E-015

ピアノフレーズ練習の AI 採点のための音特徴量比較方式
Sound Features Comparison Method for AI Scoring of Piano Phrase Practice

細田 真道[1)]　内山 匡[2)]　最知 庸[1)]　小林 丈之[1)]
Masamichi Hosoda　Tadasu Uchiyama　You Saichi　Takeyuki Kobayashi
笹生 恵理[3)]　山内 竣平[4)]　野口 啓之[4)]　阪内 澄宇[1)]
Eri Sasao　Shumpei Yamauchi　Hiroyuki Noguchi　Sumitaka Sakauchi

1　はじめに

我々はピアノ教室で生徒に課す宿題の効果を大きくするため自宅練習を補助する hiketa システム [1] の実現を目指しており，そのためには指導者不在でも演奏良否が判定できる AI 採点をする必要がある．hiketa では宿題としてある程度の長さがある曲そのものの練習を課すのではなく，曲の課題要素を分解し，課題要素ごと・難易度ごとに分類した短い 4〜8 小節程度の練習フレーズを数多く用意して宿題として課す．そしてそれが「弾けた」か，あるいはどこをどのように間違ったかによって修正アドバイスを提示し次にどの課題要素・難易度の練習フレーズへ進むかを決定する．これにより継続的に達成感を得て，宿題の効果を高め効率よく曲そのものを弾けるように補助する．

生徒の自宅環境にあるピアノが電子ピアノやキーボードであれば生徒演奏を MIDI 収録して詳細に比較採点することができる [1]．しかし，古くから普及しており数が多いと思われるアコースティックピアノ（アップライトピアノやグランドピアノ）は，ごく一部に MIDI 機能を搭載したものもあるが多くはそのような機能を持たず MIDI 採点ができないという課題がある．

本稿は MIDI 非対応である通常のアコースティックピアノであっても広く AI 採点が利用できるように，生徒演奏をマイク録音して特徴量を抽出し，予め登録した模範演奏や複数の間違い演奏のモデルと比較して良否やどの間違いであったかを検出する，音による採点方式を提案する．そして有効性を確かめるため複数の環境における演奏を判定する評価実験を実施する．

2　関連研究

hiketa システムや課題要素の分解，練習フレーズ，MIDI 収録による採点方式（以下，MIDI 採点）については文献 [1] に記載している．本稿では MIDI 非対応のアコースティックピアノに対応するため，マイク録音による採点方式（以下，音採点）を提案する．音採点を実現する方法の一つとして，音を MIDI へ変換して MIDI 採点する方法が考えられる．変換方法には例えば FFT を用いた方法 [2] や，大規模なデータセットによる機械学習を用いた方法 [3] などがあり，文献 [1] の MIDI 採点は小節や拍を検出するビートトラッキングが不要なのでこれらの変換結果をそのまま使用することも可能と思われる．しかし，採点結果は変換精度の影響を大きく受ける上，精度を上げるため大規模なデータセットでの機械学習をしようとすると，データセットの用意が難しい，学習済みモデルが大きくなり推論だけでも大きな計算コストが必要になる，などの課題がある．本稿では音を MIDI へ変換せず特徴量の抽出にとどめ，広く普及しているスマートフォンやタブレット端末などでも大きな負荷をかけずに処理が可能となることを目指す．

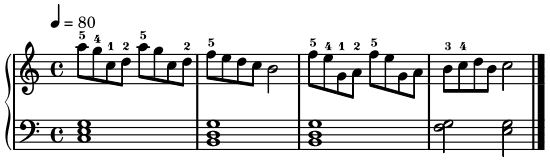


図 1　練習フレーズ「指広げ/指縮め Op. 13248」level. 12

3　練習フレーズ

図 1 に練習フレーズの一つ Op. 13248 を示す．図 2 に想定される典型的な間違い演奏を示す．これらの楽譜は MIDI 採点方式 [1] で示したものと同じである．

4　モデル

まず，MIDI 採点方式 [1] と同様，楽譜作成プログラム LilyPond [4] で譜刻（浄書）と同時に生成した Standard MIDI File（以下，SMF）[5] をモデル SMF とする．これは MIDI 採点方式と同じなので共通化が可能で，楽典やピアノ指導の知識と経験があり，練習フレーズとその典型的な間違い演奏の楽譜を作成することができれば，統計学や情報処理などのリテラシーがほとんど無くても作成することができる．

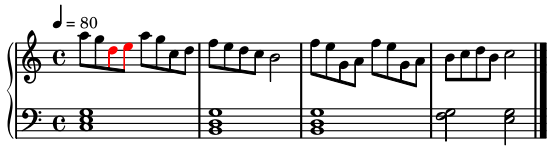
次に，音特徴量の比較元となるモデル WAV を生成する．本稿ではソフトウェアシンセサイザ TiMidity++ [6] で Fluid R3 GM を用い，音量補正有効，サンプリング周波数 44.1 [kHz]，モノラル，量子化ビット数 16 [bit]，リニア PCM としてモデル SMF からモデル WAV を得た．

5　収録

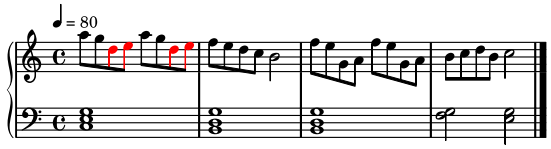
生徒演奏はスマートフォンやタブレット端末のマイクで録音する．サンプリング周波数などはモデル WAV と同様とする．手で演奏したものを録音したものなので手弾き WAV と呼ぶことにする．本稿では録音した WAV をそのまま PC へ移してその後の処理を行っているが，実際に本システムを利用する場合には，家庭内での録音をそのままクラウドなどに転送して処理することはプライバシーなどを考慮すると避けるべきと考えている．スマートフォンやタブレット端末内で音特徴量の抽出を実施し，演奏の採点はできるが元の音声を復元することは難しい状態にした上で，クラウドなどに転送して処理することは可能であると考えている．

1）東日本電信電話株式会社 デジタル革新本部 デジタルデザイン部．Digital Design Department, Digital Transformation Headquaters, NTT East Corporation.
2）NTT アドバンステクノロジ株式会社．NTT Advanced Technology Corporation.
3）株式会社 東音企画．TO-ON Kikaku Co., Ltd.
4）一般社団法人 全日本ピアノ指導者協会（ピティナ）．The Piano Teachers' National Association of Japan.

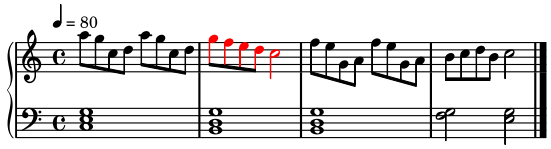
■ミス A: 1 小節 2 拍目「レミ」と弾いてしまう



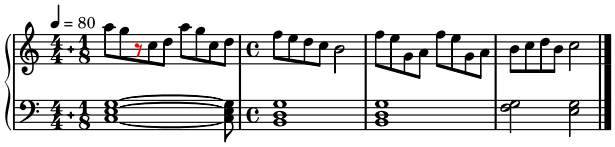
■ミス B: 1 小節 2，4 拍目「レミ」と弾いてしまう



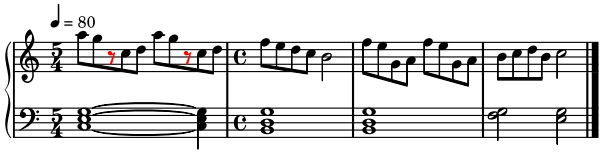
■ミス C: 2 小節目「ソファミレド」と弾いてしまう



■ミス D: 1 小節 2 拍目がつっかえる



■ミス E: 1 小節 2，4 拍目がつっかえる



■ミス F: 2 小節 3 拍目がつっかえる（指が足りない）

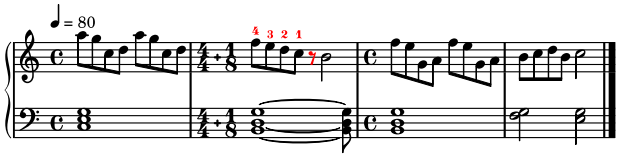


図 2　想定される典型的な間違い演奏（1，2 小節）

6　採点方式

WAV 同士の距離を求め，評価対象となる手弾き WAV を，模範演奏および各間違い演奏のモデル WAV とそれぞれ比較し，最も距離が近かったモデル WAV の演奏をしたものとして判定する採点方式を提案する．

6.1　特徴量抽出

短時間フーリエ変換（Short-Time Fourier Transform，以下 STFT）のような処理で，WAV にウィンドウサイズ t_w [s] のウィンドウ関数を，スライディングサイズ t_s [s] ずつずらしながらかけて周波数解析し，得られた周波数成分（パワー・スペクトル）のベクトル $\boldsymbol{f}_i$ を t_s 間隔の時系列 $(i = 1, 2, 3, \ldots)$ で並べたものを特徴量として使用する．通常の STFT で用いられる周波数解析は離散フーリエ変換（Discrete Fourier Transform，以下 DFT）の一種である高速フーリエ変換（Fast Fourier Transform，以下 FFT）であるが，DFT や FFT で得られる周波数成分は周波数が等間隔の等差数列になっており，その間隔，つまり周波数分解能は t_w の逆数なので，周波数分解能と時間分解能はトレードオフの関係と言える．一方，音楽で通常用いられる音高，つまり 12 平均律による半音階は，半音上がるたびに周波数が $\sqrt[12]{2}$ 倍となる等比数列になっている．そのため低音域では分解能が足りず，高音域では分解能が高すぎることになる．分解能が高すぎる方については，メルフィルタバンクをメル尺度の代わりに半音階の周波数に置き換えたような処理を行えばまとめることができるが，分解能が足りないものを補間することはできない．

一方で音楽の分析に使う周波数分析手法として定 Q 変換（Constant-Q transform，以下 CQT）[7] がある．これは半音階の周波数で周波数成分を計算しつつ t_w を周波数に応じて低音域では長く，高音域では短くしたものである．半音階の周波数成分が直接得られるので音楽の分析に使いやすいと言えるが，特に低音域で t_w が長くなることによって計算コストが大きくなってしまう．

そこで CQT のように半音階の周波数で周波数成分を計算するが，DFT のように t_w をすべての周波数で固定にして計算する方法 [8] を採る．この方法は通常の DFT などと異なり周波数成分から元の音声を得る逆変換ができないが，採点には必要ないので問題ない．本稿では IPN* 表記の A4 を 440 [Hz]，ウィンドウ関数を blackman，$t_\mathrm{w} = 0.1$ [s]，$t_\mathrm{s} = 0.05$ [s]（つまりウィンドウを半分ずつ重ねる）とする．なお，予備実験の結果，高音域の強度が弱かったため MIDI ノート番号 i の強度を $2^{i/12}$ 倍になるよう補正する．

図 3 に Op. 13248 の正しい楽譜から生成した模範演奏モデル WAV について，この方法で抽出した特徴量を可視化したスペクトログラムを示す．縦軸は音高で MIDI ノート番号（0〜127），横軸は時刻，輝度は強度 0 を黒，最大を白として $\gamma = 0.45$ でガンマ補正したものである．同様に図 4 にミス A モデル WAV，図 5 に正しい楽譜を演奏して録音した手弾き WAV，図 6 にミス A 楽譜を演奏して録音した手弾き WAV のスペクトログラムと，ミス A で誤った音符の部分を矢印で示す．いずれも低音域では周波数分解能が足りず，音符のある音高の上下に強度が漏れてぼやけている．中音域以上では周波数分解能が足りるため上下の漏れは無くなるが，音符のある音高ではない倍音成分が目立つようになってくる．低音域の分解能不足も中音域以上の倍音も MIDI 変換するならば課題となるが，本稿のようにモデル WAV と手弾き WAV を比較するのであれば双方の WAV で同様に表れるものを比較すればよいため課題にはならない．

6.2　比較

2 つの時系列ベクトルを Dynamic Time Warping（以下，DTW）という DP マッチングによる方法で比較する．これは 1980 年代の音声認識で使用されていたもので，2 つの時系列ベクトルの時間軸方向の伸縮やズレを許容した各時刻の対応関係，つまり片方のある時刻がもう片方のどの時刻に対応しているかの関係（以下パス）と，2 つの時系列ベクトル全体同士の距離（以下，DTW 距離）を求めることができる [9, 10]．

DTW を計算するには，2 つの時系列ベクトルについ

* International Pitch Notation

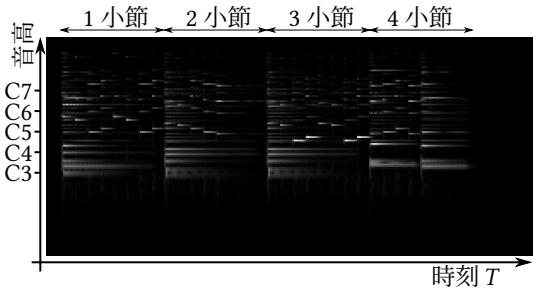


図 3　模範演奏モデル WAV のスペクトログラム

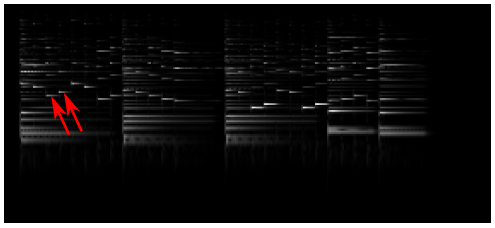


図 4　ミス A モデル WAV のスペクトログラム

て，片方の任意の時刻ともう片方の別の任意の時刻との間の距離を計算できる必要がある．この点と点の距離のことをコストと呼ぶことにする．ここでモデル WAV の特徴量を $\boldsymbol{f}_{\text{model}\,i}$，$(i = 1, 2, 3, \dots)$，手弾き WAV の特徴量を $\boldsymbol{f}_{\text{foreval}\,j}$，$(j = 1, 2, 3, \dots)$ とする．

まず音色の違いによるコストとして，周波数成分が類似していればコストを低く，そうでなければ高くすることを考える．単純にユークリッド距離を使う方法も考えられるが，双方で録音環境が異なり音量に大きな差がある可能性があり，ユークリッド距離ではこの影響を受けてしまう．そこで強度を無視してベクトルの類似度を計算できるコサイン類似度を使用し，音色コスト $C_{\text{timbre}\,i,j}$ を以下の式で計算する．

$$C_{\text{timbre}\,i,j} = 1 - \frac{\boldsymbol{f}_{\text{model}\,i} \cdot \boldsymbol{f}_{\text{foreval}\,j}}{|\boldsymbol{f}_{\text{model}\,i}||\boldsymbol{f}_{\text{foreval}\,j}|} \tag{1}$$

次に音量の違いによるコストとして，音量が近ければコストを低く，そうでなければ高くすることを考える．録音環境の違いによる音量の差を補正するため，無音が 0，最大音量が 1 になるように正規化した上で計算する．モデル WAV と手弾き WAV の音量（強度）を $P_{\text{model}\,i} = |\boldsymbol{f}_{\text{model}\,i}|$，$P_{\text{foreval}\,j} = |\boldsymbol{f}_{\text{foreval}\,j}|$ で計算し，それぞれの最大値を $P_{\text{model\_max}} = \max_{i=1,2,3,\dots} P_{\text{model}\,i}$，$P_{\text{foreval\_max}} = \max_{j=1,2,3,\dots} P_{\text{foreval}\,j}$ として，音量コスト $C_{\text{power}\,i,j}$ を以下の式で計算する．

$$C_{\text{power}\,i,j} = \left| \frac{P_{\text{foreval}\,j}}{P_{\text{foreval\_max}}} - \frac{P_{\text{model}\,i}}{P_{\text{model\_max}}} \right| \tag{2}$$

そして，無音の考慮をする．双方とも無音に近い場合はコストも低くすべきだが，$C_{\text{timbre}\,i,j}$ はコサイン類似度を用いており，ほぼ無音同士でもベクトルの向きが異なると非常に大きくなる．そこで，そういった場合にコストを小さくできる項を作るため，双方の平均 $M_{i,j}$ を以下の式で計算する．これに $0 < \alpha_{\text{weak}} < 1$ となる指数を用いれば無音に近くなった時は急激に小さな値にできる．

$$M_{i,j} = \frac{1}{2}\left(\frac{P_{\text{model}\,i}}{P_{\text{model\_max}}} + \frac{P_{\text{foreval}\,j}}{P_{\text{foreval\_max}}} \right) \tag{3}$$

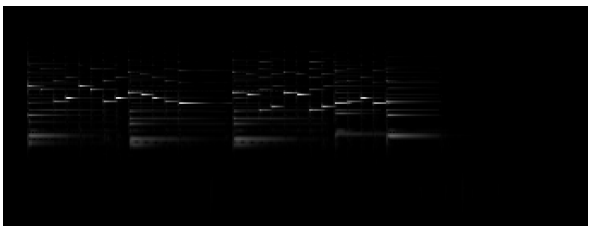


図 5　正しい楽譜の手弾き WAV スペクトログラム

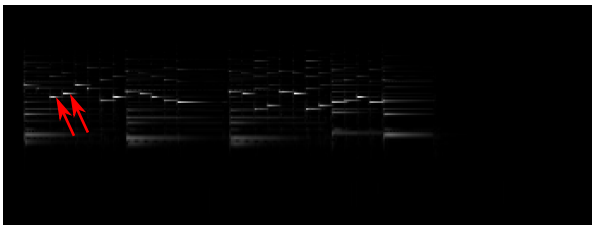


図 6　ミス A 楽譜の手弾き WAV スペクトログラム

以上 3 つの要素と調整用の係数・指数を組み合わせてコスト $C_{i,j}$ を以下の式で計算する．

$$C_{i,j} = \left(a_{\text{timbre}} C_{\text{timbre}\,i,j}{}^{\alpha_{\text{timbre}}} + a_{\text{power}} C_{\text{power}\,i,j}{}^{\alpha_{\text{power}}} \right) M_{i,j}{}^{\alpha_{\text{weak}}} \tag{4}$$

6.3　パス

図 7 に模範演奏モデル WAV と正しい楽譜を演奏した手弾き WAV を DTW で比較して得られたパスを示す．縦軸がモデル側時刻，横軸が手弾き側時刻で，パスとして互いの対応している時刻同士を線で結んで可視化した．また，パス上にモデル SMF のノート ON 時刻を丸いマーカーで示した．パスは一部直線になっていないところがあるが，マーカー同士の傾きはほとんど一定であり，ほぼ直線上にマーカーが並んでいる．つまり，打鍵（ノート ON）タイミングの伸び縮みはほぼ一定でありタイミングの問題は無かったことがわかる．

7　評価実験

提案方式の有効性を確かめるため，複数の環境で生徒演奏を模した演奏を収録し判定する評価実験を実施する．

7.1　収録

環境は楽器 4 種類（グランドピアノ・アップライトピアノ・電子ピアノ–重い鍵盤・キーボード–軽い鍵盤）と空調有無 2 種類の組み合わせとした．これらに対して Op. 13248 の正しい楽譜とミス A〜F 楽譜の 7 種類の楽譜をそれぞれ 3 回ずつ演奏した．録音端末は 6 種類（Android スマートフォン 3・Android タブレット 1・iPhone 1・iPad 1）として 6 台同時に録音した．つまり $4 \times 2 \times 7 \times 3 \times 6 = 1008$ 個の手弾き WAV を録音した．

7.2　パラメータ

係数などは以下の設定とした．

$$a_{\text{timbre}} = 1, \qquad \alpha_{\text{timbre}} = 1,$$
$$a_{\text{power}} = 0.25, \qquad \alpha_{\text{power}} = 1,$$
$$\alpha_{\text{weak}} = 0.2$$

7.3　全モデルとの比較

1008 個の手弾き WAV を，模範演奏モデル WAV とミス A〜F モデル WAV の全 7 種類と比較し，どのモデル WAV との DTW 距離が最も近くなるか評価した．

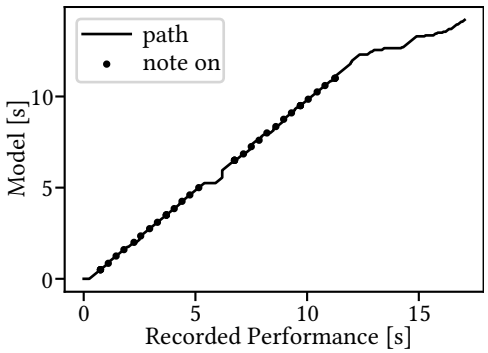


図 7　模範演奏モデルと正しい楽譜演奏のパス

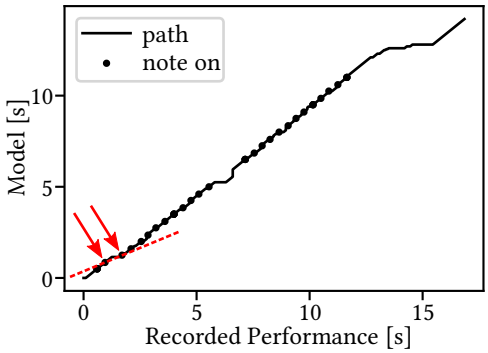


図 8　模範演奏モデルとミス D 演奏のパス

7.3.1　結果

1008 個中 774 個（約 77 %）の手弾き WAV が正解（演奏に使用した楽譜で生成したモデル WAV との DTW 距離が最も近い）となり，残り 234 個（約 23 %）の手弾き WAV を誤判定した．

7.3.2　考察

正解率は約 77 %で，誤判定の多くはミス D，E，F の手弾き WAV であった．ミス A，B，C は模範演奏と構成音が異なるところがある「構成音ミス」なのに対して，ミス D，E，F は構成音が異なる瞬間が無く，タイミングだけが異なる「タイミングミス」であるという違いがある．DTW は時間軸方向の伸縮やズレを許容して比較するアルゴリズムであるため，タイミングミスの判定が困難であったと考えられる．

そこで，判定用モデル WAV をタイミングミスを除いた 4 種類に限定し，DTW 距離が模範演奏モデル WAV と最も近くなったら，パスから模範演奏かタイミングミスかを判断する方法が考えられる．

7.4　限定モデルとの比較

模範演奏モデル WAV とミス A，B，C モデル WAV の 4 種類に限定して比較し，どのモデル WAV との DTW 距離が最も近くなるか評価した．ミス D，E，F のタイミングミス手弾き WAV については模範演奏モデル WAV との距離が最も近くなった場合に正解とする．

7.4.1　結果

1008 個中 998 個（約 99 %）の手弾き WAV が正解となり，残り 10 個（約 1 %）の手弾き WAV を誤判定した．

7.4.2　考察

図 8 に模範演奏モデル WAV とミス D 手弾き WAV を DTW で比較して得られたパスとノート ON 時刻，矢印で示す 2 つのマーカー同士の傾きを破線で示す．図 7 で示した正しい楽譜演奏の場合と異なり，破線で示した傾きが他より小さくなっている．つまりこの部分の打鍵（ノート ON）間隔が手弾き WAV 側で大きくなってしまっていることがわかる．これはミス D と同じ特徴なので，これを検出することでミス D か否かの判定が可能と思われる．同様にミス E，F や他のタイミングミスについても，どのマーカー同士の傾きが他よりも小さいか大きいか検出することで判定が可能と思われる．

未登録の未知ミスもタイミングのミスであればマーカー同士の傾きで検出できるのではないかと考えられる．構成音の未知ミスは同様に傾きが乱れるならば検出できるし，他にもパス上のコストが局所的に高くなれば検出できるのではないかと考えられる．また，MIDI 採点と同様，許諾を得て手弾き WAV の特徴量を収集できれば，クラスタリングによって想定していなかった典型的ミスを発見して登録することも可能と考えている．

8　おわりに

本稿は MIDI 非対応のアコースティックピアノであっても広く AI 採点が利用できるように，生徒演奏をマイク録音して特徴量を抽出し，予め登録した模範演奏や複数の間違い演奏のモデルと比較して良否やどの間違いであったかを検出する，音による採点方式を提案した．そして有効性を確かめるため複数の環境における演奏を判定する評価実験を実施し，構成音ミスであればほぼ検出可能，タイミングミスは後処理を追加することにより検出可能となる見込みがあることを示した．

今後は，タイミングミス判定の追加，練習フレーズや間違い演奏モデルを増やし手弾き WAV の収録を進めて採点方式の改良やパラメータ調整などを進めたい．

参考文献

[1] 細田真道，最知庸，小林丈之，笹生恵理，山内竣平，野口啓之，阪内澄宇：ピアノ宿題練習のための AI 採点方式，FIT2022（第 21 回情報科学技術フォーラム），No. CE-007 (2022).
[2] 市來健吾：音楽と数理 才能にたよらない耳コピ，ZENKEI AI FORUM (2020). 技術書典 9.
[3] Hawthorne, C., Simon, I., Swavely, R., Manilow, E. and Engel, J.: Sequence-to-Sequence Piano Transcription with Transformers, *Proc. of the Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 246–253 (2021).
[4] LilyPond 開発チーム：LilyPond —みんなの楽譜作成,（オンライン），入手先 ⟨https://lilypond.org/⟩（参照 2022-06-06）.
[5] （社）音楽電子事業協会：4. スタンダード MIDI ファイル 1.0，MIDI1.0 規格書（オンライン），入手先 ⟨https://amei.or.jp/midistandardcommittee/MIDIspcj.html⟩（参照 2022-06-06）.
[6] 出雲正尚：TiMidity++, (online), available from ⟨http://timidity.sourceforge.net/⟩ (accessed 2022-06-06).
[7] Brown, J. C.: Calculation of a constant Q spectral transform, *J. Acoust. Soc. Am.*, Vol. 89, No. 1, pp. 425–434 (1991).
[8] 松岡保静，渡部瑞季：音楽のコード認識技術とその応用，NTT DOCOMO テクニカルジャーナル， Vol. 25, No. 2 (2017).
[9] 河原達也：音声認識技術の変遷と最先端—深層学習による End-to-End モデル—，日本音響学会誌， Vol. 74, No. 7, pp. 381–386 (2018).
[10] 渡辺隆夫：音声認識におけるパタンマッチングの手法，日本音響学会誌， Vol. 42, No. 9, pp. 725–730 (1986).