2C3-OS-20a-4

演奏未経験者のためのスマートフォンセンサーを用いた即興合奏支援システムの試作

Prototype Support System of Improvisational Ensemble Using Smartphone Sensors for Person who have the experience to perform music

水野 創太[*1] 一ノ瀬 修吾[*1] 白松 俊[*1] 北原 鉄朗[*2]

Mizuno Souta Ichinose Shugo Siramatsu Shun Kitahara Tetsurou

[*1] 名古屋工業大学 工学部 情報工学科 [*2] 日本大学 文理学部 情報科学科

[*1] Department of Computer Science, School of Engineering, Nagoya Institute of Technology

[*2] Department of Information Science, College of Humanities and Sciences, Nihon University

We aim to develop a system enabling non-experts of musical performance to play improvisational ensemble. Although non-experts of music cannot determine a pitch suitable for a particular chord, they can express pitch contour by their body motion. Our previous system recognizes users' body motion on the basis of a sensor camera. In this paper, we aim to develop a method for recognizing users' body motion on the basis of smartphone sensors in order to enabling multiple users to easily join improvisational ensemble. We collected training data of correspondence between melody and body motion. Moreover, we developed three Bayesian Network models for estimating pitch name, vertical movement of pitch, and attack time.

1. はじめに

旋律歌唱や旋律聴取における 3 つの処理側面として，リズム，旋律線，調性がある[1]. これら 3 要素のうち，音楽未経験者の即興合奏を難しくするのは調性判断である. そのため本稿では，音楽未経験者が即興合奏するうえで困難となる調性判断をシステムが行い，ユーザの身体動作からスマートフォンに搭載されたセンサーを用いてリズムと旋律概形 (pitch contour) の入力を行うことで，スマートフォンを用いた音楽未経験者による即興合奏を支援するシステムの開発を目指す. なお，菅[2]は旋律線の手なぞりなどの身体動作が音楽理解を促進する方法 として有効であるとしていることから，旋律概形の入力 にユーザの身体動作を用いる手法が有効である.

本システムを実装するために，まずはスマートフォンに搭載されたセンサー（加速度センサー，ジャイロセンサー等）で計測した値からスマートフォンの上下動をト レースする単純なポジショントラッキング手法を実装した. しかし，この手法でのユーザの手の上下動の推定精度が低いことが課題となった.

そのため本稿では，北原らの手法[3]を参考に，ベイジアンネット ワークの確率モデル推定を用いて，トレースした動きのデータを訓練データとしてユーザの動き，その際に出力するべき音高を推測する手法を提案する.

また，本稿では予備段階として背景楽曲として流れている曲とのセッションを目指し，流れている曲に合わせてスマートフォンを動かすことで背景楽曲のコード進行，ユーザの動きに適した音が出力されることを目的とし，ポジショントラッキングによってトレースしたユーザの動きからベイジアンネットワークによって発音タイミング，出力するべき音名，ユーザ動作の上下動を推測する手法について述べる.

連絡先：氏名，所属，住所，電話番号，Fax 番号，電子メイルアドレスなど

2. ベイジアンネットワーク

本稿では，センサーによって計測された値を入力値としたベイジアンネットワークを用いて出力する音高やタイミングを決定する. その推定にあたり本稿では，音の出力を以下の 3 つの要素に分け，3 つのベイジアンネットワークを用いる手法を提案した.

- 発音タイミング
- 出力する音名
- 音高の上下動

2.1 訓練データの作成

本稿では，ベイジアンネットワークを作成するための訓練データサンプルを作成するため実際に，背景楽曲に合わせてユーザがスマートフォンを動かした際，約 5ms ごとの加速度 a，速度 v，速度の変化量 vc，移動距離 p，重力加速度 g を取得する予備実験を行った. 本稿では，約 5ms ごとに取得した値を 1 サンプルとする. 実験を行う際，発音タイミングの指定方法として以下の 2 つの方法を仮定した.

1. シェイク動作: スマートフォンを振るシェイク動作によって発音タイミングを指定（加速度，ジャイロ センサー等）
2. タッチ動作: スマートフォンの画面内に配置されたボタンにタッチして発音タイミングを指定

これらの実験をそれぞれ被験者 5 人に対して行い，約 65000 サンプルが得られた. 得られたサンプルから各ベイジアンネットワークに適当な特徴量を用いることでベイジアンネットワークモデルを作成するための訓練データを作成した.

2.2 発音タイミングの推定

図 1 に発音タイミングを推定するベイジアンネットワークモデルを示す. このモデルでは，ポジショントラッキングによって求めた加速度 a，速度 v，速度の変化量 vc，重力加速度 g を入力値として音を出力するタイミングを表す t を{0,1}によって推測する.

正しい発音タイミングから±30ms のサンプルを発音タイミングとして扱うものとする. 発音タイミングの推測の場合, 加速度を x 軸方向の加速度 a_x, y 軸方向の加速度 a_y の 2 種類を用いる.

なお, ボタンにタッチして発音タイミングを指定する方法については, センサーを用いておらずシステムによって配置したボタンタッチに連動して確実に発音タイミングを指定できるため, 発音タイミングを推定できるため, 発音タイミングを予測する必要がないものとする.

2.3 半音毎の音名推定

図 2 に出力する音名を推定するベイジアンネットワークモデルを表す. ポジショントラッキングによって加速度 a, 速度 v, 速度の変化量 vc, 重力加速度 g に加えて移動距離 p, 発音タイミングモデルの予測結果となる t と直前 m サンプルの予測結果から一番多く予測された結果を表す r_m, 1 つ前のノートの音名を表す n_{i-1} を入力値として用いて出力するノートの音名 n_i を予測する.

スマートフォンの画面内に配置されたボタンをタッチする演奏方法の場合, 発音タイミングを予測するモデルを用いていないため, 発音タイミングの予測結果を表す t は用いないものとする.

2.4 音高の上下動推定

図 3 にユーザ動作の上下動を推定するベイジアンネットワークモデルを表す. ポジショントラッキングによって求めた加速度 a, 速度 v, 速度の変化量 vc, 重力加速度 g, 加えて移動距離 p を入力値として用いて, 1 つ前の発音時からユーザの持つスマートフォンの上下動を表す h を{変化なし, 上昇, 下降} = {0, 1, 2}によって推測する.

このベイジアンネットワークによって推定されたユーザ動作の上下動は出力する音高の上下動を表し, ユーザの上下動と出力する音名の推定結果から出力する音のオクターブを変化させる.

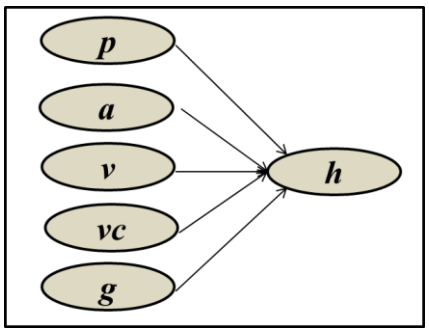


図 1 発音タイミング推定のためのベイジアンネット

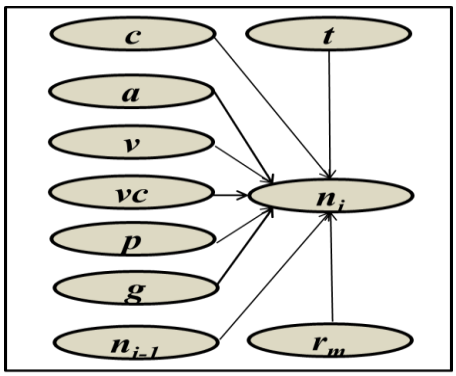


図 2 出力する音名推定のためのベイジアンネット

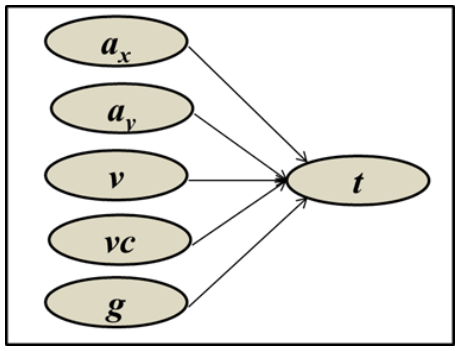


図 3 ユーザ動作の上下動推定のためのベイジアンネット

3. ベイジアンネットワークの予測精度評価

2.1 の予備実験で得た訓練データサンプルから作成した(1)発音タイミング, (2)出力する音名, (3) ユーザ動作の上下動の 3 つのベイジアンネットワークについて学習に用いた訓練データと同じ訓練データを用いて, それぞれの予測精度を確認した.

3.1 発音タイミングの推定精度

発音タイミングについてモデルが判定したノー ト数の再現率と適合率を表 1 に示す. 本稿における再現率とは, 発音タイミングから±30ms を 1 つのノートの発音タ イミングとして, システムが予測した発音タイミングと 本来の発音タイミングが一致している数と本来の発音タ イミングの割合とする.

また, 適合率とは, システムが予測した発音タイミングから 60ms までを 1 つの発音タイミングとして, シス テムが予測した発音タイミングと本来の発音タイミングが一致している数とシステムが予測した発音タイミングの数の割合とする. 表 1 では, 再現率と比較し適合率の精度が極端に低下していることから, このベイジアンネットワークによる発音タイミング予測では, 本来の発音タイミングではないタイミングで誤って発音タイミングと予測される場合が多いことがわかる.

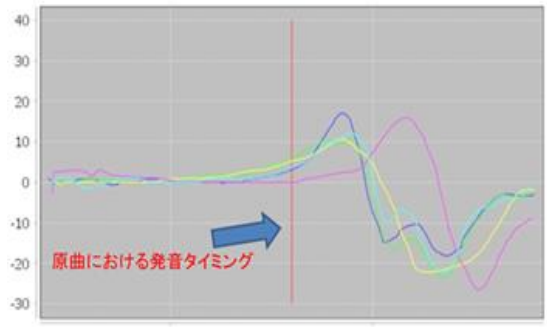
図 4 に訓練データに用いたデータサンプルにおける 1 つ目のノートの発音タイミングにおける, 発音タイミングと各被験者の x 軸方向の加速度の変化を時系列で表したグラフを示す. このグラフよりそれぞれの被験者の意図する発音タイミングと本来の発音タイミングにずれが生じていることがわかる.

したがって現在, 原曲における本来の発音タイミングから±30ms を各ノートにおける発音タイミングと定義しているが, 各被験者のデータサンプルの値から閾値を定義し, その値に基づいた発音タイミングを定義する必要があると考えられる.

表 1 発音タイミング推定の再現率と適合率

	再現率	適合率
シェイク動作	0.63 (171/270)	0.26 (171/661)

図 4 発音タイミングと被験者の x 軸方向の加速度の関係図



3.2 出力する音名の推定精度

出力する音名推定実験では，m=10 に設定し，直前 10 サンプルの予測結果のうち一番多く予測された結果を m の値とした

表 2 に，原曲と全く同一の音が推定されたサンプル数の割合を示す.

表 3 に，原曲と全く同一の音が推定されたノート数の割合を示す. これは，発音タイミングの最初のサンプルが原曲と同一の音名だった場合を正解とした精度である.

表 2,3 の結果について，表 2 と表 3 の精度を比較した際，表 3 における精度が大きく低下していることから発音タイミングの瞬間の 1 サンプルにおける予測結果の正答率が低く，音が持続している間に予測結果の正答率が高まる傾向にあることが予測される. そこで本稿では，発音タイミングの瞬間の 1 サンプルだけの予測結果ではなく，発音タイミングの瞬間から数サンプルを考慮することで精度向上の可能性があると考えて手法の改善を試みたが，実験の結果，表 3 からの顕著な精度向上は認められなかった. これは，発音タイミングの最初のサンプルから 9 サンプルまでの r_{10} は 1 つ前のノートの予測結果を考慮してしまっているために間違った予測情報を用いているためであると考えられる.

ここでは原曲と全く同一の音名を予測した精度を示したが，本研究の目的は原曲と同一の音名を予測することではなく，コード進行と不協和にならず，ユーザ動作に合った音高を出力することである. よって，本来はコードと不協和にならない音名を正解とすべきであり，精度も表 2, 3 の値より高くなるはずであるが，これについては 今後の課題とする.

表 2 出力する音名推定の精度 (サンプル単位)

	原曲と同一の音が予測されるサンプル数/全体のサンプル数
シェイク動作	0.70 (45888/65278)
タッチ動作	0.82 (53764/65251)

表 3 出力する音名推定の精度 (ノート単位)

	原曲と同一の音が予測されるノート数/全体のノート数
シェイク動作	0.45 (121/270)
タッチ動作	0.56 (152/270)

3.3 ユーザ動作の上下動推定

ユーザ動作の上下動についてモデルが本来のユーザ動作の上下動と同じ上下動の推定を行うサンプル数の割合を表 4，ノート数の割合を表 5 に示す. 本稿では，発音タイミングの最初のサンプルが正しい上下動を判定している場合，そのノートは正しい上下動が推定されているものとして扱う.

3.2 で言及したように本研究では，原曲と同一の音を出力することではなく背景楽曲のコード進行と不協和にならず，ユーザ動作に合った音を出力することである. 従って，ユーザ動作の上下動を正しく推定できるようになることが本研究における本来の目的において重要であると考えられる.

そこで本稿では，ポジショントラッキングによって求める移動距離 p の定義を変更することによってベイジアンネットワークによる推定精度の向上を試みた. 現在の定義と提案した定義を以下に示す.

移動距離(旧)=演奏開始時からの移動距離

移動距離(新)=1 つ前の発音からの移動距離

この手法での移動距離の推定は発音タイミングの決定を確実に行う必要があるため本稿では，この移動距離の定義を発音タイミングの決定を確実に行うことのできるタッチ動作において用いることでその予測精度の変化を確認した.

表 4 ユーザ動作の上下動推定の精度 (サンプル単位)

	正しい上下動予測がされるサンプル数/全体のサンプル数
シェイク動作	0.66 (42975/65278)
タッチ動作	0.78 (50909/65264)

表 5 ユーザ動作の上下動推定の精度 (ノート単位)

	正しい上下動予測がされるノート数/全体のノート数
シェイク動作	0.56 (152/270)
タッチ動作	0.68 (184/270)

3.4 新定義による予測精度

この手法では，タッチ動作による演奏において実際にボタンタッチを行った瞬間のデータサンプルのみを訓練データとして用いる. 従って，サンプルごとの予測精度がそれぞれのノートの原曲と同一の音を予測した予測結果を表す.

表 6 に出力する音名推定の精度とユーザ動作の上下動推定の精度を示す.

表より，ユーザ動作の上下動推定が高精度で行えていることがわかる. よって，この移動距離の新定義を用いることで本研究の本来の目的に沿った演奏が行えるようになる可能性が示唆された.

表 6 新定義を用いた場合の予測精度 (ノート単位)

	正しい予測がされるノート数/全体のノート数
出力する音名推定精度	0.71 (192/270)
ユーザ動作の上下動推定精度	0.91 (247/270)

4. 評価実験

本稿では実験 1 として，ベイジアンネットワークによる出力する音名と発音タイミングの推定を行うことで背景楽曲と即興合奏を行う演奏インターフェースを開発し，ユーザがあらかじめ定めた上下動作を行った際の音高とユーザの動きそれぞれの上下動が一致するかどうかの実験を(1)シェイク動作(2)タッチ動作(3)本稿における新しい定義を使った際のタッチ動作それぞれ 3 手法での実験を行った結果，音高とユーザの動きの上下動が一致しない場合が多かった.

この原因を考察した際，現在それぞれのベイジアンネットワークに用いる訓練データを作成するために用いた背景楽曲が 1 曲のみであることから原曲における音名のつながりに一定のパターンが存在ため，出力する音名予測における入力値として 1 つ前のノートの音名を表す n_{i-1} を入力値として用いていることか

らユーザの動きよりも，音名のつながりのパターンが重視されていると考えられる．

従って本稿では実験 2 として，図 2 の出力する音名推定を行うベイジアンネットワークにおいて入力値として n_{i-1} を用いない場合において実験 1 で記録したユーザの動きのデータを用いて音高とユーザの動きの上下動が一致するかどうかの実験を行った．

実験 1 と実験 2 の結果を表 7 に示す．この結果より，ユーザ動作の上下動推定が高精度で行える(3)の手法による出力する音名推定では入力値 n_{i-1} を用いないことでユーザの上下動に基づいた音高予測がさらに正確に行えるようになる可能性が示唆された．

本稿では，1 つ前のノートを入力値として用いないことでユーザの上下動に基づいた音高予測の精度を向上されることを試みたが，訓練データの作成を複数の楽曲によって行い，様々の音名のつながりのパターンを学習することでも予測精度は向上すると考えられる．従って，様々な楽曲で訓練データの作成を行うことを今後の課題とする．

表 7 各演奏手法でのユーザ動作と音高変化の一致精度

	n_{i-1}(1 つ前のノート)入力	n_{i-1}(1 つ前のノート)未入力
シェイク動作	0.48 (31/64)	0.53 (34/64)
タッチ動作	0.55 (35/64)	0.51 (33/64)
タッチ動作(新定義)	0.53 (34/64)	0.75 (48/64)

5. おわりに

本稿では，スマートフォンのセンサーとベイジアンネットワークを利用した直感的動作による演奏支援システムを開発した．背景楽曲に合わせたユーザの動きから得たセンサーの値を利用した訓練データを用いて学習することによって演奏における出力する音名や発音タイミングの推定を行うことや，ユーザの上下動と一致した音高の上下動の予測精度を向上させる可能性を示した．

今後は，複数のデバイスを用いて複数ユーザでの同時合奏機能の実装や複数楽曲による訓練データの作成，LSTM などの時系列データに適した予測手法も検討する予定である．

参考文献

[波多野 2007] 波多野誼余夫(編): 音楽と認知，Vol.8,東京大学 出版会 ,2007.

[菅 2008] 菅道子: 身体表現を取り入れた参加型音楽コンサートの可能性: カノンの理解を目指した「追いかけっこをしよう」の事例から，和歌山大学教育学部教育実践総合センター紀要 , 2008.

[北原 2009] 北原鉄朗，戸谷直之，徳綱亮輔，片寄晴弘 : BayesianBand: ユーザとシステムが相互に予測し合うジャムセッションシステム，情報処理学会論文誌，2009.