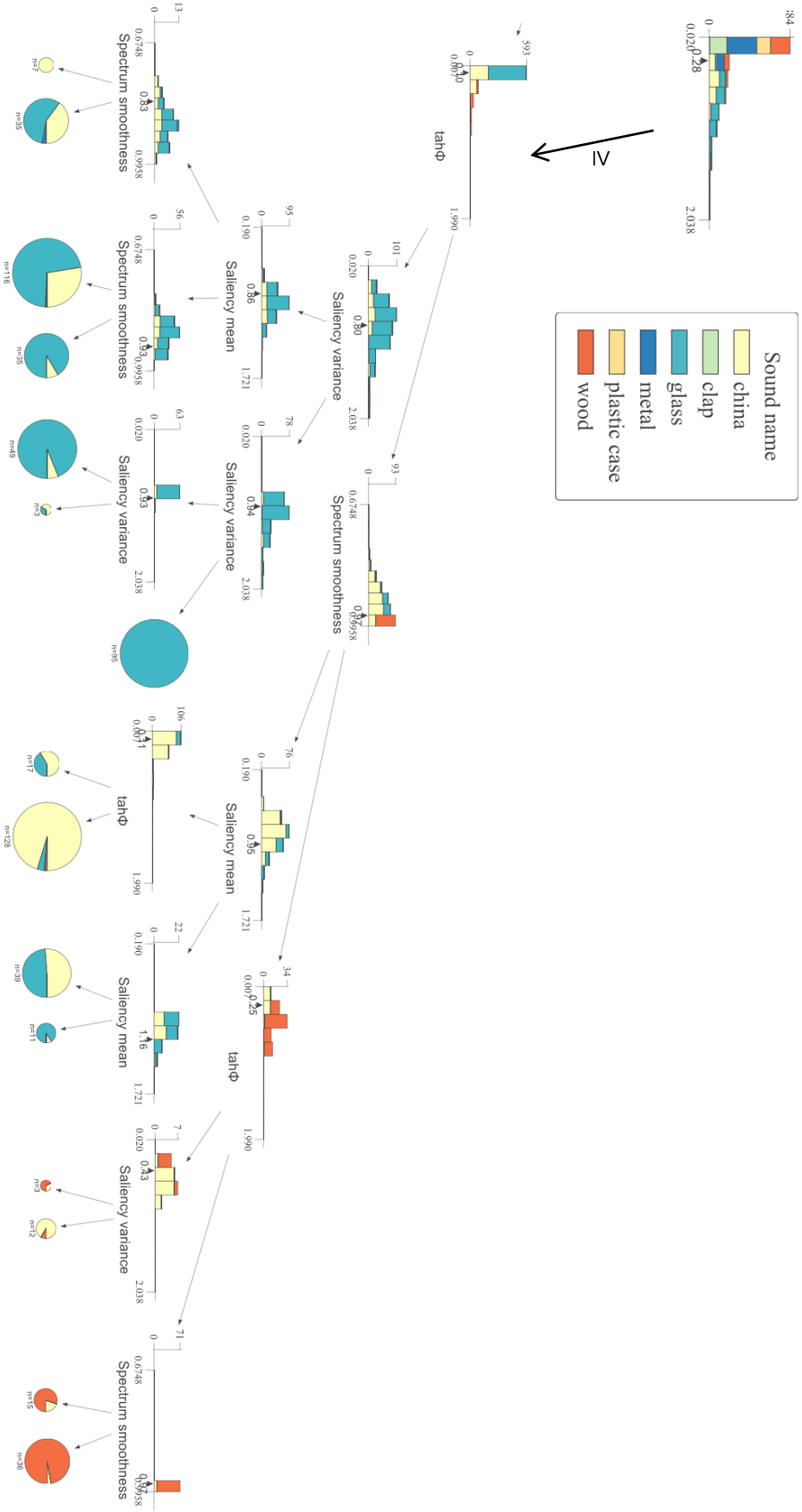
Processing (ICASSP), pp. 2986–2990, 2017.
[23] 井本桂右, “音響イベントと音響シーンの分析,” *日本音響学会誌*, vol. 74, no. 4, pp. 198–207, 2018.
[24] Y. Tokozume, Y. Ushiku, and T. Harada, “Learning from Between-class Examples for Deep Sound Recognition,” *International Conference on Learning Representations*, 2018.
[25] Y. Aytar, C. Vondrick, and A. Torralba, “SoundNet: Learning Sound Representations from Unlabeled Video,” *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp.892-900, 2016.
[26] 三木一浩, 西浦敬信, 中村哲, 鹿野清宏, “HMMを用いた環境音識別の検討,” *情報処理学会研究報告音声言語情報処理（Slp）*, vol. 1999, no. 108. pp. 79–84, 1999.
[27] 岡本 亜紗子, 林田 亘平, 中山 雅人, 西浦 敬信, “環境音認識のための最尤状態数の検討,” *情報処理学会研究報告（Slp）*, vol. 2013. no. 8, pp. 1-6, 2013.
[28] G. Muhammad and K. Alghathbar, “Environment Recognition from Audio Using MPEG-7 Features,” *2009 Fourth International Conference on Embedded and Multimedia Computing*, pp. 1–6, 2009.
[29] D. P. W. Ellis, X. Zeng, and J. H. McDermott, “Classifying soundtracks with audio texture features,” *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5880–5883, 2011.
[30] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, “Noise robust voice activity detection based on periodic to aperiodic component ratio,” *Speech Commun.*, vol. 52, no. 1, pp. 41–60, Jan. 2010.
[31] 厨川 守 他, オーディオと音楽のための音質のすべて. 誠文堂新光社, 1981.
[32] N. Saint-Arnaud and K. Popat, “Analysis and Synthesis of Sound Textures,” *Computational Auditory Scene Analysis*, USA: L. Erlbaum Associates Inc., pp. 293–308, 1998.
[33] J. H. McDermott and E. P. Simoncelli, “Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis,” *Neuron*, vol. 71, no. 5, pp. 926–940, Sep. 2011.
[34] B. Gygi, G. Kidd, and C. Watson, “Spectral-temporal factors in the identification of environmental sounds,” *J. Acoust. Soc. Am.*, vol. 115, pp. 1252–1265, Mar. 2004.
[35] D. Villamizar, D. Battaglino, D. G. Muratore, R. Hoshyar, and B. Murmann, “Sound Classification using Summary Statistics and N-Path Filtering,” *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2019.
[36] S. Sivasankaran and K. M. M. Prabhu, “Statistics based features for unvoiced sound classification,” *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2013.
[37] C. Zheng and D. L. James, “Rigid-Body Fracture Sound with Precomputed Soundbanks,” *ACM Trans. Graph.*, vol. 29, no. 4, Jul. 2010.
[38] B. L. Giordano, “Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates,” *Acoust. Soc. Am. J.*, vol. 119, no. 2, p. 1171, Jan. 2006.
[39] E. Krotkov, R. Klatzky, and N. Zumel, “Robotic Perception of Material: Experiments with Shape-Invariant Acoustic Measures of Material Type,” *Experimental Robotics IV*, Springer-Verlag, Jan. 1996.
[40] S. Okubo, Z. Gong, K. Fujita, and K. Sasaki, “Recognition of Transient Environmental Sounds Based on Temporal and Frequency Features,” *Int. J. Autom. Technol.*, vol. 13, no. 6, pp. 803–809, 2019.
[41] M. Aramaki, M. Besson, R. Kronland-Martinet, and S. Ystad, “Controlling the Perceived Material in an Impact Sound Synthesizer,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 2, pp. 301–314, 2011.
[42] T. Koumura and S. Furukawa, “Context-Dependent Effect of Reverberation on Material Perception from Impact Sound,” *Sci. Rep.*, vol. 7, no. 1, p. 16455, 2017.
[43] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *Proceedings of the 2003 IEEE Bioinformatics Conference*, pp. 523-528, 2003.
[44] X. Wang, H. Zhou, Z. Liu, and Y. Gu, “Large Scale Environmental Sound Classification Based on Efficient Feature Extraction,” *Proceedings of the International Conference on Parallel Processing Workshops*, vol. 2016-September, pp. 421–425, 2016.

[45] Y. Suzuki and A Futoshi, *et al.*., “An optimum computer-generated pulse signal suitable for the measurement of very long impulse response,” *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1119–1123, 1995.
[46] B. H. Repp, “The sound of two hands clapping: An exploratory study,” *The Journal of the Acoustical Society of America*, vol. 81, no. 4, pp. 1100-1109, 1987.
[47] A. Akay, “A review of impact noise,” *J. Acoust. Soc. Am.*, vol. 64, no. 4, pp. 977–987, 2013.
[48] 水澤富作, 近藤八重, 滝沢宣人, 河原田豊, “球体の衝突を受ける平板から発生する衝撃音に関する基礎的研究,” 土木学会論文集, vol. 2004, no. 766, pp. 47–57, 2004.
[49] J. C. Brown, “Calculation of a constant Q spectral transform,” *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.
[50] T. Kobayashi and J. Ye, “Acoustic feature extraction by statistics based local binary pattern for environmental sound classification,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 3052–3056, 2014.
[51] J. Ye, T. Kobayashi, M. Murakawa, and T. Higuchi, “Robust acoustic feature extraction for sound classification based on noise reduction,” *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5944–5948.
[52] J. Ren, X. Jiang, and J. Yuan, “Sound-event classification using pseudo-color CENTRIST feature and classifier selection,” *Proc.SPIE*, vol. 10011, 2016.
[53] J. Ren, X. Jiang, J. Yuan, and N. Magnenat-Thalmann, “Sound-Event Classification Using Robust Texture Features for Robot Hearing,” *IEEE Trans. Multimed.*, vol. 19, no. 3, pp. 447–458, 2017.
[54] B. Gygi, G. R. Kidd and C. S. Watson,“Similarity and categorization of environmental sounds,” *Perception & Psychophisics*, vol. 69, no. 6, pp. 839-855, 2007.

付録

付録A 決定木の識別規則




Sound name
china
clap
glass
metal
plastic case
wood
IV
tahΦ
Saliency variance
Spectrum smoothness
Saliency mean
Saliency variance
Spectrum smoothness
Saliency variance
Spectrum smoothness
tahΦ
Saliency mean
Saliency mean
tahΦ
Saliency variance
Spectrum smoothness

付録B 実行環境

Pyhon 3.6.2

・librosa 0.6.0
・numpy 1.16.2
・scipy 1.2.1
・pandas 0.24.2
・scikit-image 0.15.0
・scikit-learn 0.20.3
・pytorch 1.0.1

謝辞

はじめに，指導教員の佐々木健教授には学部四年の頃から 3 年に渡って多大なご教授，ご指導をして頂きました．ここに厚く感謝致します．

また，研究会など多くの場面で貴重な助言を下さった保坂寛教授，森田剛准教授に感謝致します．

研究テーマが関連深いということで研究について数多くの助言を頂き，相談に乗って下さった正田さん，Gong さん，大久保さん，金君，小松君，井上さん，研究テーマは違えども，居室を共にし，研究の進め方などの手本を示して下さった佐々木研，保坂研，森田研の皆様には大変お世話になりました．心から感謝いたします．

様々な手続きや備品の管理を行ってくださった秘書の三枝さん，森本さん，下川さんに感謝いたします．

最後に大学生活を通して変わらず支えて下さった家族と友人達へ心から感謝の意を表し、本論文の結びと致します。

2020 年 1 月

藤田　健斗