

運指を考慮したピアノ演奏ジェスチャーからの リアルタイム音高系列推定手法

三田 ちひろ^{1,a)} 渡邊 研斗² 中塚 貴之² 中野 倫靖² 後藤 真孝² 伊藤 貴之^{1,b)}

概要：本研究では、実体のある鍵盤を必要としないピアノ演奏を可能にする仮想ピアノの実現を目的として、指の動きと運指情報に基づいて音高系列を確率的に推定する手法を提案する。本手法により、専用の鍵盤装置を用いず、手の動作のみでピアノ演奏を行う環境の構築が可能となる。従来の仮想鍵盤システムでは、鍵盤位置の正確な認識のために視線追跡や物理的な触覚フィードバックが必要であり、簡易な機材による実現が困難であった。本研究ではこの課題を解決するために、指の打鍵位置の変化と運指の遷移情報を入力として、音高系列を逐次的に推定する確率モデルを構築した。提案手法は、 N -gram に基づく音高系列の言語モデルと、指間距離および運指の組み合わせに基づく翻訳モデルを統合したものであり、指の動きから高精度な音高系列推定を可能にする。さらに、応用例として、MediaPipe Hands を用いたハンドトラッキングと提案モデルを統合し、手の動きのみで演奏可能な仮想ピアノを開発した。

1. はじめに

近年、VR 技術やジェスチャーインタフェースの発展により、物理的な鍵盤を用いずにピアノ演奏を行う仮想鍵盤システムが注目されている。しかし、これらの多くは、打鍵位置を正確に認識するために視覚フィードバックや触覚デバイスが必要としており、さらに演奏者が鍵盤を視認できるように、鍵盤に相当する視覚的表示をするための専用の機材や高度な環境構築を伴うものが多い [1-3]。そのため、日常的な利用や簡易な実装には適さず、ユーザへの負担も大きいという課題がある。そこで、本研究では演奏者が鍵盤を視認しないで演奏できる仮想ピアノ演奏の実現を目指す。

通常のピアノ演奏では、演奏者は視覚的に鍵盤の位置を把握しながら打鍵することで高い演奏精度を実現している。特に初見演奏では、視覚フィードバックが演奏精度に与える影響は大きいとされている [4,5]。

一方で、実際の演奏環境では常に鍵盤を注視できるとは限らない。読譜や共演者との視線のやり取りなど、視覚的な負荷は多岐にわたる [6-8]。このように、視線の配分にはトレードオフが存在し、鍵盤を視認できない状況では演奏精度が低下するという課題がある。

本研究ではそうした課題を踏まえて、実体のある鍵盤を必要としない仮想ピアノ演奏を実現させるために、視覚に依存せず音高を選択できる支援手法を提案する。具体的には、打鍵位置の変化と運指情報をもとに、音高系列を確率的に推定する統計的機械翻訳 (SMT) に基づくモデルを構築した。さらに、本手法をスマートフォンでの撮影によるリアルタイムハンドトラッキング技術 (MediaPipe Hands) と組み合わせることで、手の動きのみで演奏可能な仮想ピアノ「InvisiblePiano」を構築した。InvisiblePiano は右手の動きのみで操作可能であり、単音かつ白鍵での演奏に対応している。

本手法は、従来のピアノ演奏支援におけるミスタッチ修正 [9,10] や深層学習ベースのジェスチャー演奏技術 [11] とは異なる特徴を有する。従来の LSTM や Transformer に基づくミスタッチ修正手法は、MIDI 楽器など専用ハードウェアを前提とするものが多い [9,10]。一方、本手法はカメラによるジェスチャー入力のみを用いてリアルタイムに音高系列を推定可能であり、スマートフォン等の汎用デバイス上で動作する点において利便性が高い。

また、近年登場している深層学習によるジェスチャー演奏の研究では、膨大なアノテーションデータや計算資源が必要であること、推定根拠の可視化が困難であること、演奏可能な音高に制限があることなどが課題とされている [11]。これに対して、本研究では SMT に基づくモデルを採用しており、少量データでの推定が可能で、推定結果

¹ お茶の水女子大学

² 産業技術総合研究所

^{a)} g2120539@is.ocha.ac.jp

^{b)} itot@is.ocha.ac.jp

の解釈性にも優れる。

本研究では、以下の2つの実験を通して提案手法の有効性を検証した。

- 実験1：開発した SMT モデルの音高系列推定精度をテストデータ上で定量的に評価（非リアルタイム）
 - 実験2：InvisiblePiano を用いたユーザ実験（ $N = 18$ 名）を通じて、視覚非依存型演奏の実現可能性を評価
- 本研究の成果により、鍵盤の有無にかかわらず音高推定と演奏が可能となり、演奏支援や楽譜生成、ジェスチャーによる音楽検索、エア楽器演奏など、幅広い応用が期待される。特に、楽器の用意が困難な教育現場や高齢者施設、さらには楽譜を用いず演奏動作指導によって伝承される民族音楽の記録支援など、将来的な展望も広がる。

本研究の主な貢献は以下のとおりである。

- ピアノ演奏時の打鍵位置変化と運指情報から音高列を推定する統計的機械翻訳（SMT）に基づく確率モデルを新たに提案した。
- 手の動きのみで演奏可能な仮想ピアノ「InvisiblePiano」を開発し、楽器や鍵盤を必要としないピアノ演奏を実現した。
- 提案手法の音高推定精度を非リアルタイム環境で定量的に評価するとともに、18 名の実験参加者を対象とした InvisiblePiano のユーザ実験を通して、視覚に依存しない演奏支援効果を実証した。

2. 関連研究

ミスタッチ修正、仮想鍵盤システム、ジェスチャー演奏の3つの項目に分けて関連研究を紹介する。

2.1 ミスタッチ修正

本研究は、従来のピアノ演奏におけるミスタッチ修正を主眼とした研究とは異なり、物理的鍵盤を必要としない仮想的演奏環境の構築を目的とする点において新規性を有する。ミスタッチ検出モデルとしては Chou ら [9] による Transformer ベースの非リアルタイムシステムの Polytune や、Marinov ら [10] による RNN によるリアルタイムモデルが挙げられる。これらは、主に MIDI 入力を通じたミスタッチ検出・修正を目的としており、高い精度を達成しているが、いずれも物理的鍵盤の操作を前提としている。

一方で本研究は、スマートフォンカメラを用いた手指ジェスチャーの認識に基づいて音高系列を推定するモデルを提案しており、鍵盤の物理的存在を前提としない。この点において、既存研究とは異なる対象領域に貢献するものである。

2.2 仮想鍵盤システム

近年、VR 技術やジェスチャーインタフェースの発展により、物理的な鍵盤を用いずにピアノ演奏を行う仮想鍵盤

システムが注目されている。しかし、これらの多くは打鍵位置を正確に認識するために視覚フィードバックや触覚デバイスを必要としており、ディスプレイや専用機材など高度な環境構築を伴うものが多い [1–3]。そのため、場所の制約にとらわれない日常的な利用や簡易な実装には適さず、ユーザへの負担も大きいという課題がある。そのような問題を解決するために、Begum ら [12] は、紙に描かれた鍵盤をスマートフォンカメラで認識し、指の位置を画像処理によって特定することで音高を決定するシステムを提案した。また、Wang ら [13] もピアノ演奏を簡易的に実現するために単眼の逆向きカメラとピアノ鍵盤の模様を描いた透明な板を用いたバーチャルピアノシステムを提案した。単眼の逆向きカメラを使用することで従来の画像処理での仮想ピアノシステムで課題となっていた手の相互被り、大きな深度検出誤差、リアルタイム性能の低さを改善した。しかし、Begum ら [12] も Wang ら [13] も演奏者に鍵盤を視認させるために鍵盤に相当するものを用意する必要がある。その点、本手法は演奏者が鍵盤を視認する必要がないために、鍵盤に相当するものを用意する必要がない。また、スマートフォンを利用し、かつジェスチャーのみを入力とすることで、簡易なピアノ演奏環境の構築が可能である点で優位である。

2.3 ジェスチャー演奏

近年では深層学習によりジェスチャー演奏を実現する研究も見られる。例えば、Gusita ら [11] は手首を基準とした指の位置関係を識別し、対応する音高を判別する3層の全結合ニューラルネットワークを鍵盤とするモデルを提案した。シンプルなネットワークを採用することで、計算負荷を抑えることができ、リアルタイムでの音高推定が可能となった。しかし、このモデルは学習において大規模なアンテーション付きデータと高い計算資源を必要とするうえ、推定結果の解釈性が低いという課題がある。また、対応可能な音高数は10音に限られていた。これに対して、本研究では、統計的機械翻訳（SMT）に基づく確率モデルを採用しており、学習に必要なデータ量が比較的少ないことに加え、実用的な環境への適用が可能であり、推定の根拠も解釈しやすいという特長を持つ。

さらに、ジェスチャーを入力する仮想鍵盤楽器の例として Yang ら [14] の MuGeVI が挙げられる。同システムでは、MediaPipe Hands を用いてハンドサインをリアルタイムで識別し、ハンドサインに対応する音や和音を出力する。しかし、MuGeVI の演奏動作は既存の楽器における演奏動作とは異なる。アコースティック楽器とデジタル楽器を比較した際、身体性に違いがあり、デジタル楽器は身体性が低く愛着が湧きにくいと指摘されている [15]。一方、本手法では触覚フィードバックはないものの、アコースティック楽器と同じ演奏動作ジェスチャーを入力とする点で先行

研究と比べると身体性が高いと言える。

3. 提案手法

本研究では、実体のある鍵盤を必要としないピアノ演奏を目的として、演奏時のジェスチャー情報（打鍵位置変化および運指情報）から音高系列を推定する確率モデルを提案する。

本提案手法では、演奏時のジェスチャーの入力系列から対応する音高系列への変換問題を、いわゆる sequence-to-sequence タスクと位置づけ、ある言語（単語系列）から別の言語（単語系列）へ変換する機械翻訳手法の枠組みを応用し、演奏時のジェスチャーに対応する音高系列への「翻訳」タスクとして定式化した。

なお、本手法は単旋律（モノフォニックな旋律）を対象としており、重音（同時に複数の音高を含む和音的構造）の推定には対応していない。また、片手のみかつ白鍵の演奏を対象とし、両手での演奏や黒鍵の演奏は対応していない。これは、モデルが時系列的に1つの音高のみを出力する設計となっているためであり、多声的な演奏データを扱うには別途、並列的な出力構造やマルチラベル化が必要となる。重音推定への対応は、今後の拡張課題とする。

3.1 問題設定

本手法の全体構成を図1に示す。入力として与えられるのは、ピアノ演奏動作における打鍵位置変化 (Δg) および運指情報 (F) である。これらの情報は、2打鍵間における指の移動ベクトルおよび使用された指番号の系列として記録される。詳細は3.2節で後述する。

本モデルでは、初期音高 p_0 が与えられていると仮定し、それ以降の系列長を T とし、音高列 $H = (p_0, p_1, \dots, p_n, \dots, p_T)$ の音高 $p_n (n \geq 1)$ を逐次的に推定する。

3.2 入力情報の定義

提案手法の入力情報である打鍵位置変化 (Δg) および運指情報 (F) を定義する。これらの情報は、演奏者の動作系列から得られるものであり、音高列の推定における翻訳モデルの入力として用いられる。

3.2.1 打鍵位置変化 Δg

打鍵位置とは、演奏者がある時刻に鍵盤を打鍵したときの指先の空間的位置を指す。この位置は、カメラ映像に基づくハンドトラッキング手法（例：MediaPipe Hands）を用いて検出される3次元座標 (x, y, z) の x 座標のみで表現される。 x 座標のみを採用した理由は、鍵盤が x 軸に水平に配置されていることを前提とし、水平方向の位置変化のみが打鍵の音高に影響を与えると仮定したからである。

時刻 n における打鍵位置を $g_n \in \mathbb{R}^1$ 、時刻 $n-1$ における打鍵位置を $g_{n-1} \in \mathbb{R}^1$ とすると、打鍵位置変化ベクトル Δg_n は次式で定義される。

$$\Delta g_n = g_n - g_{n-1} \quad (1)$$

このベクトルは、 x 座標において直前の打鍵から次の打鍵までに指がどの方向にどれだけ動いたかを示す空間的特徴量である。鍵盤上で右方向へ移動すれば正、左方向なら負となる。

3.2.2 運指情報 F

親指から小指までの指番号を意味する1~5の整数値のペアで、運指情報 F を表す。例えば、 $F = (1, 2)$ は人差し指から親指への移動を、 $F = (1, 1)$ は親指の連続打鍵を表現する。時刻 n における連続する2打鍵間における指番号のペア (2-gram) F_n は次式で定義される。

$$F_n = (f_n, f_{n-1}) \quad (2)$$

3.2.3 前提条件

本モデルでは、系列の初期状態として時刻 $n=0$ における初期音高 p_0 を与える。また、時刻 $n \geq 1$ においては、それ以前の時刻までの系列 $\{p_x\}_{x=n-N}^{n-1}$ と、対応する打鍵位置変化 Δg_n と指番号ペア F_n をもとに、次の音高 p_n を逐次的に推定することが可能となる。

3.3 翻訳モデルと音高言語モデルの構成

本手法では、SMT (Statistical Machine Translation) に基づき、打鍵位置変化や運指情報といった身体的ジェスチャーから音高系列を推定するモデルを翻訳モデル (3.3.1節) と音高言語モデル (3.3.4節) を組み合わせて構築する。

3.3.1 翻訳モデル $P(p_n | \Delta g_n, F_n, p_{n-1})$

翻訳モデルは、打鍵位置変化 (Δg_n) および連続する指番号ペア (F_n) を入力として、対応する音高変化 ($\Delta p_n = p_n - p_{n-1}$) の確率分布を推定するモデルである。ここで、直前の音高 p_{n-1} が既知であると仮定し、時刻 n における音高 p_n の条件付き確率を以下で定義する。

$$P(p_n | \Delta g_n, F_n, p_{n-1}) := P(\Delta p_n = p_n - p_{n-1} | \Delta g_n, F_n) \quad (3)$$

この確率分布は、ベイズの定理を用いて以下のように近似できる。

$$P(\Delta p_n | \Delta g_n, F_n) \propto \underbrace{P(\Delta g_n | \Delta p_n)}_{\text{動作モデル}} \cdot \underbrace{P(\Delta p_n | F_n)}_{\text{運指モデル}} \quad (4)$$

ここで、 $P(\Delta g_n | \Delta p_n)$ は本来は $P(\Delta g_n | \Delta p_n, F_n)$ だが、 F_n に依存しないと考えて近似した。 $P(\Delta g_n | \Delta p_n)$ は音高変化に対して打鍵位置変化がどの程度一致するかを表す「動作モデル」、 $P(\Delta p_n | F_n)$ は運指情報に基づく音高変化の出現傾向を表す「運指モデル」と呼ぶ。本手法では、動作モデルを正規分布で近似し、運指モデルは学習データから統計的に構築する。

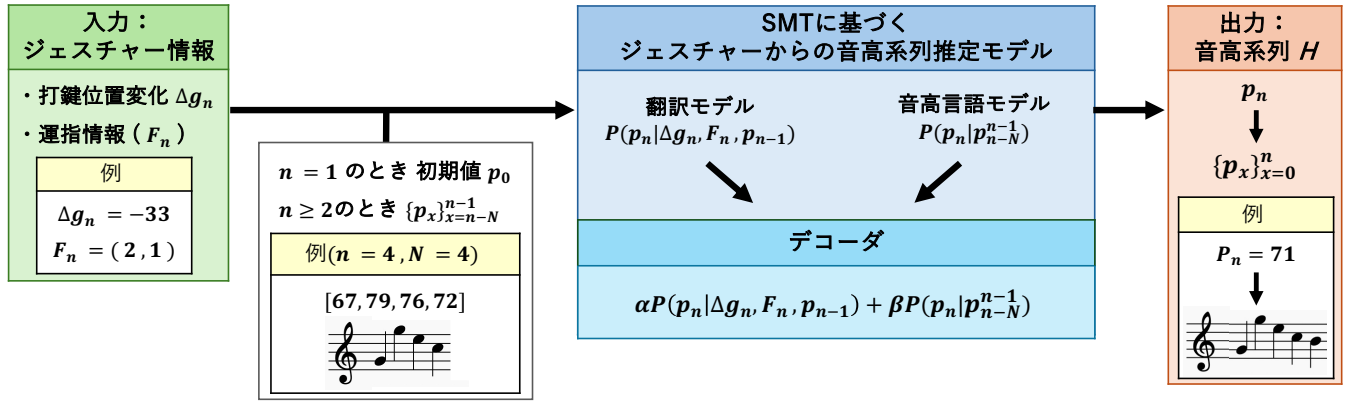


図 1 SMT に基づく音高推定モデルの全体構成図 (n : 時刻, N : N -gram モデルの文脈長)

3.3.2 動作モデル $P(\Delta g_n | \Delta p_n)$

Yonebayashi ら [16] は、隠れマルコフモデルを用いた自動運指推定モデルを構築する際、鍵盤の移動と指運びは正規分布で近似した。従って、本手法でも動作モデル $P(\Delta g_n | \Delta p_n)$ を正規分布で定義し、打鍵位置変化 Δg_n の生成過程を確率的に表現する。具体的には、各音高変化 Δp_n に対して、それに対応する打鍵移動距離（設定する鍵盤幅から算出）を平均値 $\mu_{\Delta g_n}$ とする正規分布で、 Δg_n の分布を定義する。

$$P(\Delta g_n | \Delta p_n) = \mathcal{N}(\Delta g_n | \mu_{\Delta g_n}, \sigma_{\Delta g_n}^2) \quad (5)$$

ここで、標準偏差 $\sigma_{\Delta g_n}$ は移動量の大きさに応じて以下のように設定した。白鍵 1 個分の幅を W とする。

- $|\Delta g| \leq W$: $\sigma_{\Delta g_n} = 0.5W$
- $W < |\Delta g| \leq 4W$: $\sigma_{\Delta g_n} = W$
- $4W < |\Delta g| \leq 7W$: $\sigma_{\Delta g_n} = 1.5W$
- 上記以外 (1 オクターブより大きい): Δg_n は 1 オクターブ分の鍵盤幅にクリップされ、 $\sigma_{\Delta g_n} = 1.5W$ とする

これは、視覚的な情報が得られない条件下において、音高遷移の大きさに応じて打鍵位置の予測が困難になる傾向があるという仮定して設定した。例えば実際は C4 から E4 への移動を意図している、 $\Delta g_n = 4W$ のような移動になることが考えられるので、鍵盤幅感覚の乱れを考慮してモデル化した。この設定は経験的には妥当と考えられるが、最適な設定の検証は今後の課題である。

なお、本研究における「白鍵 1 個分の幅」は、演奏環境や記録方法に応じて自由に設定可能なパラメータとした。これは、ピクセル単位の画像座標を用いるスマートフォン映像による推定や、mm 単位での実測座標を用いる高精度センサによる推定など、様々な入力環境への適用を考慮したためである。加えて、従来の物理的な鍵盤幅に捉われることなく、演奏者の身体的特徴に応じて、より適した鍵盤幅を柔軟に設定できるという利点もある。ここでは白鍵についてしか議論していないが、黒鍵を考慮する動作モデル

の構築は今後の課題とする。

図 2 は、動作モデルにおける正規分布の例である。本手法では正規分布での近似を用いたが、Gan ら [17] の楽譜データから演奏ジェスチャーの 3D 動画を生成するモデルを用いて演奏ジェスチャーのデータを収集し、それに基づいて統計的に動作モデルを構築できる可能性もあり、今後は正規分布以外への拡張の余地がある。

3.3.3 運指モデル $P(\Delta p_n | F_n)$

手の構造的制約を考慮してモデル化するために、連続する 2 つの指番号からなる 2-gram の運指ペアに基づく音高変化 Δp_n の確率分布を、運指モデルとして構築した。具体的には、全 25 通りの運指ペア $F_n = (f_n, f_{n-1})$ に対して、学習データ中における音高変化の出現頻度をもとに確率分布を推定している。

本モデルでは、打鍵位置変化量 Δg_n の符号に応じて、異なるカテゴリ集合に基づくカテゴリカル分布として定義した。具体的には $\Delta g_n \geq 0$ の場合、 $\Delta p_n \in \{0, 1, 2, \dots, 12\}$ に対するカテゴリカル分布、 $\Delta g_n < 0$ の場合、 $\Delta p_n \in \{0, -1, \dots, -12\}$ に対するカテゴリカル分布としてモデル化した。

3.3.4 音高言語モデル $P(p_n | p_{n-N}^{n-1})$

音高言語モデルは、音楽的文脈を考慮して、音高系列の整合性を評価するモデルである。本稿では、学習データから出現頻度をもとに N -gram モデルを構築する。このモデルにより、局所的な動作情報だけで推定した場合に生じうる不自然な音高列の発生を抑制し、より音楽的に自然な音高系列の生成を促す。

3.3.5 翻訳モデルと言語モデルの統合とデコード

音高列のデコードには、ビームサーチに基づく逐次探索アルゴリズムを用いた。推定は、SMT に基づく確率モデルを用いて行われる。具体的には、翻訳モデルと音高言語モデルの 2 つを組み合わせさせたスコア $\text{Score}(p_n)$ に基づいて、もっとも尤もらしい音高列をデコードする。 $\text{Score}(p_n)$ は以下のように定義する。

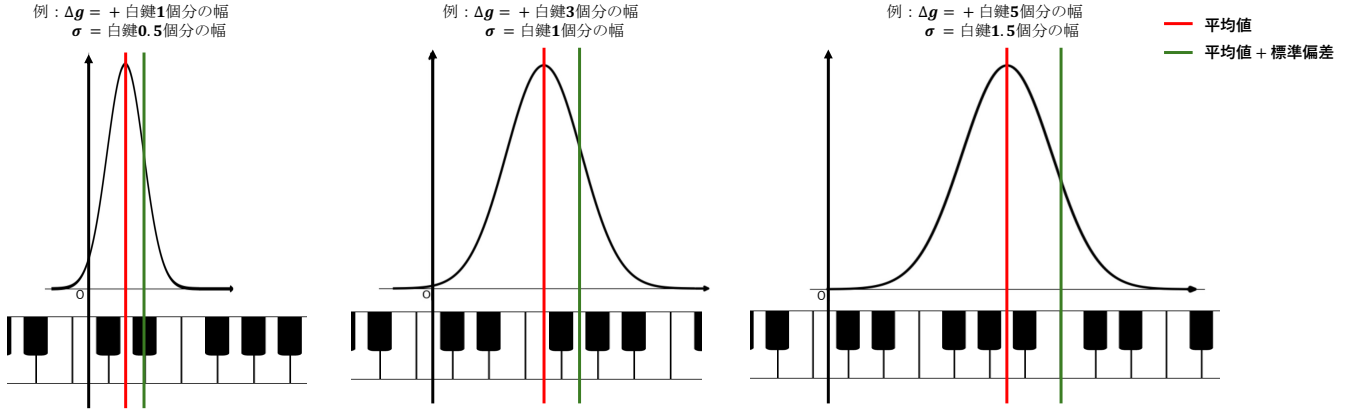


図 2 動作モデルにおける正規分布の例。縦の矢印は始点を表す。

表 1 記号の説明

記号	説明
n	演奏動作のステップを表す離散的なインデックス ($n \in \mathbb{Z}_{\geq 0}$)。実際の時間に対応するものではなく、打鍵のたびに 1 つずつ増加する。
p_n	時刻 n に対応する音高 (MIDI 番号)
Δp_n	音高の変化 $p_n - p_{n-1}$ (相対ピッチ)
g_n	時刻 n における打鍵位置 (座標)
Δg_n	打鍵位置変化 $g_n - g_{n-1}$ (移動ベクトル)
f_n	時刻 n に使用された指番号 (親指=1~小指=5)
F_n	指番号のペア (f_n, f_{n-1})
α, β	言語モデルと翻訳モデルの重み係数 (ハイパーパラメータ)
N	N -gram モデルにおける文脈長

$$\text{Score}(p_n) = \alpha \underbrace{P(p_n | \Delta g_n, F_n, p_{n-1})}_{\text{翻訳モデル}} + \beta \underbrace{P(p_n | p_{n-N}^{n-1})}_{\text{音高言語モデル}} \quad (6)$$

各要素は表 1 の通りに定義される。本モデルでは、このようにして入力された演奏時のジェスチャーの入力系列から確率的に音高系列を推定することが可能となる。

3.3.6 InvisiblePiano の提案

本研究では、物理的な鍵盤や専用ハードウェアを用いずに、右手でのピアノの演奏ジェスチャーのみで演奏可能な仮想ピアノシステム InvisiblePiano を提案する。InvisiblePiano では、MediaPipe Hands [18] によるハンドトラッキング技術を活用して、カメラ内蔵のスマートフォンを用いて手の動きをリアルタイムに計測する。この構成により、外部センサや専用ハードウェアを用いたデジタルピアノ [19] とは異なり、簡易的であり実用可能性が高い。

図 3 は InvisiblePiano の全体構成を示す。トラッキングにより得られた手の骨格情報から、各指の打鍵動作を検出し、時刻 n における打鍵開始時の x 座標 (左右方向) を $x_{n\text{start}}$ 、打鍵終了時の x 座標を $x_{n\text{stop}}$ 、および打鍵を行った指の指番号を f_n としてリアルタイムに取得する。打鍵

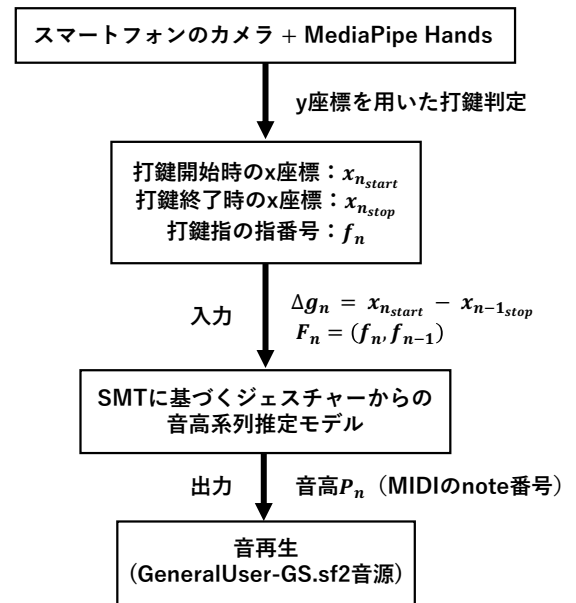


図 3 InvisiblePiano の全体構成

状態の判定には、演奏前にすべての指を机上に置いた静止状態における各指の y 座標 (上下方向) を取得し、5 本の指の中での最大値と最小値の差に基づいて閾値を定める。演奏中に取得された各指の y 座標とこの閾値を比較することで、打鍵の有無を判定する。音高系列推定には、提案手法を用いて打鍵位置変化量 $\Delta g_n = x_{n\text{start}} - x_{n-1\text{stop}}$ と運指情報 $F_n = (f_n, f_{n-1})$ を入力として処理を行う。推定された音高 p_n は fluidsynth を用いて GeneralUser-GS.sf2 音源からリアルタイムに音を再生することで、演奏が可能となっている。

4. 実験 1：提案手法の性能評価

本実験が掲げる主なリサーチクエスチョン (RQ) は以下の通りである。

- **RQ1:** 提案手法は、音高系列をどの程度正確に推定できるか？

実験 1 では、**RQ1** に対する回答を得ることを目的として、提案手法の精度を、シミュレーションにより定量的に評価

した。つまり、手法の性能評価のみを行い、ユーザによる操作やリアルタイム性は考慮しない。評価には、音高言語モデルの学習データに含まれる曲と、学習データに部分的に含まれる曲、学習データに含まれない曲を用いた。これらの楽曲に対し、本モデルでの音高系列の推定を行った。この実験では、モデルが出力した音高系列と正解との一致率を主要な評価指標とし、提案手法の音高系列推定性能を検証する。既存手法と比較をしないのは、[9,10]とは入力形式とタスクが異なり、[11]とは入力形式が異なることから、直接的な比較が難しいからである。

評価には、後述する13曲分のテストデータを用い、モデルが出力した音高系列と正解との一致率を主要な指標として性能を検証した。

4.1 学習データの構成

本研究で使用した音高言語モデルの学習データは、『童謡・唱歌の世界』^{*1}に掲載されている楽曲の中から、ハ長調に移調可能であった392曲（1曲あたり平均約200音）の旋律部分で構成されている。これらの楽曲はいずれも単旋律であり、重音（和音）は含まれていない。

本モデルは白鍵の演奏のみを対象としているため、モデル学習にあたっては、対象楽曲をすべてハ長調に移調したうえで、臨時記号を含む音は排除し、白鍵内の音高のみを対象とした。また、音高 p に関しては、オクターブ違いの音高は同一のものとして扱った。

音高言語モデルとしては、高次の N -gram モデルにおいてスパース性（未出現系列の存在）が問題となる場合に備えて、2-gram から 10-gram までの複数の N -gram モデルを構築した。この設計により、観測されていない高次系列に対しても、より信頼性の高い低次系列の統計に基づく確率推定が可能となる。実際に適用する際には、一般的なバックオフスムージングのようにすべての N -gram モデルを加重的に統合するのではなく、該当する N -gram が存在する範囲で可能な限り高次のモデルを単独で用いる方式を採用した。これにより、推定精度を維持しつつ計算効率を向上させている。

一方、翻訳モデルにおける運指情報に基づく音高の条件付き確率分布 $P(\Delta p_n | F_n)$ は、PIG (PIano fingerinG) データセット [20] を用いて学習した。このデータセットには、多様なピアノ楽曲に対して付与された実演者の運指データが含まれており、本研究では、ある特定の運指 f に対して選択される音高 p_n の出現頻度に基づいて確率分布を構築した。

特定の運指ペア F_n に対して、指の間隔 $d = f_n - f_{n-1}$ と、打鍵動作に伴う鍵盤上で移動した白鍵の数 ($\Delta g_n / W$) が $|d - (\Delta g_n / W)| = 0$ の場合には、対応する音高 p_n の確



図4 『朧月夜』（作曲：瀧廉太郎）における1フレーズの運指例。a, b, c, d は運指パターン。

率に 1.15 倍の重みを、また $|d - (\Delta g_n / W)| = 1$ の場合には 1.10 倍の重みを掛けた。これらの補正を反映したのち、全体の確率分布が 1 となるよう正規化を行った。この補正は、学習データがクラシック音楽に特化しており運指の傾向がジャンルに偏っている可能性があるため、より一般的な自然運指の傾向を補う目的で導入したものである。なお、この補正方針は「自然な運指では、指の間隔と実際に打鍵する白鍵の間隔が一致する傾向がある」という考えに基づいている。

本実験では、 $W = 20\text{mm}$ と仮定した。ハイパーパラメータ α は $2/3$ 、ハイパーパラメータ β は $1/3$ とした。

4.2 テストデータの構成

本実験では、音高言語モデルの学習データに対するフレーズの含有状況に基づき、13 曲を以下の3カテゴリに分類し、テストデータとして使用した。

- **完全含有曲**：音高言語モデルの学習データに含まれている曲（5 曲、総音数 448 音、ジャンル：童謡・唱歌）
- **部分含有曲**：音高言語モデルの学習データに一部のフレーズや音高系列が含まれている曲（3 曲、総音数 326 音、ジャンル：童謡・唱歌）
- **非含有曲**：音高言語モデルの学習データに一切含まれていない曲（5 曲、総音数 455 音、ジャンル：クラシックピアノ曲）

各曲に対し、楽譜から算出された打鍵位置変化量 (Δg_n) に乱数ノイズを加えたものと、ピアノ経験 10 年以上の奏者 4 名がそれぞれ独立に付与した運指列 (F_n) を組み合わせ、合計 50 データセット（13 曲 $\times$ 4 運指パターン、うち 2 曲において運指パターンが重複）を構成した。

乱数ノイズは以下の規則に基づき、白鍵幅に換算して一様分布から連続値として付与した。

- $\Delta g_n \leq W$ ： $-0.5W \sim +0.5W$ 個分の範囲
- $\Delta g_n \leq 4W$ ： $-W$ 個分 $\sim +W$ 個分の範囲
- 上記以外の場合： $-1.5W$ 個分 $\sim +1.5W$ 個分の範囲

図4には、テストデータの運指情報の一例として、完全含有曲に分類される『朧月夜』（作曲：瀧廉太郎）の1フレーズに対して付与された運指を示す。

4.2.1 ビームサーチにおける安定化処理

本手法はリアルタイム推定を想定しており、逐次的に得

*1 <https://www.s-pst.info/douyou-syouka/>

表 2 補正の有無によるビームサーチ path の比較

step	0%のときの path	15%のときの path
4	[67, 64, 67, 62]	[67, 64, 67, 62]
5	[67, 67, 72, 65, 70]	[67, 67, 72, 65, 70]
6	[67, 64, 67, 64, 69, 69]	[67, 64, 67, 62, 67, 67]
7	[67, 64, 67, 64, 69, 67, 70]	[67, 64, 67, 62, 67, 67, 67]
8	[67, 67, 67, 67, 72, 71, 71, 71]	[67, 64, 67, 62, 67, 67, 67, 67]
9	[67, 67, 67, 67, 72, 71, 71, 69, 69]	[67, 64, 67, 62, 67, 67, 67, 67, 67]
10	[67, 67, 67, 67, 72, 71, 71, 69, 69, 69]	[67, 64, 65, 62, 67, 67, 72, 67, 64, 64]

られるスコアの変動により、推定される音高系列が不安定になる場合があった。この問題に対処するため、ビームサーチにおいてその時点で最もスコアの高い経路 (path) に対して、対応するスコアを 1.15 倍する補正を加えた。

この補正の効果を確認するために、学習データやテストデータには含まれていない Telemann 作曲のピアノ曲冒頭 10 音を対象とした比較実験を行った。この区間に対して、ピアノ経験 10 年以上の演奏者が付与した運指は [5, 3, 5, 2, 5, 1, 5, 4, 3, 2] である。また、対応する正解の音高系列は [67, 64, 67, 62, 67, 60, 72, 71, 69, 67]、打鍵に伴う指先の移動量 (打鍵変化量) は、前述の規則に従ってノイズを付与し [0, -30.86, 28.46, -53.79, 55.98, -96.37, 128.73, -10.91, -26.52, -10.94] (単位: mm) であった。

重み補正を行わない場合、各ステップで最大スコアを保持していた path は頻繁に切り替わり、推定系列が不安定であった。一方、補正として 15% の重みを与えた場合、path の切り替わりが抑えられ、より一貫した推定が得られた。

表 2 は、補正の有無による path の推移を比較したものである。step4 までは両者同じであるため表は step4 から記載している。補正あり (右側) は、途中で path が極端に切り替わることが少なく、推定の一貫性が保たれていることがわかる。

また、最終的な推定音高系列は、補正なしでは [67, 64, 67, 62, 70, 69, 70, 71, 69, 69]、補正ありでは [67, 64, 67, 62, 70, 67, 67, 67, 64] となった。いずれも打鍵変化量は共通であるが、補正ありの方が推定音高の上下方向が実際の手の移動方向とより整合しており、推定の安定性が向上していると考えられる。

このような傾向から、本研究では 15% の補正重みを用いることとした。ただし、本手法における補正量の最適値については、現時点では統計的な検証を行っておらず、今後の課題とする。

4.3 評価方法

提案手法の音高推定性能は、モデルが出力した音高系列と正解音高系列との音高およびその時刻 (出現順序・位置) が一致した割合 (以下、一致率) を用いて評価した。一致

表 3 含有レベル別の平均一致率

含有レベル	平均一致率 (%)
完全含有曲	98.43
部分含有曲	32.63
非含有曲	18.07

表 4 完全含有曲における音高推定精度 (一致率, %) の比較
提案手法 vs 運指モデルなし。(下線は一致率の高い方を表している)

	運指	曲名 (ID)	提案手法	運指モデルなし
a		1	<u>100.00</u>	18.75
		2	<u>100.00</u>	21.43
		3	<u>92.70</u>	23.36
		4	<u>100.00</u>	<u>100.00</u>
		5	<u>100.00</u>	<u>100.00</u>
b		1	<u>100.00</u>	3.12
		2	<u>100.00</u>	8.57
		3	<u>100.00</u>	11.68
		4	<u>93.02</u>	27.13
		5	<u>100.00</u>	<u>100.00</u>
c		1	<u>100.00</u>	<u>100.00</u>
		2	<u>91.43</u>	42.86
		3	<u>100.00</u>	<u>100.00</u>
		4	96.12	<u>100.00</u>
		5	<u>100.00</u>	26.09
d		1	<u>100.00</u>	6.25
		2	<u>100.00</u>	22.86
		3	<u>100.00</u>	<u>100.00</u>
		4	95.35	<u>100.00</u>
		5	<u>100.00</u>	26.09

率は、以下の式で定義される。

$$\text{一致率} = \frac{\text{一致した音高の数}}{\text{モデルが推定した音高の総数}} \quad (7)$$

ここで、モデルが推定した音高系列の数と正解音高の数は、常に一致する。

4.4 結果と考察

表 3 に、テストデータ全体における音高系列推定の一致率を含有レベル別に示す。完全含有曲においては、一致率が 98.43% と非常に高く、提案手法の有用性が示す結果となった。一方で、学習データに含まれないフレーズを多く含む非含有曲では一致率が 18.07% にとどまり、提案手法の汎化性能に課題が残ることが示された。

以下、それぞれの含有レベルにおける推定結果について詳細に検討する。

4.4.1 完全含有曲の分析

完全含有曲では、音高言語モデルがすべての音高系列をカバーしており、また翻訳モデルとの統合により、わずかな打鍵位置のズレに対しても頑健に音高系列を復元できた。

表 4 に、完全含有曲に対する音高系列推定精度の比較結果を示す。左側は翻訳モデルに運指情報を含めた提案手法

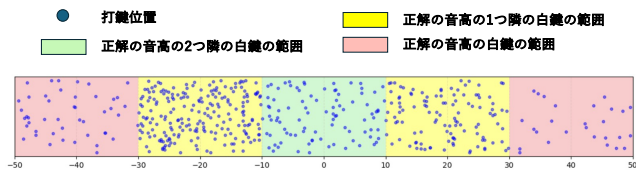


図 5 打鍵位置の変化量 (Δg) の分布

の結果、右側は運指モデルを用いずに運指を考慮できない結果である。下線は、各曲において提案手法と比較手法のうち一致率が高かった方を示す。

結果から明らかなように、提案手法では全曲・全演奏者に対して安定して 100% 近い一致率が得られている一方、運指モデルなしでは演奏者や曲によって一致率が大きく低下している。たとえば曲 1 では、運指 a, b, d において著しく一致率が下がっており、動作モデル単独では運指の違いや手の形状に伴う多様な打鍵位置変化を適切に解釈できないことがわかる。

この結果は、打鍵位置だけで音高系列を推定するには限界があり、運指という中間的な意味表現を仲介することが、より安定かつ精度の高い音高系列復元に不可欠であることを示している。すなわち、翻訳モデルによる打鍵位置から運指への変換が、音高言語モデルと連携する上で重要な役割を果たしていると言える。

ただし、曲 4 では運指 c, d においては提案手法の方が低い一致率であった。これは、運指 c, d において 3 小節目にある D4 から E4 への遷移を薬指から中指で移動していることが要因と考えられる。D4 から E4 への遷移の間に休符があるため、運指自体は不自然ではないが本手法では休符を考慮できていないために運指モデルを利用すると一致率が下がるという結果になった。従って、休符の処理が今後の課題である。

加えて、図 5 に、表 4 で比較手法の一致率が 30% 未満であり、かつ提案手法の一致率が 100% であった時の打鍵位置の変化量 (Δg) を散布図として可視化した。横軸は打鍵位置の変化量 (Δg) mm である。縦軸に意味はなく、データ点を視認しやすくするためにランダムに配置している。この可視化により打鍵位置にノイズが含まれる状態においても、提案手法がどの程度正確に音高を復元できたかを直感的に示される。図 5 より、隣の鍵盤範囲を打鍵している場合でも、翻訳モデルが演奏ジェスチャーを適切に推定し、その上で音高言語モデルが文脈を補完することにより、大きくずれた入力からでも正しい音高系列が高精度に復元されていることが視覚的に確認できる。

4.4.2 部分含有曲の分析

部分含有曲について、演奏者（運指設計者）ごとの音高系列推定精度を表 5 に示す。

表から分かるように、運指によって推定精度にばらつきが見られ、特に曲 1 における d の一致率は 22.06% と他の

表 5 部分含有曲における演奏者別一致率 (%)

曲名 (ID)	a	b	c	d
1	42.65	42.65	42.65	22.06
2	53.42	52.02	52.02	43.84
3	12.50	16.96	5.36	5.36

表 6 部分含有曲における相対音高（音高遷移）の一致率比較 (%)

pp は percentage point を意味し割合の差であり改善量を示す。

曲名 (ID)	運指 (ID)	10-gram (%)	2-gram (%)	pp
1	a	44.78	53.73	+8.95
	b	44.78	64.18	+19.40
	c	44.78	61.19	+16.41
	d	16.42	55.22	+38.80
2	a	62.76	76.55	+13.79
	b	67.59	71.03	+3.44
	c	71.03	80.69	+9.66
	d	62.07	66.21	+4.14
3	a	19.82	65.77	+45.95
	b	21.62	53.15	+31.53
	c	28.83	59.46	+30.63
	d	28.83	56.76	+27.93

運指と比較すると 20% 以上低い。この要因として、始めの A4 から C5 への遷移を薬指から小指で移動しており、薬指と小指で白鍵 2 個分を跳躍するといった特殊な運指を使っていたことが考えられる。したがって、多様な運指にも対応できる翻訳モデルを拡張させることが課題としてあげられる。また、音高言語モデルが高次 (10-gram) で構築されていることから、未知の系列に遭遇した際のスパース性が影響している可能性があるため、検証する。

4.4.3 部分含有曲における n-gram 階層の影響と相対音高精度の評価

音高言語モデルのスパース性の影響を検証するため、音高言語モデルを 10-gram から 2-gram に変更し、再評価を行った。この際、前の音高の推定ミスがその後の推定結果に連鎖的な影響を与える可能性があるため、評価においては音高そのものの一致ではなく、相対音高（音高差）、すなわち $p_n - p_{n-1}$ の一致率に基づいて行った。これにより、局所的な推定ミスによる全体的な音高のずれを無視しつつ、提案手法の性能を評価できる。

表 6 に、2-gram 使用時の相対音高一致率を示す。結果として、多くの演奏者・楽曲において、音高の絶対的一致率は低下したものの、音楽的に意味のある遷移（相対音高）は高精度に復元されていることがわかった。これは、翻訳モデルが打鍵位置と運指の関係から適切な方向を導き、N-gram の次数を下げることで学習データがスパースでない音高言語モデルの確率を利用できるからだと考えられる。

以上より、本実験では提案手法が、完全含有曲に対しては高精度な推定が可能である一方、部分含有曲に対する汎

表 7 非含有曲における音高一致率 (%)

曲 (ID)	A	B	C	D
1	22.50	32.50	30.00	-
2	11.11	11.11	17.59	-
3	7.50	20.00	6.25	13.75
4	12.50	12.50	23.86	12.50
5	25.42	25.42	16.95	23.73

化性能に課題があることが明らかになった。今後の改良点として以下の点が挙げられる。

- 音高言語モデルの N -gram 階層を柔軟に切り替えることで、スパース性と汎化性能のトレードオフを調整すること
- 音高言語モデルにおいて、相対音高系列を含む拡張的なスコアリングを導入すること
- 運指情報の誤差に対する頑健性を高めるための拡張翻訳モデルを導入すること

4.4.4 非含有曲の分析

非含有曲 (学習データに全く含まれていない未知の旋律) に対する推定精度を分析したところ、全体として一致率は低く、最大でも 32.50%、18 曲中 11 曲は 20% 以下であった。各曲における演奏者ごとの一致率を表 7 に示す。

曲 1 は b と d が、曲 2 は c と d が同じ運指であった。よって、d では実験を行わなかった。一致率全ての平均値は 18.07% であり、提案手法の現状のモデルが未知系列に対しては十分な汎化性能を発揮できていないことが示唆される。

また、一致率が平均未達であった演奏データを分析した結果、推定精度の低下につながった可能性のある運指系列には以下があった。

- ポジション移動のために同一音を不自然に異なる指で連続して打鍵している例 (例: C4 を親指の後、小指で打鍵)
- 休符やフレーズ切り替えを用いたポジション移動によって、運指と音高の対応が文脈から逸脱する例 (例: C4 を小指で弾いた後、E4 を親指で打鍵)

これらの運指設計は学習データに多く存在しないため、系列全体の推定精度が下がると考えられる。

また、推定対象の系列長が短い場合にモデルがより安定して動作するようであった。非含有曲の推定結果において冒頭 12 音に限定した一致率を算出し、その結果を表 8 に示す。表より全体として冒頭部分では推定精度が表 6 と比較して高く、短系列に対して本モデルがより安定して動作する可能性が示唆される。すなわち、初期の短いコンテキストであれば翻訳モデルおよび言語モデルの一致率が高く、正答に導かれやすいことが考えられる。

一方で、長い系列では音高言語モデルの予測が支配的となり、特に訓練データとは異なるジャンルの曲に対しては、

表 8 非含有曲における冒頭 12 音の音高一致率 (%)

曲 (ID)	a	b	c	d
1	58.33	66.67	50.00	-
2	83.33	83.33	83.33	-
3	58.33	41.67	41.67	33.33
4	58.33	58.33	66.67	58.33
5	83.33	83.33	83.33	75.00

文脈予測の汎化性能に課題があるとわかった。

以上の結果から、今後は翻訳モデルにおける運指情報の表現形式の最適化に加え、音高言語モデルのジャンル横断的な学習や拡張も重要な改善方策となると考えられる。

5. 実験 2: InvisiblePiano のユーザスタディ

本実験が掲げる主なりサーチクエスション (RQ) は以下の通りである。

- **RQ2:** 鍵盤を目視できない状況において、一般的なピアノ演奏と同様に、鍵盤の打鍵位置のみに基づいて音高を制御するシステムと比較して、提案する InvisiblePiano は、演奏精度およびユーザ体験にいかなる影響を与えるか?

実験 2 では、**RQ2** に対する結果を導くために、実際の演奏環境において InvisiblePiano の有効性を検証するためのユーザ実験を行った。この実験では、鍵盤を視認できない状況下で、2 種類のデジタルピアノ (打鍵位置のみに基づく音高系列推定システムと InvisiblePiano) を用いて同一のメロディ演奏を行ってもらい、その演奏の正解音高系列との一致率と主観的な体験評価 (アンケート) に基づいて比較・分析を行った。実験 2 はお茶の水女子大学の倫理審査委員会の承認のもとで実施された。

5.1 目的と比較条件

本実験の目的は、2 つのシステムによる演奏パフォーマンスおよび主観的な体験を比較することで、InvisiblePiano の有効性を検証することである。

- **ベースライン条件 (Baseline):** 打鍵位置のみに基づいて音高を制御するモデルを用いた演奏
- **提案手法条件 (InvisiblePiano):** 提案手法を用いた演奏

いずれの条件も、演奏者は鍵盤の位置を視認できないように視線を固定した状態で演奏を行う。これにより、視覚に依存しない状況における演奏支援の効果を純粋に比較可能とした。

5.2 Baseline の設計

Baseline は 3.3.6 節で述べた InvisiblePiano と比較すると、音高系列推定の手法だけ異なる。Baseline の音高系列推定では、打鍵時の x 座標に基づいて音高系列を決定する。

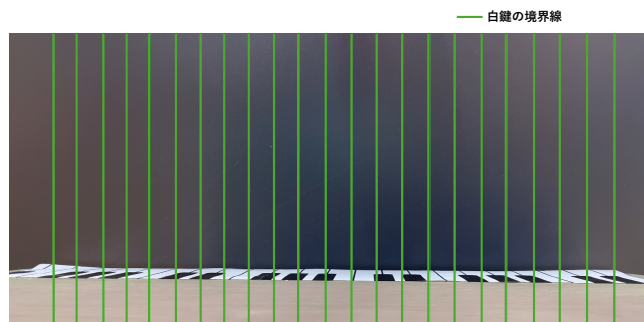


図 6 Baseline における鍵盤範囲と打鍵位置の対応

MediaPipe Hands を用いて取得したランドマーク座標は 1920×1080 ピクセルの画像上で取得し、白鍵 1 個分の幅は 66.92 ピクセルに設定した。これは後述の InvisiblePiano でも同じである。鍵盤幅に基づき各キーの位置を定義した。図 6 に示すように、打鍵位置がどの鍵盤範囲に含まれるかを判定し、対応する音高を割り当てる。図中の緑色の x 軸方向の線は、設定した鍵盤の境界を示している。なお、図中には目安として紙に印刷された鍵盤の図を重ねているが、この紙に示された鍵盤境界と設定した境界には、わずかな位置のズレが見られる。しかし、各鍵盤の中心点はすべて対応する設定範囲内に含まれており、このズレは認識上の誤差として無視できると判断した。ハイパーパラメータ α は $2/3$ 、ハイパーパラメータ β は $1/3$ とした。InvisiblePiano のハイパーパラメータも同様に設定した。

5.3 InvisiblePiano の音高系列推定モデル

InvisiblePiano における提案手法の学習データは 4.1 節で述べた実験 1 の学習データから後述の課題曲を除いたものとした。Baseline と同様に、MediaPipe Hands を用いて取得したランドマーク座標は 1920×1080 ピクセルの画像上で取得し、白鍵 1 個分の幅は 66.92 ピクセルに設定した。また、InvisiblePiano でも 4.2.1 節で述べた安定化処理を行った。

5.4 実験環境

実験環境は図 7、図 8 の通りとした。詳細は表 9 の通りとした。実験参加者は机の前に座り、机上の紙製鍵盤の上に手を置いて机を叩くように演奏をした。紙製鍵盤の白鍵 1 個分の幅は 2.0mm とした。記録用のスマートフォンは倍率 0.5x で動画を撮影した。実験には Apple 社製の iPhone 13 を用いた。

5.5 参加者

本実験には、20 代女性 18 名が参加した。ピアノ演奏歴に関しては、10 年以上の経験者が 4 名、1 年以上 10 年未満の経験者が 7 名、1 年未満の初心者が 7 名であった。

参加者が全員 20 代女性ではあるが、実験参加者 18 名の

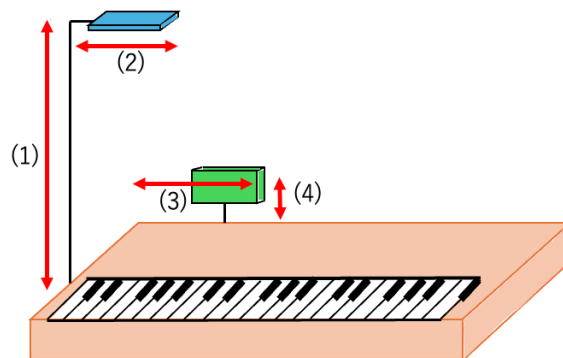


図 7 正面から見た実験環境

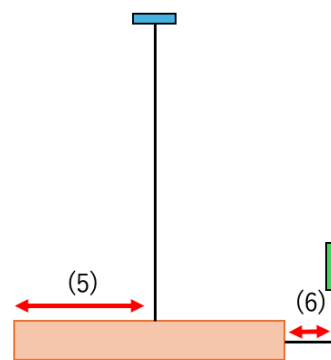


図 8 横から見た実験環境

表 9 図 1, 2 の詳細

図 1, 図 2 で示す対象	詳細
オレンジ色の直方体	机
青色の直方体	記録用のスマートフォン
緑色の直方体	ハンドトラッキング用のスマートフォン
(1)	51cm
(2)	20cm
(3)	24cm
(4)	4cm
(5)	20cm
(6)	6cm

手幅（親指から小指までの最大開張長）を測定し、産業技術総合研究所の公表する身体寸法データ [21] と比較した結果、男性の手幅も一部カバーできていた。手幅の分布を図 9 に示す。

5.6 課題と実施手順

演奏課題として、ハ長調の『ハッピーバースデイトゥーユー』の一節（全 11 音）を課題曲と設定した。図 10 は課題曲の楽譜である。課題曲は音高言語モデルの学習データに含まれないが、同一ジャンル（童謡・唱歌）に属する曲である。番号は運指番号を表しているが、実験参加者は必要

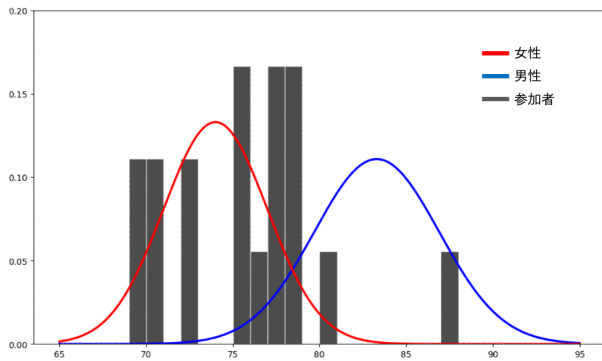


図 9 実験参加者の手幅分布と成人男性・女性の手幅の正規分布との比較 (縦軸：確率密度，横軸：手幅 (mm))



図 10 ハ長調の『ハッピーバースデートゥーユー』の一節

に応じて運指を変更した。また、弾き直しは禁止とした。本実験では、演奏される音の数は常に 11 音に限定し、すべての実験参加者が必ず 11 音を弾くように指示した。これにより、演奏結果の分析においては、余剰や欠損といった変動要因を排除し、11 音すべてについて正解との一致／不一致のみを判定対象とすることが可能となった。

参加者はまず目視によって楽譜を覚えた後、鍵盤が書かれた紙の上で聴覚フィードバックがない状態で練習をもらった。参加者が「覚えた」と判断した時点で練習を終了し、鍵盤が書かれた紙から視線を外し、手元を見ない状態で、それぞれのシステム上で課題曲を演奏してもらった。実験参加者は全員 Baseline, InvisiblePiano の順に演奏した。演奏時の手の様子は、上方からカメラで撮影し記録した。打鍵開始時および終了時の x 座標、打鍵した指の指番号、音高系列も記録した。

演奏後には、各条件ごとに主観評価アンケートに回答してもらい、以下のユーザ体験評価を収集した。

- どちらのシステム (Baseline / InvisiblePiano) の方が弾きやすかったと感じたか
- それぞれのシステムで演奏してみた感想 (操作性、違和感、気づいた点など)

5.7 評価指標

本実験における演奏精度は、実験 1 と同様に一致率で評価した。また、ユーザ体験に関しては、演奏後に実施した主観評価アンケートの回答内容をもとに分析を行った。具体的には、「どちらのシステムが弾きやすかったか」という選択肢の集計および、自由記述による各システムの操作性や違和感、気づきに関する感想を質的に評価し、

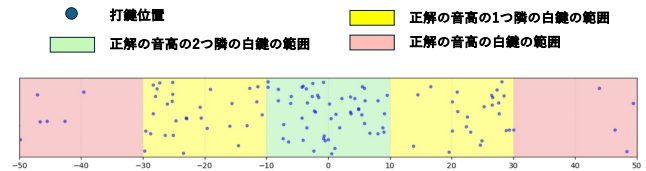


図 11 InvisiblePiano の一致率 100% の
実験参加者の打鍵位置の変化量 (Δg) の分布

InvisiblePiano の有効性に関する知見を得ることを目的とした。

5.8 結果と考察

表 10は、提案手法 InvisiblePiano と Baseline における音高系列の一致率 (精度) および主観的な使いやすさ評価の結果である。括弧内の数値は、InvisiblePiano における打鍵位置のみに基づいた場合の一致率を示しており、補正を行わなかった場合の参考値として提示している。

InvisiblePiano の補正後の一致率は多くの実験参加者で 100 % を達成しており、打鍵位置に誤差があった場合でも意図した演奏が実現可能であることが示された。特に、実験参加者 e, g, i, k, m, n, o, p, q, r, s, v では一致率が 100 % であり、補正機構が打鍵位置のずれを効果的に吸収していることが確認できる。これに対して、打鍵位置のみに基づく一致率は平均して低く、InvisiblePiano による補正が重要な役割を果たしていることが明らかである。図 11では、InvisiblePiano で出力された音高系列が一致率 100 % であった実験参加者の打鍵位置の変化量 (Δg) を散布図として可視化した。横軸は打鍵位置の変化量 (Δg) (mm) を示しており、縦軸には意味はなく、重複するデータ点を視認しやすくするためにランダムに配置している。この図は、打鍵位置にノイズが含まれる状態においても、提案手法がどの程度正確に音高を復元できたかを直感的に示している。図中の点のばらつきが大きいかにかかわらず、高一致率が維持されていることから、InvisiblePiano は打鍵の不確かさを吸収し、演奏意図を正確に音として出力できる柔軟性を有していることがわかる。なお、実験 1 と比較して本実験で高い推定精度が得られたのは、課題曲が音高言語モデルの学習データには含まれていないものの、同一ジャンル (童謡・唱歌) であり、音数も少なかったため、推定精度が高かったと考えられる。

さらに、主観評価では 19 名中 16 名の実験参加者が「InvisiblePiano の方が弾きやすい」と回答しており、提案手法の使用感の良さも確認された。ピアノ初心者であった実験参加者 m は、「通常は鍵盤に音名を書いて確認しながらでないと不安であったが、InvisiblePiano では打鍵する指と音の高低の移動方向さえ覚えれば良かったため、自信を持って演奏できた」とコメントしており、初心者にとっても高いアクセシビリティを有することが示唆される。

表 10 Baseline と InvisiblePiano における一致率の比較と主観評価結果

	e	f	g	h	i	j
Baseline(%)	64	64	82	91	82	64
InvisiblePiano(%)	100 (36)	64 (55)	100 (73)	91 (45)	100 (55)	30 (45)
弾きやすさ	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	Baseline
	k	l	m	n	o	p
Baseline(%)	64	64	9	55	55	64
InvisiblePiano(%)	100 (64)	45 (55)	100 (64)	100 (64)	100 (9)	100 (18)
弾きやすさ	<u>InvisiblePiano</u>	Baseline	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>
	q	r	s	t	u	v
Baseline(%)	64	45	64	36	64	82
InvisiblePiano(%)	100 (36)	100 (36)	100 (64)	73 (55)	91 (64)	100 (55)
弾きやすさ	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	<u>InvisiblePiano</u>	どちらも同じ	<u>InvisiblePiano</u>

※ InvisiblePiano 列の括弧内数値は、打鍵位置から得られた音高系列との一致率であり、補正処理を行わなかった場合の参考値である。

一方で、課題も明らかとなった。主な指摘として、(1) InvisiblePiano の応答速度が Baseline に比べてやや遅く感じられた点がある。これは、補正処理による計算遅延に起因すると考えられるが、一部の実験参加者からは「実際のピアノ鍵盤のようなタッチの重さを感じた」というポジティブな捉え方も見られた。また、(2) ジェスチャー認識精度に関しては、実験参加者 f, j, l, t において、打鍵していない指が誤って打鍵として認識されるケースが確認されており、この誤認識が InvisiblePiano の一致率の低下に影響を与えたと考えられる。本システムでは MediaPipe Hands を用いて打鍵指を判定しているが、判定精度向上が、今後の技術的課題として挙げられる。

さらに、(3) InvisiblePiano の補正により意図した音が出ることで、実験参加者は打鍵位置のズレに気づきにくくなる傾向が確認された。実際、Baseline では聴覚フィードバックをもとに演奏中に修正が行われていたが、InvisiblePiano では修正の必要性を感じにくいため、結果として実際の打鍵位置が大きくずれる傾向があった。これは、特にピアノ練習用途においては誤った学習に繋がるリスクがある点として注意が必要である。

ただし、ピアノ演奏の目的は必ずしも技術向上に限られるものではなく、音楽療法や情緒的な自己表現といった非専門的・非学習的な文脈でも広く実践されている。実際に、楽器演奏を含む能動的な音楽活動が自己効力感や主観的幸福感の向上に寄与することが報告されており [22]、また、ピアノ演奏が不安やストレスの軽減、感情表出の手段としても効果的であることが指摘されている [23]。本実験によって、InvisiblePiano は初心者にも高いアクセシビリティを有する可能性が示唆されたため、今後、Well-being のための音楽体験支援ツールとしての応用が期待される。

また、ピアノ演奏の目的は、単に音高制御能力を高めることにとどまらない。音高の正確性に加えて、身体を通じた音楽表現の重要性が指摘されている。たとえば、演奏者の身体動作に着目した研究では、音を正確に出すことだけ

でなく、音楽的表現につながる身体の使い方を重視する指導法への転換が提案されている [24]。さらに、James [25] はピアノ演奏における上半身の動きの重要性を強調し、それが演奏効率や表現の質に密接に関わることを示した上で、より身体的感覚に基づく練習方法の必要性を指摘している。これらの知見を踏まえると、ピアノ演奏において重要なのは単なる音高制御能力の向上にとどまらず、演奏者自身の内在的な音楽性を身体を通じて表現する力であると考えられる。

InvisiblePiano は、打鍵位置が不正確であっても意図した音を出すことが可能であり、演奏者がミスタッチを気にせず音楽表現に集中できる点において、表現力育成の支援となる可能性がある。

現時点での InvisiblePiano は、音の強弱や音色の制御を考慮していないため、今後の課題として演奏ジェスチャーに基づく表現制御機能の拡張が挙げられる。演奏評価における視覚情報の役割を調査した研究では、音を除去した映像のみの提示条件において、聴覚情報を伴う条件よりも演奏の質をより的確に評価できたことが示されている [26]。また、視覚情報が演奏表現のニュアンス、特にタッチや強弱に関する情報伝達に寄与する可能性が報告されている [27]。これらの先行研究に基づけば、演奏者のジェスチャーから音楽表現の意図を推定し、それに基づいて音の表情を動的に制御する拡張インタフェースの実現は十分可能であると考えられる。

6. まとめ

本研究では、実体のある鍵盤を必要としないピアノ演奏を可能にする仮想ピアノの実現を目的として、指の動きと運指情報に基づいて音高系列を確率的に推定する手法を提案した。本節では、提案手法の精度検証を通じて、RQ1「提案手法は、音高系列をどの程度正確に推定できるか？」と RQ2「鍵盤を目視できない状況において、一般的なピアノ演奏と同様に、鍵盤の打鍵位置のみに基づいて音高を制御

するシステムと比較して、提案する InvisiblePiano は、演奏精度およびユーザ体験にいかなる影響を与えるか？」に対する結論を述べる。

実験 1 では、提案モデルが学習データに含まれる楽曲に対して高い音高系列推定精度（正解音高系列との一致率）を示した。一方で、未知の演奏者や楽曲に対しては推定精度がやや低下する傾向が確認された。これは、運指情報と音高言語モデルの両方において、訓練データに依存した傾向が見られるためであると考えられる。また、音高言語の文脈長（n-gram）の設定によって精度が大きく変動することが示された。特に非含有曲に対しては、音高遷移の観点からの評価において n を小さく設定したモデルの方が汎化性能が高くなる傾向が確認され、学習バイアスと汎化性能のトレードオフが示唆された。

実験 2 では、新たに提案する InvisiblePiano システムを評価し、鍵盤の打鍵位置のみに基づいて音高を制御するシステムと比較した。その結果、InvisiblePiano は鍵盤を目視できない状況でも、演奏者が意図した音高を忠実に再現できることが確認された。さらに、ユーザ体験に関する評価では、InvisiblePiano の方が弾きやすいと評価された実験参加者が多く、初心者でも使いやすいシステムであるとの意見が得られた。

提案モデルの今後の課題としては、翻訳モデルにおける運指情報の表現方法の最適化や、音高言語モデルのジャンル横断的な学習・拡張が挙げられる。また、InvisiblePiano に関しては、応答速度の向上や、ジェスチャー認識における打鍵判定精度の改善が必要である。さらに、音の強弱や音色のコントロールを実現するための拡張も今後の課題となる。

謝辞 実験参加者の皆様に深く感謝いたします。

参考文献

- [1] Feng, Y.: Design and research of music teaching system based on virtual reality system in the context of education informatization, *PLoS ONE*, Vol. 18, No. 10, p. e0285331 (2023). Article e0285331, total 16 pages.
- [2] Fu, Y., Liu, J., Li, X. and Sun, X.: A haptic feedback glove for virtual piano interaction, *Virtual Reality & Intelligent Hardware*, Vol. 7, No. 1, pp. 95–110 (2025).
- [3] Liang, H., Wang, J., Sun, Q., Liu, Y.-J., Yuan, J., Luo, J. and He, Y.: Barehanded music: Real-time hand interaction for virtual piano, *the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games(I3D '16)*, pp. 87 – 94 (2016).
- [4] Banton, L. J.: The role of visual and auditory feedback during the sight-reading of music, *Psychology of Music*, Vol. 23, No. 1, pp. 3–16 (1995).
- [5] Loerch, L.: The Association of Eye Movements and Performance Accuracy in a Novel Sight-Reading Task, *Journal of Eye Movement Research*, Vol. 14, No. 4 (2021). Article 5, total 30 pages.
- [6] Kawase, S.: Gazing behavior and coordination during piano duo performance, *Attention, Perception, & Psychophysics*, Vol. 76, pp. 527–540 (2014).
- [7] Bishop, L., Cancino-Chacón, C. and Goebel, W.: Moving to Communicate, Moving to Interact: Patterns of Body Motion in Musical Duo Performance, *Music Perception: An Interdisciplinary Journal*, Vol. 37, No. 1, pp. 1–25 (2019).
- [8] Vandemoortele, S., Feyaerts, K., Reybrouck, M., Bievre, G. D., Brône, G. and Baets, T. D.: Gazing at the partner in musical trios: a mobile eye-tracking study, *Journal of Eye Movement Research Journal of Eye Movement Research*, Vol. 11, No. 2 (2018).
- [9] Chou, B. S.-H., Jajal, P., Eliopoulos, N. J., Nadolsky, T., Yang, C.-Y., Ravi, N., Davis, J. C., Yun, K. Y.-J. and Lu, Y.-H.: Detecting Music Performance Errors with Transformers, *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*, Vol. 39, No. 22, pp. 23687–23695 (2025).
- [10] Marinov, G.: Real-time Error Correction and Performance Aid for MIDI Instruments, arXiv:2011.13122 (2020). arXiv preprint arXiv:2011.13122.
- [11] Gușită, C., Stănescu, D., Gușită, B., Gaina, L. I. and Ghergulescu, I.: A Novelty Real-Time Gesture Recognition Model for Air-Hand Piano Playing Using MediaPipe, *2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP2024)*, IEEE, pp. 3–10 (2024).
- [12] Begum, H., Shaheen, S., Moetesum, M. and Siddiqi, I.: Digital Beethoven — An Android Based Virtual Piano, *Proceedings of the 2017 13th International Conference on Emerging Technologies (ICET2017)*, IEEE, pp. 1–5 (2017).
- [13] Wang, Y. and Song, L.: Virtual Piano System Based on Monocular Camera, *Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021, Shenzhen, China, August 9–12, 2021, Proceedings*, Lecture Notes in Computer Science, Vol. 12837, Springer, pp. 112–121 (2021).
- [14] Yang, Y., Wang, Z. and Li, Z.: MuGeVI: A Multi-Functional Gesture-Controlled Virtual Instrument, *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME2023)*, pp. 536–541 (2023).
- [15] Magnusson, T. and Mendieta, E. H.: The Acoustic, the Digital and the Body: A Survey on Musical Instruments, *Proceedings of the 2007 Conference on New Interfaces for Musical Expression (NIME07)*, pp. 94–99 (2007).
- [16] Yonebayashi, Y., Kameoka, H. and Sagayama, S.: Automatic decision of piano fingering based on hidden Markov models, *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pp. 2915–2921 (2007).
- [17] Gan, Q., Wang, S., Wu, S. and Zhu, J.: PIANOMOTION10M: Dataset and Benchmark for Hand Motion Generation in Piano Performance, *International Conference on Learning Representations (ICLR2025)* (2025). Accepted as a conference paper at ICLR 2025.
- [18] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, S., Sung, C.-L., Chang, M. G. and Grundmann, M.: MediaPipe Hands: On-device Real-time Hand Tracking, *arXiv preprint arXiv:2006.10214* (2020).
- [19] Wijaya, F., Tseng, Y.-C., Tsai, W.-L., Pan, T.-Y. and Hu, M.-C.: VR Piano Learning Platform with Leap Motion and Pressure Sensors, *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW2020)*, IEEE, pp. 585–586 (2020).

- [20] Nakamura, E., Saito, Y. and Yoshii, K.: Statistical Learning and Estimation of Piano Fingering, *Information Sciences*, Vol. 517, pp. 68–85 (2020).
- [21] 産業技術総合研究所 (AIST): AIST 日本人の手の寸法データ-寸法項目一覧 (2010).
<https://www.airc.aist.go.jp/dhrt/hand/data/list.html>,
最終閲覧: 2025-08-02.
- [22] Zhang, Q., Park, A. and Suh, K.-H.: Relationship between Playing Musical Instruments and Subjective Well-Being: Enjoyment of Playing Instruments Scale, *Behavioral Sciences*, Vol. 14, No. 9, p. 744 (2024). Article 744, total 14 pages.
- [23] Shipman, D.: A Prescription for Music Lessons, *Federal Practitioner*, Vol. 33, No. 2, pp. 9–12 (2016).
- [24] Meng, C.: Expressive Movements in Piano Performance: The Inducing Factors, *Journal of Human Movement Science*, Vol. 5, No. 1 (2024).
- [25] James, B.: Pianism: Performance Communication and the Playing Technique, *Frontiers in Psychology*, Vol. 9, No. 2125 (2018).
- [26] Tsay, C.-J.: Sight over sound in the judgment of music performance, *Proceedings of the National Academy of Sciences*, Vol. 110, No. 36, pp. 14580–14585 (2013).
- [27] Aroso, N.: Role of Movement and Gesture in Communicating Music Expressiveness to an Audience: An experiment on dynamics perception after a contemporary percussion performance, *Convergences - Journal of Research and Arts Education*, Vol. 13, No. 25, pp. 83–94 (2020).