動画計測ラベルを用いた
深層学習によるピアノ演奏におけるサステイン・ソフトペダル深度推定

石坂玲生
法政大学情報科学部 ディジタルメディア学科
reo.ishizaka.5y@stu.hosei.ac.jp

Abstract

This study proposes a method to estimate continuous sustain and soft pedal depths from piano performance audio. Existing studies have excluded continuous pedal depth estimation for the soft pedal and have required high-cost data collection. Our approach works to estimate continuous pedal depth from piano audio, includes the soft pedal, and enables low-cost data collection in real performance environments. The method feeds acoustic features into a convolutional neural network and a bidirectional recurrent network and estimates sustain and soft pedal depth at each frame. During training, the method classifies pedal depth into 256 levels, and inference converts the class outputs into continuous values. Pedal depth labels are collected from video recordings of markers attached to the pedal mechanism and aligned with the audio in time. The process works without dedicated sensors or internal MIDI data and enables low-cost data collection in real performance environments. The best model achieved RMSE 0.191 for sustain pedal depth and 0.202 for soft pedal depth, with three temporal convolution layers (kernel width 15) followed by a bidirectional GRU. In terms of MSE, the best model slightly outperformed study[1] on the sustain pedal (0.0363 vs. 0.0425).

1 はじめに

1.1 ピアノ演奏におけるペダル表現

ピアノ演奏において、楽譜に書かれた音符や記号だけでは表しきれない演奏者の意図や個性が、打鍵速度・テンポの揺れ・ペダル操作といった細かな演奏表現に現れる。これらの中でもペダル操作は、和声のつながりや音色の質感を大きく左右する重要な要素である。ペダルの組み合わせや踏む込みのタイミング・ベロシティを変化させることでピアノ演奏に彩りを加える。

ピアノはサステインペダル、ソフトペダル、ソステヌートペダルの３つのペダルを備える。サステインペダルはすべてのダンパーを弦から離すことで、音の減衰を遅らせ、倍音成分を長く保持する。一方ソフトペダルはハンマーの打弦位置をずらすことで、フェルトの硬度や弦の接触条件を変化させ、こもった丸い音色を作る。Bragagnolo ら [2] の論文ではソフトペダルを踏むと高域の倍音が抑えられてスペクトル重心が低くなることでこもったやわらかい音になると述べられている 。これらのペダルは単なる ON/OFF 操作ではなく、踏み込みの深さや踏み替えのタイミングによって音響的効果が連続的に変化する。実際の演奏では、flutter pedal のようにダンパーと弦がわずかに接触する状態を維持する微細な操作や、ペダルの踏み込み量をミリメートル単位で調整する表現が用いられることも多い。このようなペダル表現は、演奏者の感覚に強く依存しており、外部から客観的に把握することが難しい。

1.2 楽譜とペダル表現解釈

楽譜上のペダル記号は、通常ペダルの ON/OFF を示す二値的な情報として記述される。そのため、実際に演奏者がどの程度の深さでペダルを踏み込んでいるか、あるいはどのタイミングで微細な踏み替えを行っているかといった情報は、楽譜から直接読み取ることができない。また、演奏者は必ずしも記譜されたタイミング通りにペダルを操作するとは限らず、フレーズ全体の流れや響きの調整に応じて実際の操作は変化する。さらに、バロック時代のようにペダル機構が存在しなかった楽曲や、現代のポップスのようにそもそもペダルの記譜が存在しない楽譜も多く、こうした楽曲においては演奏者の感性によってペダルが自由に用いられている。そのため、既存の楽譜情報だけでは演奏中のペダル操作を再現することが難しい。

また、演奏音を聴取することでペダル操作を推定することも一つの方法であるが、これにも限界がある。たとえばサステインペダルの ON/OFF 程度であれば耳で判別可能な場合もあるが、踏み込みの深さや、他のペダルとの組み合わせによる音響変化を、人間の聴覚だけで正確に識別するのは極めて困難である。例えば、Bragagnolo ら [2] はソフトペダルの音響的解析を行っているが、F3 の音に対してソフトペダルを使用した際、6 倍音以降 (特に 4kHz 付近までの周波数帯) において、通常時よりもエネルギーレベルが低下することが報告している。このような広帯域の倍音分布の変化を人間の聴覚が定量的に把握することは難しい。筆者自身の聴覚的な印象としても、ソフトペダルの有無の違いは単音において比較的判別しやすい一方で、踏み込みが浅い場合と深い場合といった連続的なベロシティの差異については、単音を聴いただけでは明確に判断できないと感じている。このように、楽譜情報や人間の聴覚のみに基づいて演奏中のペダル操作を正確に推定することは難しい。

1.3 従来研究と課題

演奏音からピアノのペダル操作を推定する研究は、これまでにもいくつか報告されている。Liang ら [3] は、物理モデル音源によって生成した合成データを用いてニューラルネットワークを学習し、その後、実演奏音へ転移学習を行うことで、サステインペダルの踏み込みと踏み離しをフレーム単位で検出する手法を提案した。同研究では、実演奏データに対して F1 スコア 0.89 の性能を達成しており、ペダル検出における転移学習

Supervisor: Prof. Katunobu Itou

の有効性を示している。一方で、ペダル操作を二値（on/off）として扱っており、実演奏において重要となる踏み込み量の連続的な違いは推定対象としていない。

これに対し、Fang ら [1] はサステインペダルの踏み込み深さを連続値で推定するモデルを提案し、従来の二値検出では捉えられない微細なペダル操作を回帰推定で検出した。連続的なペダル深度推定という課題に取り組んだ点で重要な研究である。しかし、そのデータ作成には、MAESTRO データセットのような、内部 MIDI 情報を取得可能な特殊なピアノを用いた取得難易度の高い方法がとられている。このようなデータ取得手法は、精密なラベルを得るうえで有効である一方で、MIDI 1.0 ではペダル値の分解能に制約があることと、演奏者自身が日常的に使用しているピアノ環境を前提とした分析を必ずしも想定していないという問題がある。MIDI1.0 形式では、サステインペダルは 128 段階, ソフトペダルは 0-1 の 2 値で情報が保持される。また、実際の演奏環境では、ピアノは機種や個体差、調律状態、設置環境によって音響特性が大きく異なる。演奏者が自分のピアノで行った演奏を対象として、ペダル操作を分析することを目的とする場合、そのピアノ固有の状態を反映したデータを、同一環境下で取得できることが重要である。

また、演奏音からソフトペダルの操作量を直接推定する研究は、少なくとも筆者の調査した範囲では確認されていない。その理由として、ソフトペダルやソステヌートペダルはサステインペダルに比べて使用頻度が低く、十分な演奏データを収集することが難しい点が考えられる。また、いずれのペダルも倍音構造に影響を与えるという点で音響的効果が類似しているため、代表的なペダルとしてサステインペダルのみが研究対象とされてきた可能性もある。しかし、実際のピアノ演奏においては、複数のペダルが組み合わされることで多様な音色表現が生み出される。

よって、ピアノ練習環境で手軽に取得可能なデータを用い、サステインペダルに加えてソフトペダルの踏み込み深さを連続値として推定する枠組みを構築することが、重要な課題であると考えた。

1.4 本研究の目的と概要

本研究の目的は、演奏者自身が日常的に使用しているピアノ環境において取得可能なデータを用い、演奏音から複数ペダルの操作量を推定するシステムを構築することである。特に、サステインペダルに加えて、これまで推定研究の対象とされてこなかったソフトペダルの踏み込み深さも対象に含めて、実演奏におけるペダル表現をより包括的に捉えることを目指す。

本研究では、特別な内部センサや MIDI 情報に依存せず、カメラ撮影によるペダル機構のモーション計測を用いて連続的なペダル深度ラベルを生成する手法を採用する。この方法により、一般的なピアノ演奏環境においても、低コストかつ再現可能な形でペダル操作のラベル付与が可能となる。また、生成したラベルと演奏音を時間的に対応付けることで、演奏音からペダル深度を連続値推定する学習データを構築する。

本研究の新規性は，サステインペダル と ソフトペダル の踏み込み深さを演奏音から連続値として推定することと，ペダル機構を撮影した動画から連続的な踏み込み深さラベルを作成し，それを演奏音と対応付けた実演奏データを用いる点にある。

図 1: 提案手法における学習フェーズ（左）および推論フェーズ（右）の全体フロー

2 提案手法

2.1 手法概要

本研究では、ペダル操作条件に応じた音響的特徴を抽出し、サステインペダル、ソフトペダルの 2 種類のペダルにおける踏み込みの深さを分類により推定するモデルを構築する。

図 1 に，本研究で用いる提案手法の全体的な処理フローを示す。左側は学習フェーズ，右側は推論フェーズを表している。

学習フェーズでは，演奏データとして音声と映像を同時に収録し用意する。その後映像計測でペダル踏み込み深さの連続値ラベルを生成する。一方，音声信号からは音響特徴量を抽出する。両者を時間的に対応付けたうえでクリップ単位の学習データを構成する。これらの対応付けられた入出力データを用いて，音響特徴量を入力，ペダル深度を出力とする学習モデルを構築する。

推論フェーズでは，入力として演奏音のみを与え，学習済みモデルによって各フレームにおけるペダル踏み込み深さを推定する。

2.2 データ取得環境と収録条件

演奏データの収録は、演奏者が日常的に使用しているピアノ環境において行った。使用した楽器は一般的なグランドピアノ (ヤマハ) である。録音は防音室内で行い、温度は 26 度に一定に保つ。マイクは響板の上方に位置するように、かつ左右を一定に保つようにスタンドに固定する。録音および撮影には、ZOOM Q8n-4K を用い、音声と映像を同時に取得した。音声は WAV 形式でサンプリング周波数 48 kHz、量子化ビット数 24 bit、モノラルで記録した。映像はペダル機構を側面から撮影し、サステインペダルおよびソフトペダルの動きを明確に捉えられるように配置した (図 2)。各ペダルの可動部に、それぞれ異なる色のシールを貼付し、後述する画像処理により踏み込み量を推定するための視覚的マーカとした。

図 2: ペダル深度ラベル取得のための撮影環境．ペダル機構を側面から撮影し，サステインペダルおよびソフトペダルの可動部に異なる色のシールを貼付している．

2.3　ペダル深度ラベルの生成方法

映像はフレーム単位で処理し，各フレームにおいてマーカに対応する色領域を抽出した。抽出された領域に対して重心座標を算出し，重心位置の変位をペダル操作量として扱った。ペダル深度は正規化した。

2.4　音響特徴量の抽出

演奏音からペダル操作に関する音響的特徴を抽出するため，入力特徴量としてログメルスペクトログラムと MFCC を用意した。

両特徴量で窓幅 50ms, フレームシフト 25ms は共通で設定した。このフレームシフトの設定は，BPM 120 程度の一般的なテンポの楽曲において，16 分音符の時間長が約 125 ms であることを踏まえたものである。25 ms の時間分解能を用いることで，1 つの 16 分音符に対して約 5 フレームが対応し，演奏中のペダル踏み替えや踏み込み量の変化を時間的に十分追跡可能であると判断したため、25ms に設定した。

■ログメルスペクトログラム　ログメルスペクトログラムの算出では，音声信号をサンプリング周波数 48 kHz に統一し，各フレームに対してメルフィルタバンクを適用した。メルバンド数は 128 とし，得られたメルスペクトル M に対して $\log(M+\epsilon)$ の形で対数変換を行い，ログメルスペクトログラムを得た。

■MFCC　MFCC の算出では，音声信号を 16 kHz にリサンプルしたうえで，プリエンファシス（係数 0.97）を適用した。次に FFT で得た振幅スペクトルに対して 40 個のメルフィルタバンクを適用し，対数変換後に IDCT（Type 3）を行うことでケプストラム係数を得た。本研究では低次の 14 次元（c_0〜c_{13}）を MFCC 特徴量として用いた。

2.5　データ構成とクリップ生成

本研究では，音声・映像を同時に収録した演奏データとして，合計約 2 時間分のデータを取得した。このうち約 5 分間に相当するデータをテストデータとし，残りを学習・検証データとして 9:1 の割合で分割した。

音響特徴量およびペダル深度ラベルは，フレームシフト 25ms で時間方向に整列させた後，400 フレーム（約 10 秒）を 1 クリップとして分割した。

ペダル深度ラベルは，0〜1 に正規化された連続値として得られるが，学習の安定性を考慮し，256 段階に量子化した。各フレームは 0〜255 の整数値を持つ 256 クラスの分類ラベルとして扱う。

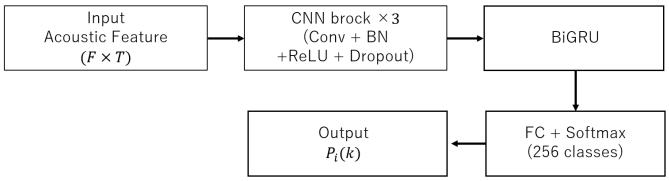

図 3: 学習モデル設計

2.6　入出力データ

入力には音響特徴量、出力には各ペダルのラベルデータを用いる。音響特徴量としては、ログメルスペクトログラムおよび MFCC の 2 種類を用意した。ログメルスペクトログラムは 128（メルバンド）×400（フレーム）の行列,MFCC は 14×400（フレーム）の行列が入力として用意される。出力は，各フレームにおけるペダル踏み込み深さであり，サステインペダル、ソフトペダルについてそれぞれ独立に推定を行う

2.7　学習モデル構成

CNN と双方向 GRU を組み合わせた時系列モデルを構築した (図 3)。基本構造は，畳み込み層の局所的な特徴抽出と，双方向 GRU による時間方向の文脈情報抽出を組み合わせた時系列モデルである。また、各畳み込み層は各畳み込み層は Conv, Batch Normalization, ReLU, Dropout から構成される。ログメルスペクトログラムを入力とする場合，周波数方向の局所的なパターンを捉えることを目的として周波数方向の畳み込みの CNN を 3 層配置した。MFCC を入力とする場合には、周波数方向の畳み込みは行わず時間方向の 1 次元畳み込みを用いた。畳み込み層で抽出された特徴系列は，双方向 GRU（BiGRU）に入力される。BiGRU の出力は，全結合層および softmax 層を通して，各フレームにおける 256 クラスの確率分布を出力する。

2.8　推定

各フレームのペダルラベル (0:1 の連続値) を 0,...255 の 256 クラスの分類ラベルに変換したものを推定する。biGRU の出力を全結合層及び softmax 層を通して各フレームにおける 256 クラスの確率分布が出力される。学習時の損失関数には、クロスエントロピーを用いた。

推論時には、softmax 出力で得られた確率分布を用いて、ペダル深度を連続値として復元する。各クラス k での連続値を

$$v_k = \frac{k}{255}$$

とし、フレーム t における推定ペダル深度 $\hat{y}_t$ を，

$$\hat{y}_t = \sum_{k=0}^{255} p_t(k)\, v_k$$

として算出する。この方法により、学習は分類問題として安定して行いつつ、推論は連続値のペダル深度を算出する。

Table 1: 各特徴量における性能（テストデータ）

ペダル	特徴量	条件	MSE	RMSE mean
Sustain	log-mel	Freq-Conv1D×3(k=3)	0.0511	0.226
Sustain	MFCC14	Time-Conv1D×3(k=15)	0.0363	0.191
Soft	log-mel	Freq-Conv1D×3(k=3)	0.0557	0.236
Soft	MFCC14	Time-Conv1D×3(k=15)	0.0407	0.202

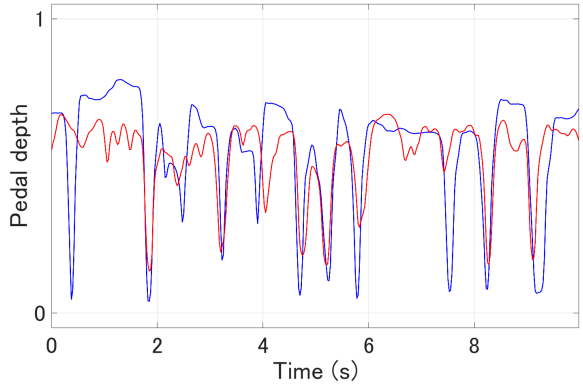

図 4: サステインペダル深度推定結果の一例（青：正解，赤：推定）

3 評価

3.1 評価方法

5 分のテストデータに対して評価を行う。評価指標には RMSE を用いる。各テストクリップについて，全フレームの誤差 $e_t = \hat{y}_t - y_t$ を計算し，

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_t^2} \quad (T = 400)$$

によりクリップ単位の RMSE を算出した。

3.2 結果

本章では，テストデータにおける推定性能を示す。評価指標には RMSE を用い，各クリップ（400 フレーム）に対してフレーム方向の平均二乗誤差を計算し，その平方根をクリップ単位の RMSE として算出した。テストデータに含まれる全クリップの RMSE の平均値を RMSE mean として算出した。

表 1 は，ログメルスペクトログラムおよび MFCC を入力特徴量として用いた場合の，テストデータにおける推定性能（RMSE mean）を示す。MFCC を用いた条件では，サステイン・ソフトの両ペダルで RMSE mean が低下し， MFCC を用いたモデルが最も良い性能を示した.

fang ら [1] のサステインペダルの連続値を推定した研究では MSE=0.0425 と報告されている。本研究で得られた MFCC での最もいい性能である RMSE=0.191 は、これを MSE に換算すると MSE=0.0363 となり、約 0.006 ほど精度が上回った。

図 4 はテストデータ中の 1 クリップに対するサステインペダルの推定結果と正解ラベルを示す。推定結果は，時間方向の大まかな変化傾向を捉えており，ペダルの踏み込みおよび戻しに対応した増減が確認できる。しかし，ペダル深度が 0.8 付近といった端の区間に対しては十分に推定できておらず、全体として中間的な値に留まる傾向が見られた。

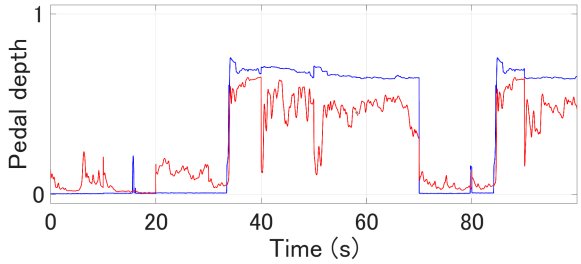

図 5: ソフトペダル深度推定結果の一例（青：正解，赤：推定）

図 5 はテストデータ中の 1 クリップに対してソフトペダルの推定値と正解値を図示したものである。大まかな時間変化は追えているものの、全体として推定値が上下に荒い傾向があった。

3.3 考察

図 4 に示した推定結果より，サステイン・ソフトペダルの両方の推定モデルはペダル踏み込み量の大まかな時間変化を追うことができていた。サステインペダルの推定では、1 に近い端の区間に対しては十分に推定できていない傾向が見られた。この要因の一つとして、学習時に用いた損失関数の設計が挙げられる。本研究では、各フレームに対してクロスエントロピー損失を等しい重み付けで学習を行っている。しかし実際の演奏データにおけるペダル深度の分布は一様ではなく、最大まで踏み込んだ状態との演奏効果は比較的出現頻度が低い。このようにクラス不均衡の状況下で損失を等しい重み付けを行ってしまっているため、頻度の高いクラスを優先的に学習し全体的に平均的な値に寄っていると考えた。また、ピアノのペダル深度がある閾値を超えた部分から効果がない可能性が考えられる。ペダルが閾値を超えて踏むと、弦とダンパーが完全に離れてしまいペダルの音響効果がなくなるからだ。

次にソフトペダルの推定では、全体として上下に荒い傾向が見られた。ソフトペダルはサステインペダルと異なり、特に音の減衰や残響などの時間方向に音響効果が出にくい操作であるため、フレームごとの予測値が荒くなったと考えた。また、今回取得したデータは自分がペダルの頻度を考慮せずに自然体で演奏したデータであるため、ソフトペダルの頻度を考慮せずに取得してしまった。そのため、今後はソフトペダルの頻度を増やしたデータ取得が必要だと考える。

次に、入力特徴量の違いについて考察する。本研究ではログメルスペクトログラムと MFCC の両方を用いて実験を行った結果、MFCC を入力としたほうが性能が高かった。ペダル操作は音の減衰特性や倍音全体のエネルギー分布、つまりスペクトル包絡の形状に影響を与える操作である。ここで、MFCC は、メルスペクトルに対して逆離散コサイン変換をすることでケプストラム係数を抽出し、そこから低次 14 次元分抽出することでスペクトル包絡の形状変化をとらえる特徴量である。よってペダル操作がもたらす音響変化と MFCC の特徴量の設計の性質が適合するため、優れた結果がでたのではないかと考える。一方、ログメルスペクトログラムは倍音成分の細かな変化や局所的な周波数構造の変化をとらえるため、演奏音高、和音構成、タッチの違いなどペダル以外の音色操作の要因にも影響を受ける。よって、ペダル操作の特徴を学習する中で不要な情報が多く含まれていた可能性がある。

今後の改善点としては、まずクラス不均衡を考慮した損失関数として、クラス頻度に基づく重み付きクロスエントロピーの導入が考えられる。また、今回 MFCC とメルスペクトログラムのそれぞれを入力として用いたが、まとめて入力とする方法も検討できる。

4　結論

本研究では、演奏者が日常的に使用しているピアノ演奏環境で手軽に取得可能な音声・映像データを用いて、演奏音からサステインペダルおよびソフトペダルの踏み込み深度を推定する手法を提案した。演奏データのうち映像から抽出したペダル深度ラベルと演奏音から抽出した音響特徴量を対応つけたデータセットを作成することで、特別なセンサや MIDI 情報を必要としない低コストなデータ取得・学習の枠組を構築した。音響特徴量に MFCC を用い、時間方向の畳み込み層 3 層と双方向 GRU を組み合わせた分類による深度推定手法により、サステインソフトペダル両方において最も良好な性能を示し、RMSE はそれぞれ 0.191, 0.202 を記録した。また、テストデータの図示により、ペダル踏み込み量の大まかな時間変化を追えていることも確認した。一方で、推定結果にはペダル深度が 1 付近といった端の区間において十分に推定できていない傾向があった。今後の課題は、ペダル深度の出現頻度を考慮した演奏データ作成に加えて頻度を考慮した損失計算で重み付きクロスエントロピーを導入することである。

参考文献

[1] Kun Fang, Hanwen Zhang, Ziyu Wang, and Ichiro Fujinaga. High-resolution sustain pedal depth estimation from piano audio across room acoustics, 2025.

[2] Alessandro Bragagnolo and Didier Guigue. Sympathetic vibration in a piano. *Journal of the Acoustical Society of America*, Vol. 147, No. 4, pp. 2462–2474, 2020.

[3] Hongru Liang, György Fazekas, and Mark Sandler. Transfer learning for piano sustain-pedal detection. In *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 2514–2518, 2018.