Vol.2022-MUS-134 No.8
Vol.2022-SLP-142 No.8
2022/6/17

Contrastive Loss を用いた打楽器練習用ゴムパッドの音色可視化

前田哲徳[1]　西村雅史[1]

概要：打楽器の基礎演奏能力の習得のためにゴムパッドを用いた練習が行われるが，ここでは，基礎演奏能力のなかでも均一な音を奏でる技術を扱い，音色の均一な音を奏でるためのフィードバックとして音色のずれの可視化に取り組んだ．打楽器の音色はラウドネス・ピッチを揃えた場合に２音の知覚の差に対応する属性と定義され，極めて多次元的な尺度である．今回，上級者の２音の知覚差を Contrastive Loss を用いて学習することで音色の違いをモデル化することを試みた．実験の結果，ゴムパッド音の音色のずれの可視化および自動評価について，その可能性が示唆された．

キーワード：ドラム，打楽器，パッド，音色，可視化

Timbre Visualization for Drum Training Pad with Using Contrastive Loss

AKINORI MAEDA[†1]　MASAFUMI NISHIMURA[†1]

1. はじめに

打楽器の基礎演奏技術は打楽器練習用ゴムパッドを用いた練習で習得されることが一般的である．以下，基礎演奏技術習得のためのゴムパッドでの練習を基礎練習と呼ぶ．ゴムパッドを使用している様子を図 1 に示す．アコースティックな打楽器は音が大きく，高価で，持ち運びが難しいものが多い．これに対し，ゴムパッドは音が小さく，安価で，持ち運びが容易である．ゴムパッドを用いることによって多くの打楽器奏者が様々な場所で基礎練習をすることが可能となる．

基礎演奏技術とは，所望の音が鳴るように，スティックを用いて打面を正確に叩く技術とする．ティンパニやドラム，マリンバなど様々な打楽器を演奏するために共通に必要となる技術である．中でも均一な音を鳴らす技術は重要となる．これは，演奏記号(強弱，速度などの指定)がない場合，常に均一な音を鳴らすことが求められるからである．

この均一な音を鳴らす技術の習得には，ずれを指摘する指導者が必要である．しかし，誰もが優秀な指導者に恵まれるわけではない．自宅での練習を強いられる場合も含め，指導者を伴わない練習機会は多く想定される．このような問題を解決しうるものとして，システムによって打楽器演奏練習を支援する研究，製品がある． 振動触覚 [1-2]や音[2]，筋電気刺激[3-5]を用いて正しいリズムの獲得を促す研究や，音によるリズムのずれのフィードバックを行う製品[6]がある．可視化によるフィードバックでは音の大きさ[7-8]，リズム[7-8]，フォームのずれ[7]を扱った研究があり，上達の推移を可視化する取り組み[8]もある．特に，均一な音を奏でる技術に着目すると，振動触覚や音，筋電気刺激によるフィードバックはどの程度ずれているかをフィードバックすることは難しく，行えても数段階の提示となる．一方，可視化フィードバックではずれのフィードバックをより詳細に行うことができる．しかし，音の大きさとリズムのずれを対象とした取り組みはあるが，音色のずれを扱ったものは未だない．本研究では均一な音を奏でる技術習得のためのフィードバックに利用可能な音色のずれの可視化を扱う．

図 1　ゴムパッドを使用している様子(パッドは YAMAHA トレーニングパッド 8 インチ TS01S)

1 静岡大学大学院 総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University

Vol.2022-MUS-134 No.8
Vol.2022-SLP-142 No.8
2022/6/17

音色とはラウドネス，ピッチを揃えた２音を聴いたときに知覚する差に対応する属性と定義される[9]．極めて多次元的な尺度である．ゴムパッドの音色のずれについては，叩く打面位置の違いやスティックと打面の接触角度の違いなどに起因すると考えられる．基礎練習では，これらの微妙な音色のずれも聞き分け，均一な音を奏でることが求められる．

これまでに VAE(Variational Auto Encoder)[10]の潜在空間を用いた音色可視化方法が提案されている[11]． ここでは潜在空間を２次元とし，２次元空間上の点の相互間距離を音色類似度と結び付けている．しかし，VAE を用いた手法では，潜在空間は人の知覚の指標を用いずニューラルネットワークが自己教師あり学習で獲得したものであるので，出力される音色可視化は人の知覚が反映されたものではない．音色は人の知覚で定義される非常に多次元的な尺度で，２次元空間上の点の相互間距離も本来，人の知覚に基づいているべきである．本研究では，Contrastive Loss[12]を用いた音色可視化手法を提案する．これにより上級者による音色の類似度判断を教師データとした音色可視化モデルの構築を試みる．

2. 関連研究

2.1 音響心理学における音色可視化

音響心理学分野では，心理物理実験をもとに低次元空間のマップを作成する手法で音色可視化が行われている．SD (Sematic Differential)法[13]での音色可視化[14-15]や MDS (Multi-Dimensional Scaling：多次元尺度法)[16-17]による音色可視化[18-21]が行われている．

SD 法では被験者は「美しい」「豊かな」「迫力のある」などあらかじめ決められた音色表現語に関して定められた段階の尺度で回答する．これらの結果から表現語間の相関係数行列を計算し，因子分析を行い，直行因子を抽出する．これに対し，MDS では，被験者は評価するすべての音のペアの類似度を回答し，これに基づいた相互間距離になるように低次元空間にマップする．その後，各軸の意味を考察する．SD 法では最初に音色表現語を定めてしまうため，これに含まれなかった要素については調査できないという性質があるが，MDS では類似度で判断するため，音色表現語を決めずに調査できる．しかし，これらの手法は，音を録音した後に心理物理実験を要するためシステムによる自動評価には組み込めない．

2.2 Contrastive Loss

Contrastive Loss は距離学習に用いられる損失関数の１つである．Hadsell らによって提案された[12].

Contrastive Loss が解くべき問題は，入力空間におけるサンプル間の近傍関係が与えられたとき，高次元の入力パターンを低次元の出力に写像する関数を見つけることである．より正確には，入力ベクトルの集合$I = \{\overrightarrow{X_1}, \dots, \overrightarrow{X_P}\}, \overrightarrow{X_i} \in \Re^D, \forall i = 1, \dots, n$とすると，$W$をパラメータとする関数$G_W = \Re^D \to \Re^d, d \ll D$を求めることである．Contrastive Loss を最小化することでG_Wを得ることができる．G_Wは以下の３つの性質をもつ．

1. 出力空間の距離尺度(ユークリッド距離など)は，入力空間における近傍関係を近似的に表現する必要がある．
2. 出力空間でのマッピングは，入力空間での距離尺度の算出を必要とせず，複雑な変換に対する不変的な規則を学習できるべきである．
3. 近傍関係が未知のサンプルに対しても有効であるべきである．

Contrastive Loss はサンプルのペアごとに算出される．$\overrightarrow{X_1}, \overrightarrow{X_2} \in I$を入力ベクトルのペアとすると，このペアに割り当てられるラベルYは２値であり，以下のように定義される．

$$Y = \begin{cases} 0 & if\ \overrightarrow{X_1}\ is\ similar\ to\ \overrightarrow{X_2} \\ 1 & otherwise. \end{cases} \quad (1)$$

$\overrightarrow{X_1}, \overrightarrow{X_2}$間の距離関数$D_W$は$G_W$の出力間のユークリッド距離として以下のように定義される．

$$D_W\left(\overrightarrow{X_1}, \overrightarrow{X_2}\right) = \left\| G_W\left(\overrightarrow{X_1}\right) - G_W\left(\overrightarrow{X_2}\right) \right\|_2 \quad (2)$$

表記を短くするために以下$D_W\left(\overrightarrow{X_1}, \overrightarrow{X_2}\right)$を$D_W$とかく．Contrastive Loss は以下の通りである．

$$\mathcal{L}(W) = \sum_{i=1}^{P} L\left(W, \left(Y, \overrightarrow{X_1}, \overrightarrow{X_2}\right)^i\right) \quad (3)$$

$$L\left(W, \left(Y, \overrightarrow{X_1}, \overrightarrow{X_2}\right)^i\right) = (1-Y)\frac{1}{2}\left(D_W^i\right)^2 + (Y)\frac{1}{2}\left\{max\left(0, m - D_W^i\right)\right\} \quad (4)$$

$\left(Y, \overrightarrow{X_1}, \overrightarrow{X_2}\right)^i$は$i$番目のラベル付けされたペア，$P$は学習ペア数である．$m$はハイパーパラメータで，$m > 0$である．類似していない入力ペアの場合，$D_W$が$m$より小さいときに損失関数に影響を与える．$W$に関して$L$を最小化すると，類似した入力ペアの$D_W$が小さく，類似していない入力ペアの$D_W$が大きくなるように Contrastive Loss は設計されている．

3. 提案手法

本研究では音色類似度判断を教師データとする Contrastive Loss を用いた音色可視化モデルを提案する．図 2 に提案する音色可視化学習モデルを示す．音色類似度判断は音色の定義に従い，ラウドネスを揃えた２音のパッド音を聴く心理物理実験を行い，”Similar”か”Dissimilar”の２値で判断させたものである．モデルの入力はパッド音をメルスペクトログラムに変換したものを用いる． ２つの入力は重みの共有されたサブネットワークでエンコードされ，特徴量ベクトルのユークリッド距離と２つの入力に対応する音色類似度ラベルにより Contrastive Loss が算出される．可視

Vol.2022-MUS-134 No.8
Vol.2022-SLP-142 No.8
2022/6/17

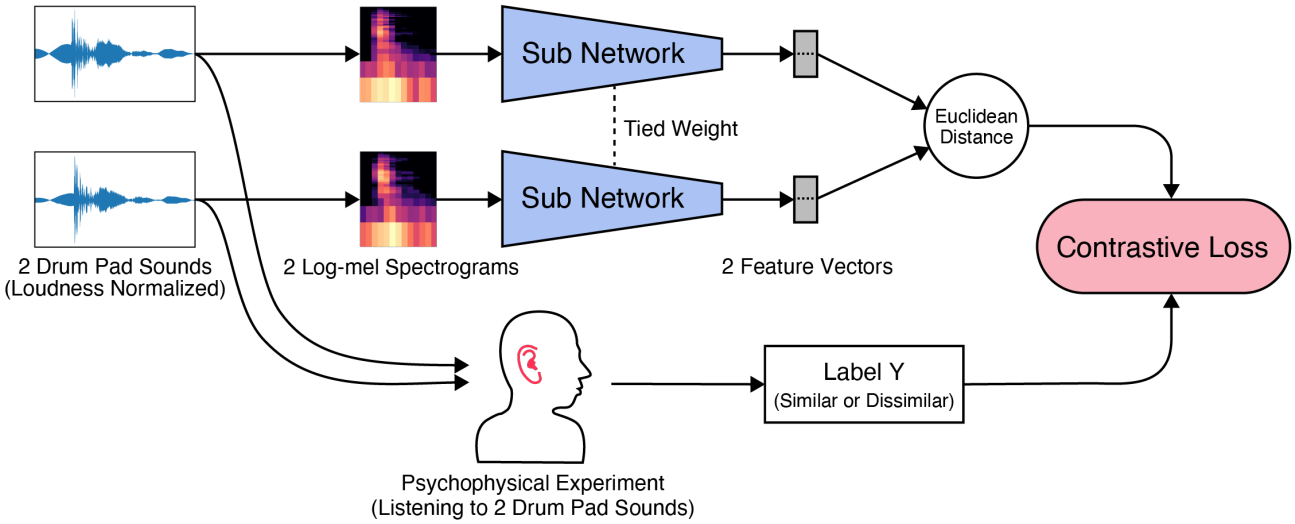

図 2　提案する音色可視化学習モデル

化時にはサブネットワーク 1 つのみを用い，出力される 2 次元特徴量ベクトルを 2 次元空間の座標と捉えることで 1 音ごとの音色可視化を行うことができる．

4. 評価実験

ゴムパッド音の録音，データ前処理，心理物理実験，モデル学習方法，評価方法，評価値Eの結果，音色可視化，考察の順に説明する．

4.1 ゴムパッド音の録音

サンプリングレート 44.1kHz でパッド音を録音した．パッドは YAMAHA トレーニングパッド 8 インチ TS01S，スティックは VICFIRTH 5A，マイクは audio-technica AT2020 を使用した．録音の様子を図 3 に示す．演奏者は，演奏前に BPM 100 をメトロノームで聴き，BPM 100 の 4 分音符のペースで均一な音が鳴るように 1 分間ゴムパッドを叩くタスクを指示された．演奏者は 19 歳から 23 歳の男性で，打楽器初心者 3 名，上級者 2 名である．それぞれ初級者 A，B，C，上級者 A，B とする．上級者 A の打楽器演奏歴は 8 年，上級者 B の打楽器演奏歴は 12 年であった．学習には初級者 A のデータを用い，音色可視化には，学習に用いていない初級者 B，C と上級者 A，B のデータを用いた．初級者 A はタスクを 3 回行い，初級者 B，C，上級者 A，B はタスクを 1 回行った．

4.2 データ前処理

パッド音は単音に切り出された後，単音内の最大値のフレームの前 1000 フレーム，後 3000 フレームの固定長で切り出された．また，ラウドネスを揃えるため，-25.0dBFS で正規化された． この処理の結果，初級者 A の音を 269 個，初級者 B の音を 86 個，初級者 C の音を 94 個，上級者 A の音を 99 個，上級者 B の音を 110 個得た．

心理物理実験には，ここで得た音を使用し，ニューラルネットワークの入力には，窓幅 1024，スライド幅 320，メルフィルタバンク数 64 で変換されたメルスペクトログラムを使用した．

4.3 心理物理実験

被験者は，ゴムパッド音の録音に参加していない上級者であり，打楽器演奏歴 14 年の著者である．本研究では，個

図 3　録音の様子

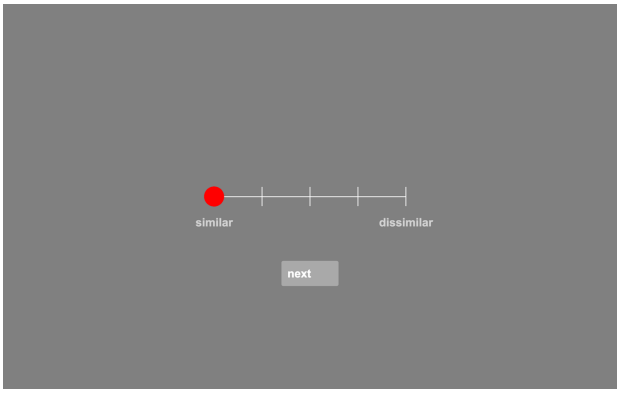

図 4　心理物理実験の回答画面

Vol.2022-MUS-134 No.8
Vol.2022-SLP-142 No.8
2022/6/17

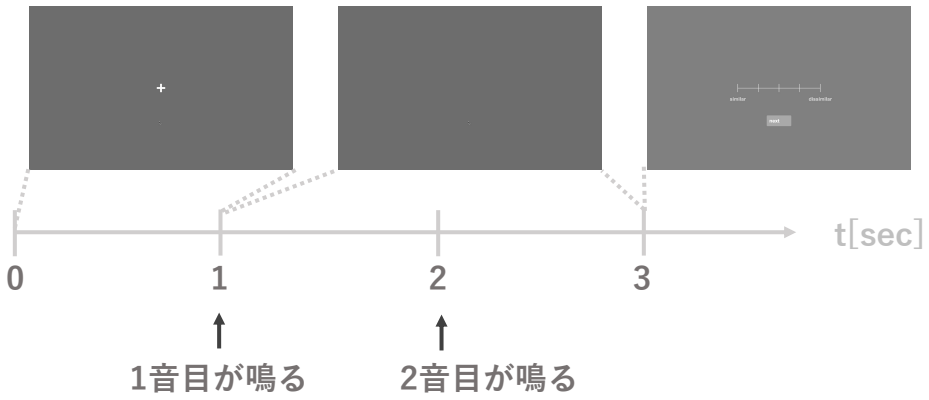

図 5　心理物理実験の 1 試行の流れ

人差によるデータのばらつきを排除して提案手法の有効性を確認したいため被験者は 1 名とした．

被験者はペアのゴムパッド音を聴き，”Similar”か”Dissimilar”の 2 値で判断する．18 音ごとに音のグループを作る．各グループにおいてすべての組み合わせをとり音のペアを作ると，1 グループあたり 18C2 で 153 種類となる．2 音の提示される順序は両方向で提示されるため被験者が聴く音のペアの総数は 306 種類となる．実験は練習試行，本試行の順に行われる．練習試行では 30 試行，本試行では 306 試行ずつ行い，本試行では 103 試行毎に休憩を挟んだ．本試行のデータのみ使用する．練習試行は 306 種類の音のペアよりランダムに選択された 30 試行がランダムな順序で提示され，本試行では，306 種類の音のペアが 1 回ずつランダムな順序で提示された．

実験は心理物理実験作成ツールの Psychopy[22]で作成された．回答画面を図 4 に，1 試行ごとの流れを図 5 に示す．まず 0s に画面の真ん中に十字が表示され，1 s に十字が消え，1 音目のパッド音が提示される． 2s に 2 音目のパッド音が提示され，その後 3s にスライダーが表示され，被験者は回答する．本稿では扱わないが 5 段階での実験も実施したためスライダーは 5 段階となっている．被験者は，スライダーの両端のみを回答するように指示され，音色類似度を 2 値で判断した．回答画面では選択した段階のメモリに赤い丸が表示されることで被験者が自身の選択を確認することができる．スライダーの下の Next ボタンを押して次の試行に移る．音色類似度の心理物理実験は MDS のための心理物理実験[18-21]を参考に設計された．

音のグループは 10 個作成された．計 180 個のゴムパッド音が必要となるが，録音した 269 個のゴムパッド音より用意した．被験者は 10 個のグループについて音色類似度判断をした．2 個目のグループ以降は練習試行をスキップした．得られた 1530 ペアに対するラベルのうち，提示順序逆で回答が一致した 1003 ペアを用いることにした．

4.4 モデル学習方法

10 個の音のグループはグループごとに異なる音で構成されている．音のグループ 2 個で 1 まとまりとし，5 つのまとまりに分ける．この分け方で 5 fold の Nested Cross Validation で 10 epoch 学習した．この分け方だと単音についてオープンな条件となる．Contrastive Loss のハイパーパラメータ mは 1 とした．サブネットワークには 2 種類の 4 層 CNN を用いた． 1 つにはプーリング層を用い，も

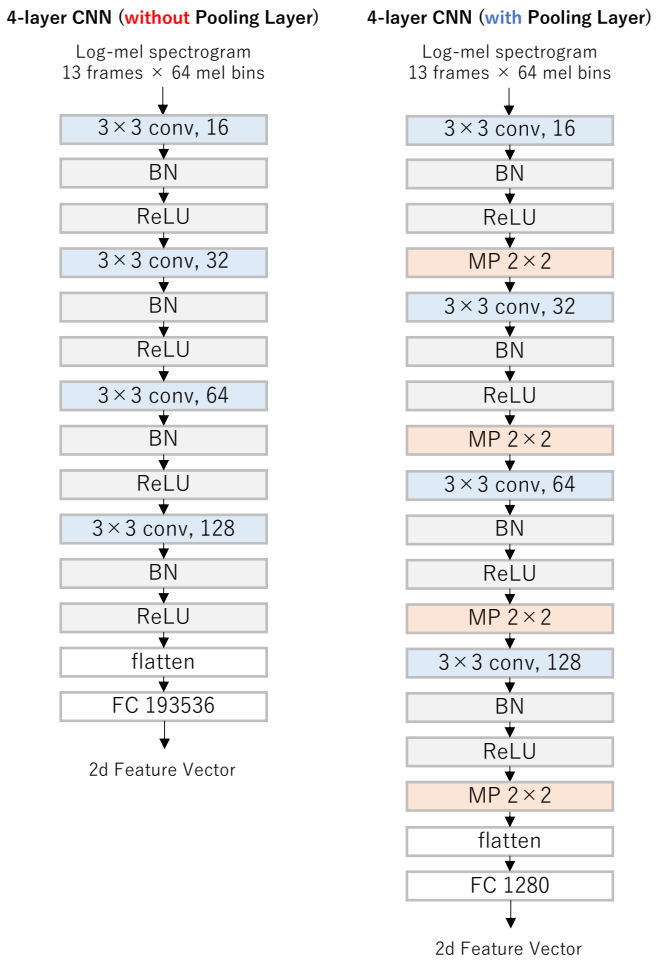

図 6　サブネットワークの構造
(左はプーリング層なしの 4 層 CNN，右はプーリング層ありの 4 層 CNN)

Vol.2022-MUS-134 No.8
Vol.2022-SLP-142 No.8
2022/6/17

う 1 つにはプーリング層を用いなかった．使用したサブネットワークの構造を図 6 に示す．

4.5 評価方法

評価値による客観的評価と可視化を行う．前者では提案手法による音色可視化がどれほど人の知覚を反映しているかを評価値によって客観的に評価し，後者では実際に可視化を行い，考察する．

評価値Eは以下のように定義され，$E > 1$の範囲で大きいほど 2 値を顕著に可視化できているとする．

$$E = \frac{MD_{dis}}{MD_{sim}} \tag{5}$$

$$MD_{sim} = \frac{\sum_{i=1}^{N_{sim}} D_{sim}^{i}}{N_{sim}} \tag{6}$$

$$MD_{dis} = \frac{\sum_{j=1}^{N_{dis}} D_{dis}^{j}}{N_{dis}} \tag{7}$$

D_{sim}^{i}はi番目の”Similar”とラベル付けされたペアのユークリッド距離，N_{sim}は”Similar”とラベル付けされたペアの数を示す．また，D_{sim}^{j}はj番目の”Dissimilar”とラベル付けされたペアのユークリッド距離，N_{dis}は”Dissimilar”とラベル付けされたペアの数を示す．MD_{sim}とMD_{dis}はそれぞれD_{sim}^{i}とD_{dis}^{i}の平均値であり，EはMD_{sim}を分母，MD_{dis}を分子とした値として算出される．

“Similar”とラベル付けされたペアの相互間距離の平均値と”Dissimilar”とラベル付けされたペアの相互間距離の平均値が等しいときにEは 1 となり，“Similar”とラベル付けされたペアの相互間距離の平均値が短く，”Dissimilar”とラベル付けされたペアの相互間距離の平均値が長くなるほど，Eが大きくなることがわかる．以上より，$E > 1$の範囲で大きいほど提案手法の音色可視化は人の知覚が反映されていることが示唆される．

4.6 評価値 E の結果

音色可視化モデルのサブネットワークがプーリング層なしの 4 層 CNN の場合とプーリング層ありの 4 層 CNN の場合について評価値Eの実験結果を表 1 に示す．

表 1　実験結果

	プーリング層なし	プーリング層あり
E	1.82	1.78

4.7 音色可視化

初級者 B の音 86 個，初級者 C の音 94 個，上級者 A の音 99 個，上級者 B の音 110 個を用いて音色可視化を行った．評価値Eの優れていたプーリング層なしの 4 層 CNN を用いた．音色可視化例を図 7 に示す．Advanced Player は上級者を，Beginner は初級者を指す．

4.8 考察

上級者が音色の均一な音叩くことができているとすると，可視化結果では，上級者は局所的に点が集中し，初級者は比較的点が散らばることが期待される．可視化例を見ると上級者 A は点が集中している傾向が見えるが，上級者 B の

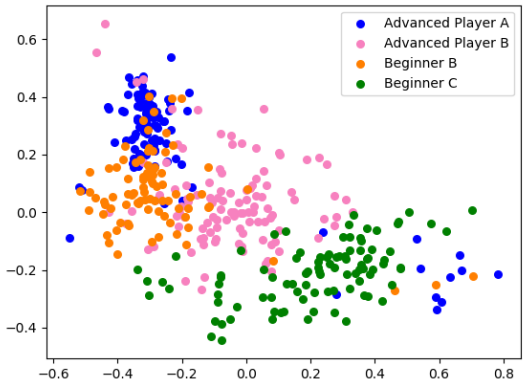

図 7　音色可視化例

点の散らばりは初級者の点の散らばりと同程度のように見える．点の散らばり具合が奏者個人内の音色のずれを表現できているかの議論にはさらに多くの上級者のデータが必要である．一方，奏者ごとの点の集中は観察しやすい．1 人のゴムパッド音の音色類似度を学習したニューラルネットワークが奏者間の音色の違いを可視化できていることになる．この可視化例より，奏者間の音色の違いが奏者個人内での音色のばらつきに比べて大きいのではないかと考える．評価値$E > 1$より提案手法の音色可視化は，人の知覚をある程度は反映した可視化となっていることが示唆されているが，奏者個人内の比較的小さな音色の違いを検出するに足る精度は得られていない．

評価値Eの結果よりプーリング層なしの条件で精度が高いことが示唆された．プーリング層の目的はそもそも位置ずれに対する頑健性向上だが，今回，入力音を切り出す際に，最大値フレームの前後固定長切り出しを行ったことにより時間方向の位置ずれが小さくなったために，あまり機能していないように考えられる．入力音が 0.1s 未満であり，入力特徴ベクトル長が小さいためプーリング層が入る CNN を深く重ねることは難しかった．しかし，プーリング層なしで精度が高いことが実験結果より確認されたため，プーリング層なしのさらに深いニューラルネットワークでの学習結果にも期待できる．

5. おわりに

ゴムパッドでの練習において音色可視化を利用するには奏者個人内の音色のずれについて評価できる必要がある．評価値Eより提案手法がある程度人の知覚を反映した可視化を行えることが示唆されたが，音色可視化例よりその制度が奏者個人内の音色のずれを表現するにあたり十分かは不明で，議論にはさらに多くの上級者のパッド音のデータが必要である．

また，提案手法の課題として，人手による音色類似度判断は手間がかかり，心理物理実験に参加可能な対象者が少ないことから大規模な教師データの作成が難しいことが挙

げられる．一方，VAE と Contrastive Loss を組み合わせた音色可視化手法もあり[23]，自己教師あり学習を適切に使用してデータセットの大きさの課題を解決したい．

謝辞　実験に協力してくださった方々や議論してくださった方々に謹んで感謝いたします．

参考文献

[1] Simon Holland, Anders J. Bouwer, Mathew Dalgleish and Topi M. Hurtig: Feling the Beat Where it Counts: Fostering Multi-limb Rhythm Skills with the Haptic Drum Kit. *Proceedings of the 4th International Conference on Tangible, Embedded, and Embodied Interaction,* pp. 25-27(2010).

[2] Hiroyuki Kanke, Tsutomu Terada and Masahiko Tsukamoto: A Percussion Learnining System Using Rhythm Internalization with Haptic Indications. *Proceedings of 12th International Conference on Advances in Computer Entertainment Technology,* No.14, pp.1-5(2015).

[3] Ayaka Ebisu, Satoshi Hashizume and Kenta Suzuki, Akira ishii, Mose Sakashita, Yoichi Ochiai: Stimulated Percussions: Techniques for Controlling human as Percussive Musical Instrument by Using Electrical Muscle Stimulation. *SIGGRAPH ASIA(SA '16)* Posters, No.37, pp.1-2(2016).

[4] Ayaka Ebisu, Satoshi Hashizume, Kenta Suzuki, Akira Ishii, Mose Sakashita and Yoichi Ochiai: Stimulated percussions: m-0ethod to control human for learning music by using electrical muscle stimulation. *Proceedings of the 8th Augmented Human International Conference(AH '17)*, No.33, pp.1-5(2017).

[5] Ayaka Ebisu, Satoshi Hashizume and Yoichi Ochiai: Building a feedback loop between electrical stimulation and percussion learning, *ACM SIGGRAPH 2018 Studio(SIGGRAPH '18)*, No. 1, pp.1-2(2018).

[6] KORG INC.: BEATLAB mini RHYTHM TRAINER EFGSJ2.

[7] Yasuhiko Tsuji and Atsuhiro Nishikata: Develoopment and Evaluation of Drum Learning Support System Based on Rhythm and Drumming Form. Electronics and Communications in Japan(Part Ⅲ: Fundamental Electronic Science). Vol.89, No.9. pp.11-21(2006).

[8] 星野将吾, 深山覚, 後藤真孝: 基礎的演奏能力向上のための打楽器練習支援システム．情報処理学会研究報告, Vol.2020-MUS-126, No.10, pp.1-7(2020).

[9] APA Dictionary of Psychology, timbre, *AMERICAN PSYCHOLOGY ASSOCIATION,* https://dictionary.apa.org/timbre.

[10] Diederik P Kingma and Max Welling: Auto-Encoding Variational Bayes. *International Conference on Learning Representations(ICLR)*(2014).

[11] Naoki Kimura, Keisuke Shiro, Yota Takakura, Hiromi Nakamura and Jun Rekimoto: SonoSpace: Visual Feedback of Timbre with Unsupervised Learning. *Proceedings of the 28th ACM International Conference on Multimedia(MM '20)*, pp.367-374 (2020).

[12] Raia Hadsell, Sumit Chopra, Yann LeCun: Dimensionality Reduction by Learning an Invariant Mapping. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR '06)*, pp.1735-1742(2006).

[13] Osgood C. E.,.Suci J. G and Tannenbaum P. H.: The Measurement of Meaning, *University of Illinois Press*(1957).

[14] 北村音壱，難波精一郎，三戸左内 :再生音の心理的評価について，電気通信学会電気音響研究会資料，1-27(1962).

[15] 曽根敏夫，城戸健一，二村忠元 :音の評価に使われることばの分析．日本音響学会誌，18(6)，320-326(1962).

[16] J. B. Kruskal: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika*, Vol.29, No.1, pp.1-27(1964).

[17] J. B. Kruskal: Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika*, Vol.29,No.2,pp.115-129(1964).

[18] John M. Grey: Multidimensional perceptual scaling of musical timbres. *The Acoustical Society of America*, Vol.61, No.5, pp.1270-1277(1977).

[19] Paul Iverson and Carol L. Krumhansl: Isolating the dynamic attributes of musical timbre. *The Acoustical Society of America*, Vol.94, No.5(1993).

[20] Stephen McAdams, Suzanne Winsberg, Sophie Donnadieu, Geert De Soete and Jochen Krimphoff: Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research,* No.58, pp. 177-192(1995).

[21] Stephen Lakatos: A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, Vol.62, No.7, pp.1426-1439(2000).

[22] Open Science Tools Ltd.: Psychopy. https://www.psychopy.org/

[23] Keitaro Tanaka, Ryo Nishikimi, Yoshiaki Bando, Kazuyoshi Yoshii and Shigeo Morishima: Pitch-Timbre Disentanglement Of Musical Instrument Sounds Based On Vae-Based Metric Learning. *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),* pp.111-115(2021).