Vol.2021-MUS-132 No.19
2021/9/17

フルート演奏ロボットを用いた機械学習による フルート演奏パラメータ推定

黒田 迅[†1,a)]　上瀧 剛[†1]

概要：楽器の練習を補助するシステムはこれまでに多く開発されている．当研究室でも，楽器演奏ロボットを利用し，演奏録音を行うだけで具体的な指導を行えるシステムの開発を行ってきたが，従来の研究はリコーダーを対象としており，息の流量のみを推定するものであった．そこで本研究では，より複雑な演奏パラメータを持つフルートを対象として，息の流量と息の角度の 2 パラメータを推定するモデルを提案する．提案モデルでは，フルート演奏ロボットの演奏録音と人間の演奏録音を同時に学習に用い，人間の演奏に適応させた．また，学習したモデルに人間の演奏録音を入力し出力を観察したところ，未経験者と初心者，中級者で出力結果に違いが見られた．

1. はじめに

楽器演奏の習得において，自主練習の質は重要なものである．そのため，自主練習の質を向上させるための練習補助システムがこれまでに多く提案されている．しかし，従来の練習補助システムには，具体的な指導を行えない，設備が大掛かりになるなどの問題点があった．

そこで，当研究室では，演奏録音から具体的な指導を可能にするシステムの開発を目指している．これを実現するために，演奏録音から演奏時の身体パラメータを推定する機械学習モデルを作成し，推定結果を初心者と上級者で比較するという手法を提案している．提案手法では，息の量や角度といった身体パラメータが人間の演奏からは取得が困難であることに対して，楽器演奏ロボットの演奏時の制御パラメータを身体パラメータとみなすことで取得を可能とした．この手法を用いて，著者らはこれまでにリコーダーを用いた息の流量 1 パラメータ推定システムの開発を行い，人間の演奏時の主観的な流量と推定結果の間に相関があることを確認してきた．

本研究では，新たにフルートを対象楽器とした，初学者向け 2 パラメータ推定システムを提案する．はじめにロボットと人間の演奏録音からデータセットを作成し，パラメータ推定モデルを学習を行った．次に，データセットに含まれていない複数人の演奏者による演奏をモデルに入力し，推定結果を観察した．

[†1] 現在，熊本大学
Presently with Kumamoto Uniersity
[a)] kuroda@navi.cs.kumamoto-u.ac.jp

2. 関連研究

2.1 従来の練習補助システム

2.1.1 演奏評価システム

音色などの評価点に対して，プロの演奏家などと同等の評価を下すことのできるシステムが Picas らによって開発されている [1]．これは，演奏に対しプロの演奏家が評価をしたものをデータセットに使用し，機械学習等でプロと同じような評価を行えるシステムを構築するものである．このシステムを利用すると，学習者は音声をシステムに入力することで，演奏に対する評価を受け取ることができる．

2.1.2 音色の視覚的フィードバック

音色の違いを特徴として抽出して，視覚的にフィードバックを行うシステムも Kimura らによって開発されている [2]．このシステムでは，奏法による微妙な音色の変化を二次元の特徴として抽出する．抽出した二次元の特徴をグラフにプロットして表示することで，音色を視覚的にフィードバックすることが可能となっている．そのため，学習者は自身の演奏やお手本の演奏のプロットを見て，お手本に近づくように練習を行うことができる．

2.1.3 バイオリンの運弓動作指導システム

センサを用いて身体的な動作や物理的な量を取得することで，具体的な指導を目指すシステムも Miyasato らによって提案されている [3]．このシステムでは，弦楽器において，腕の動きを検出する加速度センサやモーションセンサを演奏者や楽器に取り付け，演奏時の腕の動作を取得する．取得した動作をプロの動作などの基準となる値と比較することで，指導を行うことができる．このシステムを利

用すると，具体的な指導が演奏の改善につながるため，自主練習の効率を向上させる事ができる．

2.2 リコーダー練習補助システム

こうした従来の練習補助システムには，演奏録音のみで動作することと，具体的な指導を行うことを両立できないという問題点が存在した．そこで，当研究室では，楽器演奏ロボットによる演奏をデータセットとして用いることで，録音のみから演奏パラメータを推定し，具体的な指導を行うことのできる練習補助システムを提案した [4]．提案した手法では取得が困難な人間の演奏時の身体パラメータの代わりに，楽器演奏ロボットの制御パラメータを利用することで，学習データセットを作成している．これにより，録音のみから演奏時の身体パラメータを推定する回帰モデルを構築することが可能となっている．このシステムはリコーダーの演奏練習を対象としている．推定するパラメータは息の流量 1 パラメータとなっており，モデルから出力された推定結果は，人間の主観に基づいた息の流量と相関があることが確認された．

2.3 本研究のフルート練習補助システム

今回提案する練習補助システムは，従来のリコーダーモデルよりもパラメータ数が多く，複雑な楽器を対象としたものである．対象とする楽器はフルートとし，息の流量と息の角度という 2 パラメータを推定するものとなっている．

3. 提案手法

図 1 に提案する練習補助システムを示す．提案システムは，演奏録音を用いて機械学習モデルを学習させる部分と，学習させたモデルをシステムとして用いて，実際に演奏指導を行う部分に分かれている．

機械学習モデルの学習には，演奏録音を用いる．楽器を演奏することのできるロボットの演奏を，息の量や体の動作といった，演奏時のロボット制御パラメータとともに記録し，学習用のデータセットとする．そして，データセットを使用し，演奏から制御パラメータを推定することのできるモデルを構築する．その後，モデルに初心者と上級者の演奏録音を入力し，推定結果を比較することで，具体的な演奏指導を行うことが可能になっている．

また，今回の提案モデルでは従来と異なり，人間による演奏録音もデータセットとして使用し，同時に学習を行った．これにより，人間の演奏に適応したモデルの作成を行った．

本発表では，はじめにデータセットとその作成に使用したロボットや器具の説明を行う．その後，使用した機械学習モデルの説明を行う．最後に，学習済みのモデルに複数人の演奏録音を入力した際の出力結果の観察を行う．

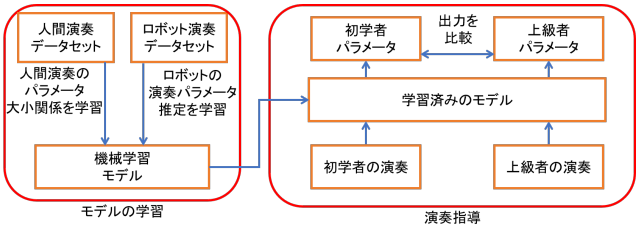


図 1: 提案する練習補助システム．ロボットと人間の演奏データを用いて，演奏時のパラメータを推定するモデルを機械学習を用いて学習する．その後，学習したモデルに人間の演奏を入力することで，人間のパラメータをロボットに置き換えて出力することができる．

4. データセット

4.1 ロボット演奏データセット

使用したロボット

データセットの作成には，著者らの研究室で作成しているフルート演奏ロボットを使用した．図 2 にデータセット作成用のフルート演奏ロボットを示す．このロボットは，アクリル板と 3D プリントされた部品で作成されており，フルートの頭部管が取り付けられている．演奏の際に送り込まれる息は比例流量弁によって無段階に調節を行うことが可能である．また，息の当たる角度は，サーボモータによってロボットに取り付けられているフルートを回転させることで調節できるようになっている．

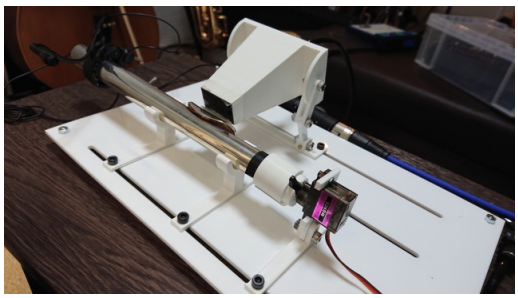


図 2: 作成したフルート演奏ロボット．このロボットは比例流量弁の制御値を変えることで送り込まれる空気の量を調節することができる．また，サーボモータを用いてフルートを回転させることで空気の当たる角度を調節することができる．

使用する音

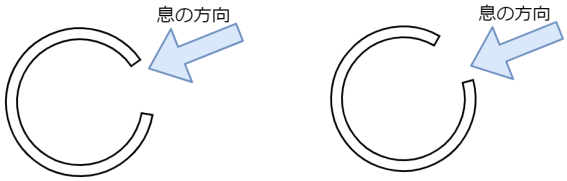
初心者がフルートを練習する際に，頭部管と呼ばれる部分のみを用いて練習を行うことがある．そのため，頭部管のみを用いた演奏に焦点を当て，システム開発を行うこととした．

録音環境

フルート演奏ロボットの録音は，防音室内で自動で行った．録音マイクには JTS MA-500 を使用し，フルート本体に固定した．

データ数と入力形式

10 秒間の録音を 1 データとし，録音を行った．パラメータは角度制御値 15 段階，流量制御値 26 段階とし，一つのパラメータにつき 5 データを作成した．図 3 にフルートを左側面から見た際の角度の大小方向について示す．フルートを手前側に，左側面から見て右向きに回転させると角度の値は大きくなり，反対に回転させると角度の値は小さくなる．また，フルートを演奏する際に生じる音程の変化に対応するために，半音を 1 としたときに 0.1，0.2，0.3 だけピッチを下げたデータも作成し，合計で 7800 データを用意した．モデルへの入力には，録音音声から作成されたメル周波数 128 の対数メルスペクトログラムを 128 × 128 の大きさになるようにランダムに時間方向から切り出したものを使用した．



(a) 角度大 (UP)　　**(b)** 角度小 (DOWN)

図 3: フルートの角度とパラメータの値．演奏者から見て手前に回転させると角度が大きく，奥に回転させると角度が小さくなるとした．また，人間の演奏に関しては，角度が大きいものを「UP」，角度の小さいものを「DOWN」と定義した．

4.2 人間演奏データセット

提案手法では，ロボットの演奏データの他に，人間の演奏データも学習に使用した．

録音に使用した器具

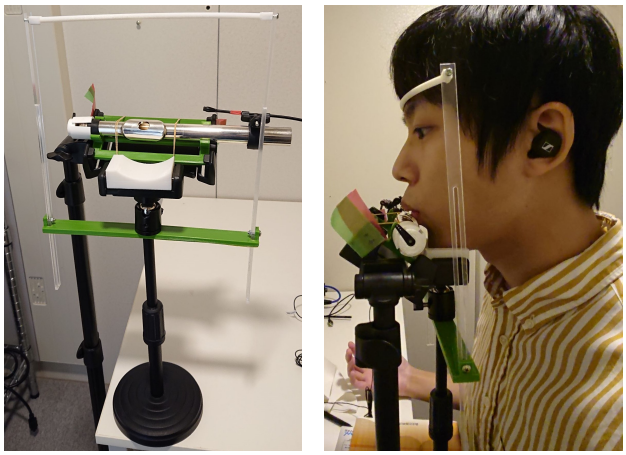
人間演奏録音を行う際，息の角度を固定するために器具を使用した．図 4 に使用した器具を示す．器具はスマートフォン用のスタンドにアクリル板と 3D プリンターで作成した部品を取り付けて作成した．この器具で額と顎を固定し，演奏を行うことで，フルートに入る息の向きを固定することができる．

録音環境

録音はロボットデータセットの作成と同じく防音室内で行い，マイクは JTS MA-500 を使用した．

データ数と入力形式

著者が演奏を行い，0.5 秒間の演奏を 1 データとして録音を行った．パラメータは角度 2 段階，流量 2 段階とした．人間演奏の角度は，図 3 に示すように手前側に回転させた場合を「UP」，反対に回転させた場合を「DOWN」としている．UP と DOWN の角度差は約 10°であり，器具に取り付けたフルートを回転させて調整した．また，人間演奏



(a) 作成した器具　　**(b)** 使用している様子

図 4: 作成した頭固定用の器具．スマートフォンスタンドにアクリル板と 3D プリントした部品を取り付けて作成している．この器具を使用して演奏することで，フルートに当たる息の角度を固定することができる．

の流量は強く吹く場合を「F」，弱く吹く場合を「P」とした．人間演奏の流量は演奏者の主観に基づいて吹き分けを行った．データはそれぞれのパラメータで約 80 個ずつ，合計約 320 個である．モデルへの入力には，0.5 秒間の録音音声から 128 × 128 の対数メルスペクトログラムを作成して使用した．

5. 演奏パラメータ推定モデル

5.1 モデル構造

図 5 に，提案手法の機械学習モデルを示す．構築したモデルは 2 パラメータを推定するものであり，パラメータごとに使用するモデルを分割し，別々に推定を行う．パラメータ推定モデルには，MLP Mixer[5] を使用した．MLP Mixer は 2 つの MLP 層からなる単純なモデルであり，画像を複数のパッチに分割し，分割したパッチ内での特徴抽出とパッチ間での特徴抽出を分けて行う．MLP Mixer は CNN や Attention 等を利用したモデルより計算数を抑えつつ，同程度の性能を得ることが可能である．MLP Mixer の実装には pytorch 用の MLP Mixer ライブラリ[*1]を用いた．モデルの最終的な出力は sigmoid 関数を通して 0 から 1 の値で出力される．

5.2 損失関数

提案するモデルでは，ロボットの演奏録音と人間の演奏録音を同時に学習に使用している．

ロボットの演奏データは正解となる流量と角度の制御パラメータがわかっているため，制御パラメータを正解の値とした回帰学習を行うことができる．そのため，ロボットデータの損失関数には平均二乗誤差 (MSE) を利用した．

[*1] https://github.com/lucidrains/mlp-mixer-pytorch

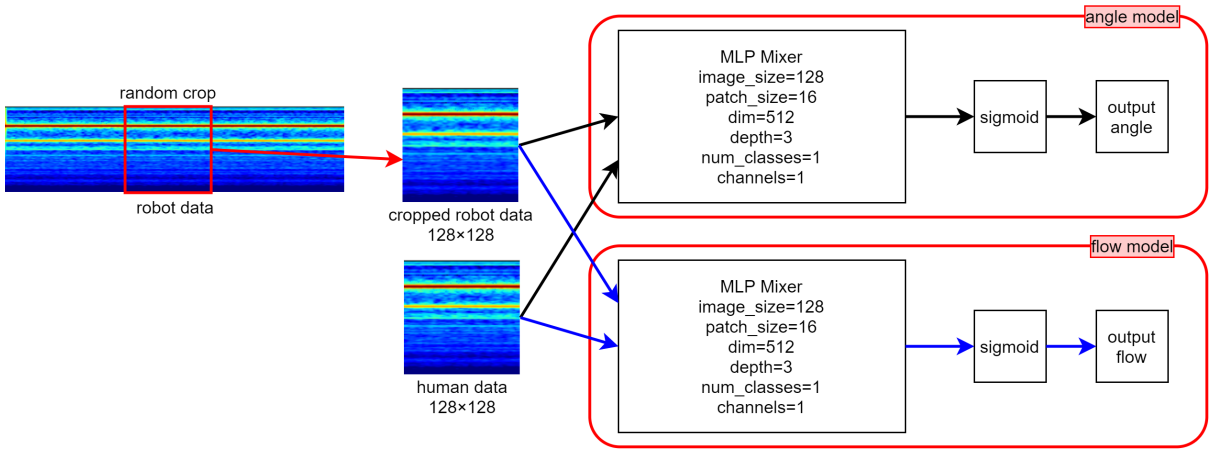


図 5: モデルの構成．モデルは MLP Mixer をベースとした回帰モデルであり，流量推定と角度推定はモデルを分けて行う．入力にはロボット演奏と人間演奏から作成した 128 × 128 の対数メルスペクトログラムを用いる．出力は 0 から 1 の値をとり，1 に近づくほど流量，角度ともに大きな値となる．

これに対して，人間の演奏データは演奏パラメータの大小はわかっているが，正解となる値はわからない．そこで，データの大小関係を学習するランキング学習手法である RankNet[6] の損失関数を導入した．RankNet では，2 つのデータ間のスコアが，ラベルの大小と一致するように学習を行う．今回使用するラベルは，流量が「F」「P」の 2 種類，角度が「UP」「DOWN」の 2 種類であり，それぞれ前者がラベル大，後者がラベル小となっている．

以上から，今回導入した損失関数は以下のようになる．

$$\text{loss} = \text{MSE}(o_r, t_r) + \text{RankNet\_loss}(o_{h1}, o_{h2}, P_{h1h2})$$

ここで，o_r はロボットデータ入力時のモデル出力，t_r はロボットの制御パラメータ，o_{h1} と o_{h2} はそれぞれ人間データ入力時のモデル出力，P_{h1h2} は 2 つの人間データ間のパラメータ大小関係である．

6. 評価実験

6.1 モデルの学習結果

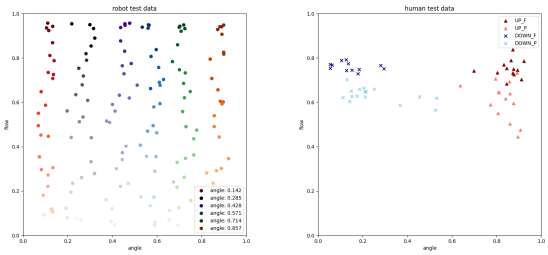
作成したデータセットでモデルの学習を行った．最適化関数には Adam を使用し，学習率は 0.0001 とした．100 エポック学習を行い，ロボットと人間演奏のテストデータを入力し，その結果をグラフにプロットした．

図 6 にそれぞれの結果を示す．グラフは横軸が角度の推定値，縦軸が流量の推定値となっている．ロボット演奏データについては，各プロット点の色が角度ラベルを表し，色の濃さが流量ラベルを表している．流量ラベルは色が濃いほど流量が大きく，色が薄いほど流量が小さいものとなっている．また，人間演奏データについては表 1 にそれぞれのラベルと色の関係を示す．

図 6 より，ロボット演奏と人間の演奏どちらにおいても正しく学習が行えていることが確認できた．

表 1: 人間演奏データのラベルと色の関係

角度	流量	色
DOWN	P	水色
DOWN	F	青
UP	P	ピンク
UP	F	赤



(a) ロボットの演奏　　**(b)** 人間の演奏

図 6: テストデータのプロット結果．プロットされた点は色相が角度ラベルの違いを表し，濃淡が流量ラベルの違いを表している．プロット結果から，ロボットデータ，人間データどちらも正しく学習できていることがわかる．

6.2 複数人の演奏の観察

次に，データセットに含まれていない複数人の演奏者による演奏を録音し，そのデータを入力した際の出力結果を観察した．演奏は P1 から P5 の 5 人の参加者に行ってもらった．参加者のうち P1 は中級者，P2 と P3 は初心者，P4 と P5 は未経験者である．録音にはデータセット作成に使用した固定器具を用いた．また，演奏するパラメータも人間演奏データセット作成時と同じものである．

6.2.1 中級者の演奏

図 7 に中級者 P1 の演奏に対する出力結果を示す．流量は異なるパラメータ間で区別されていることが確認できた．また，角度はフルートの固定角度に関わらず近い位置

表 2: 各参加者の推定パラメータの分散

参加者		分散 [10^{-3}]									
番号	習熟度	DOWN F		DOWN P		UP F		UP P		AVERAGE	
		角度	流量	角度	流量	角度	流量	角度	流量	角度	流量
P1	中級者	7.45	0.90	0.76	0.44	10.19	1.90	1.78	4.39	**5.05**	**1.91**
P2	初心者	0.15	1.17	0.29	0.84	11.27	1.36	2.50	3.09	**3.55**	**1.61**
P3	初心者	1.10	0.49	9.40	1.19	6.74	1.15	5.58	4.17	**5.71**	**1.75**
P4	未経験	5.17	2.65	7.87	1.25	9.14	0.70	10.37	1.36	**8.14**	**1.49**
P5	未経験	5.22	1.46	9.42	2.21	5.47	1.93	30.52	3.23	**12.66**	**2.21**

にプロットされた．これは，中級者は出したい音が明確であり，フルート自体の角度を変えたとしても狙った位置に息を吹き込む事ができているためだと考えられる．

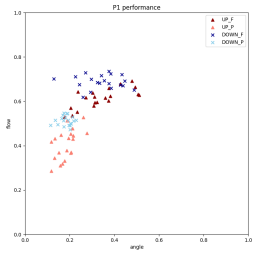
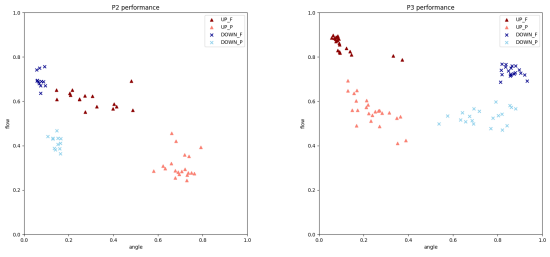


図 7: 中級者 P1 の演奏結果．流量は区別されているが，角度はパラメータによらず近い値をとっている．

6.2.2 初心者の演奏

図 8 に初心者 2 人の演奏に対する出力結果を示す．2 人ともそれぞれのパラメータを明確に吹き分けることができていた．P2 の演奏については角度 UP，流量 F の演奏が中央に寄っている．P2 によると，これは音をよく鳴らすために息の角度を意識的に変えたからであり，この結果は主観的なイメージと一致しているということだった．また，P3 の演奏に関しては角度が真逆にプロットされてしまった．P3 の演奏は音程が低く，データセット内の音域を超えてしまっていたためうまく推定できなかったと考えられる．

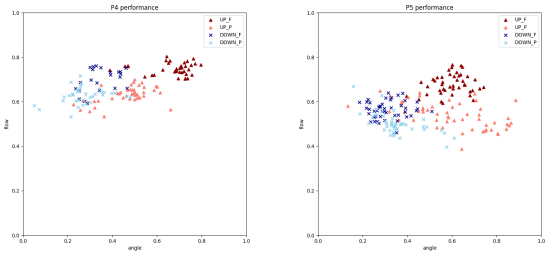


(a) 参加者 P2 の演奏　　(b) 参加者 P3 の演奏

図 8: 初心者の演奏結果．流量，角度ともに明確に吹き分けができている．しかし，P3 の演奏の推定結果は角度の大小が逆にプロットされた．P3 の演奏は音程が低く，今回用意したデータセットでは対応できなかったと考えられる．

6.2.3 未経験者の演奏

図 9 に未経験者 2 人の演奏に対する出力結果を示す．2 人とも経験者ほどパラメータの吹き分けができておらず，角度方向に広く分布する結果となった．これは，未経験者は音を出すコツなどを掴みきれておらず音が安定せず，音の鳴るパラメータを探りながら演奏しているためだと考えられる．



(a) 参加者 P4 の演奏　　(b) 参加者 P5 の演奏

図 9: 未経験者の演奏結果．中級者や初心者と比べ，流量，角度ともにうまく吹き分けができておらず，角度方向に分布が広がっている．

6.3 複数人の演奏の定量的評価

P1 から P5 の演奏を定量的に評価するために，分散と群平均法による距離計算を用いた．

6.3.1 分散による定量的評価

各パラメータをどれだけ明瞭に演奏できたかの定量的な指標として，推定されたパラメータの分散を用いた．表 2 に，各参加者の演奏パラメータごとの角度と流量の分散を示す．分散が小さいほど，各パラメータを明確に，安定して演奏することができたと考えられる．

表 2 より，角度については，経験者 P1，P2，P3 の分散の平均値が未経験者 P4，P5 の値よりも小さくなった．これは，フルートの経験者は音を出す際のコツを掴めているため，鳴らす音にばらつきが少ないためだと考えられる．これに対し，流量については未経験者と経験者の間で大きな差はなかった．これは，流量を安定させることが角度を安定させることよりも簡単であり，未経験者でもある程度は安定させることができたためだと考えられる．

6.3.2 群平均法による定量的評価

各パラメータ間で，どれだけ吹き分けを行うことができたかの定量的な指標として，群平均法によるパラメータ間の距離を用いた．群平均法とはクラスタリングに用いる距離の計算手法であり，2つのカテゴリ間のすべての組み合わせに対して距離を求め，その平均値をカテゴリの距離とする．

表3に群平均法による距離を計算するカテゴリの組み合わせと，カテゴリ間でのパラメータの同異について示す．カテゴリの組み合わせは全部で6通りあり，角度，流量共に同じラベルであるものが2つ，異なるラベルであるものが4つとなっている．これらの組み合わせをラベルの同異で分け，それぞれの距離の平均値を比較することでパラメータによる吹き分けができているかを評価する．

表4に，パラメータのラベルが同じ場合と異なる場合でのそれぞれの平均値を示す．same はカテゴリの組み合わせのうちパラメータが一致するものの平均値，different はパラメータが異なるものの平均値となっている．演奏者がパラメータによる吹き分けが行えている場合は，same の値よりも different の値が大きくなると考えられる．

表4より，参加者 P2，P3，P4 及び P5 について，角度，流量ともに同じパラメータでの距離よりも異なるパラメータでの距離の方が大きくなった．これにより，それぞれの奏者が角度や流量の吹き分けを行えていることがわかる．また，P1 については角度のパラメータによる距離差が同一パラメータの方が大きくなっていた．これは，P1 は中級者であり，フルートの角度によらず狙った位置に息を当てることができていることに加え，流量によって息を当てる位置を調節しているためだと考えられる．

表 3: カテゴリの組み合わせとパラメータの同異

カテゴリ 1		カテゴリ 2		角度の同異	流量の同異
角度	流量	角度	流量		
DOWN	P	DOWN	F	same	different
DOWN	P	UP	P	different	same
DOWN	P	UP	F	different	different
DOWN	F	UP	P	different	different
DOWN	F	UP	F	different	same
UP	P	UP	F	same	different

表 4: パラメータの同異による距離の平均

参加者		距離 $[10^{-1}]$			
番号	熟練度	角度		流量	
		same	different	same	different
P1	中級者	**1.62**	1.18	0.94	**1.92**
P2	初心者	2.36	**3.90**	1.00	**2.91**
P3	初心者	1.38	**6.23**	0.92	**2.50**
P4	未経験	1.65	**2.65**	0.50	**0.99**
P5	未経験	1.27	**2.73**	0.92	**1.13**

7. まとめ

本研究では，提案するフルート練習補助システムについての解説と，複数人の演奏に対するモデルからの出力の観察を行った．提案しているシステムは息の流量と角度という2パラメータを推定する複雑なものであり，ロボット演奏データに加えて人間データも学習に使用することで人間のデータに適応したモデルの生成を行った．また，複数人の演奏録音をモデルに入力したところ，演奏者の習熟度に応じて出力結果に違いが見られたが，音程の低すぎる演奏に対しては，正しく推定できなかった．今後は GUI を作成し，実際にシステムを使用した練習実験を行い，システムの有効性を検証していきたいと考えている．

参考文献

[1] Picas, O. R., Rodriguez, H. P., Dabiri, D., Tokuda, H., Hariya, W., Oishi, K. and Serra, X.: A real-time system for measuring sound goodness in instrumental sounds, *Journal of The Audio Engineering Society* (2015).

[2] Kimura, N., Shiro, K., Takakura, Y., Nakamura, H. and Rekimoto, J.: SonoSpace: Visual Feedback of Timbre with Unsupervised Learning, *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 367–374 (2020).

[3] 洸司宮里，聖也大城，健太郎野口，志穂子神里：慣性センサによるバイオリンの運弓動作指導の検討，情報科学技術フォーラム講演論文集，Vol. 10, No. 3, pp. 783–784 (2011).

[4] 黒田 迅，上瀧 剛：ロボットと機械学習による楽器演奏パラメータ推定と練習補助システムへの応用，第128回音楽情報科学研究発表会，No. 2 (2020).

[5] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M. and Dosovitskiy, A.: MLP-Mixer: An all-MLP Architecture for Vision, *arXiv preprint arXiv:2105.01601* (2021).

[6] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G.: Learning to Rank Using Gradient Descent, *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96 (2005).