Vol.2022-MUS-134 No.46
Vol.2022-SLP-142 No.46
2022/6/18

各楽器音に着目した楽曲間類似度学習

橋爪 優果[1,a)]　李 莉[2,b)]　戸田 智基[2,c)]

概要：自由度の高い楽曲推薦・検索を実現するためには，楽曲間類似度を計算する必要があり，その基準が重要となる．楽曲信号に対して楽曲間類似度を直接計算する枠組みとして，各楽曲のタグ情報を用いて，トリプレット損失に基づく距離学習を行うデータ駆動型の手法が提案されている．しかし，その結果得られる楽曲間類似度の基準は，様々な楽器音が混ざった楽曲全体を捉える場合が多く，例えば，類似したドラム音を含む楽曲を検索するといったことが困難であり，楽曲推薦・検索システムの機能が限定される．そこで，本報告では，より自由度の高い楽曲推薦・検索システムの実現のために，楽曲中の各楽器音に着目した楽曲間類似度計算を提案する．提案手法では，タグ情報を用いずに個別の楽器音に対して距離学習を行う．さらに，実際には各楽器音源が常に得られるとは限らないことを想定し，楽器音源分離で得られる各楽器音源を用いて距離学習を行う効果も検討する．実験の結果，(1) 各楽器音に対して異なる類似度の基準を学習できること，(2) 一部の楽器音を用いて学習した類似度は，楽曲そのものを用いて学習したものよりも正確な結果を導くことが可能なこと，(3) 分離した楽器音源を用いて学習すると性能が低下すること，(4) 提案手法で学習した類似度の基準は人間の感覚に対応する結果を示すことが明らかとなった．

Music similarity learning focusing on individual instrumental sounds

Abstract: The criteria for measuring music similarity are important for developing a flexible music recommendation system. Some data-driven methods have been proposed to calculate music similarity from only music signals, such as metric learning based on a triplet loss using tag information on each musical piece. However, the resulting music similarity metric usually captures the entire piece of music, i.e., the mixing of various instrumental sound sources, limiting the capability of the music recommendation system, e.g., it is difficult to search for a musical piece containing similar drum sounds. Besides, for greater accuracy, the tag information is commonly labeled manually, but this requires a lot of human resources and time. Towards the development of a more flexible music recommendation system, we propose a music similarity calculation method that focuses on individual instrumental sound sources in a musical piece. We adopt metric learning to individual instrumental sound source signals without using any tag information. Furthermore, we also investigate the effects of using instrumental sound source separation to obtain each source in the proposed method since each instrumental sound source is not always available in practice. Experimental results have shown that (1) different similarity metrics can be learned for individual instrumental sound sources, (2) similarity metrics learned using some instrumental sound sources can lead to more accurate results than that learned using the entire piece of music, (3) the performance degrades when training a network with the separated instrumental sounds, and (4) similarity metrics learned by the proposed method well produce results that correspond to perception by human senses.

1. はじめに

インターネット上で視聴可能な楽曲の量は膨大かつ増え続けている．このような状況下で世界中の楽曲をすべて聴いて好みの楽曲を見つけることは不可能である．そのため，ユーザが好みの楽曲を効率的に見つけるために楽曲推薦システムなどの音楽情報検索（Music Information Retrieval: MIR）技術が必要であり，楽曲間類似度を計算する適切な基準の設計が不可欠である．

楽曲間類似度を計算する一つのアプローチとして，ユーザの聴取履歴を活用する方法があり，代表的な手法として協調フィルタリングがある [1]．この手法では，あるコンテンツに対し似たような評価や同じ行動をしたユーザは，他のコンテンツに対しても似たような評価をすると仮定する．これにより，同じような行動をとった他のユーザの評価から，未知の楽曲の評価を予測することができる．しかし，新しくリリースされた楽曲は，ある程度の聴取履歴が

[1] 名古屋大学 情報学研究科
[2] 名古屋大学 情報基板センター
[a)] hashizume.yuuka@g.sp.m.is.nagoya-u.ac.jp
[b)] li.li@g.sp.m.is.nagoya-u.ac.jp
[c)] tomoki@icts.nagoya-u.ac.jp

記録されるまではほとんど推薦されない可能性があるという限界がある．また，一般に人気のある楽曲はより多くの評価を得られるため，知名度の低い楽曲に出会う機会が少ないという問題もある．

協調フィルタリングの問題を回避する従来法として，コンテンツ自身の特徴に基づく類似度を用いて推薦するコンテンツベースのアプローチがある．コンテンツベースの類似度は，一般的に音楽信号から特徴表現を抽出し，その表現間の類似度または距離を計算することで得られる．例えば，コード進行やテンポ，人手で設計された音響特徴などが特徴表現として用いられ [2]，コサイン距離やユークリッド距離などの距離基準を用いてそれらの間の類似度を測定する．コンテンツベースの手法はユーザの聴取履歴を必要としないという利点がある一方で，その性能は用いる特徴量と距離基準に大きく依存するため，人手で設計されたものを用いる際には必ずしもうまく機能するとは限らない．

前述のようなコンテンツベースの従来法に対して，深層学習の登場によりデータ駆動型の特徴抽出の研究が発展し，MIR システムの性能向上に有効であることが示されている [3], [4]．例えば，ジャンル分類器の中間層から特徴表現を抽出することが提案されている [4], [5]．また，人間が付与したタグ [6]，アーティストタグ [7]，ジャンルタグ [8]，ゼロショット学習によるタグ [9] などのタグ情報も活用した距離学習により特徴表現を学習する方法が提案されている．これらの従来手法では，通常，様々な楽器音が混ざった楽曲全体を考慮して楽曲間類似度を測定される．これに対し，より柔軟な楽曲間類似度計算の枠組みとして，ユーザが聴きたい楽曲の視点はユーザによって異なる可能性がある点に着目すると，楽曲間類似度を一つの視点のみで計算するだけでは不十分である．例えば，ドラムに着目した際に類似している楽曲を探すといった，より使いやすく柔軟な楽曲推薦・検索システムを実現するためには，より多様な特徴表現を用いて楽曲間類似度を算出することが必要である．

本報告では，より柔軟な楽曲間類似度計算を実現するために，楽曲中の各楽器音に着目した楽曲間類似度計算手法を提案する．提案手法では，データ駆動型特徴抽出技術の可能性を最大限に活かし，タグ情報を用いず，トリプレット損失を用いた距離学習により特徴表現を抽出する [10]．トリプレット損失におけるポジティブ，ネガティブサンプルは，同じ楽曲から抽出されたセグメントかどうかで定義する．また，各楽器音に着目した楽曲間類似度の計算を可能にするために，楽曲そのものではなく，単一の楽器音源を用いて類似度を計算する．さらに，実際には個別の楽器音源を得ることは困難であると想定し，楽器音源分離を適用して，ミックスされた元の楽曲からそれぞれの楽器音源を取得し，分離された音源信号を用いた際の評価も行う．客観評価実験と主観評価実験により，(1) 楽器ごとに異なる有用な類似度基準を学習できるか，(2) 分離された音源信号を提案手法の類似度計算に利用できるか，(3) 学習した類似度基準は人間の感覚に対応した結果を導くか，について検討する．

2. 関連研究

2.1 楽曲間類似度計算のための特徴表現抽出

Li ら [11] は, MFCC（mel frequency cepstral coefficients）をネットワークの入力とし，予測したジャンルを出力する CNN（convolutional neural network）を用いた特徴表現の抽出方法を提案している．最終層の入力として用いられる潜在特徴を，MFCC から抽出される特徴表現として利用する．また，MFCC の代わりにメルスペクトログラムを入力として用いる枠組みもよく研究されている． 例えば，Fathollahi と Razzazi[5] は，楽曲を 50%オーバラップを含む 3 秒のセグメントに分割し，それぞれを対応するメルスペクトログラムに変換し，これを CNN の入力として使用している．楽曲の長さを 3 秒，5 秒，10 秒と変化させ，オーバーラップがあるものとないものを用いた場合の性能を比較した結果，オーバーラップを含む 3 秒間のセグメントを使用した場合に最も良い結果が得られることが報告されている．

2.2 トリプレット損失を用いた楽曲間類似度基準学習

深層距離学習 [12] では，機械学習により特定の課題に対する距離基準を自動的に構築することを目的としており，一般に手作業で設計するよりも課題に適した距離基準を見つけることができる．トリプレット損失 [10] では，1 つのサンプルをアンカーとし，残りの 2 つのサンプルをポジティブサンプルとネガティブサンプルとするトリプレットを用いて距離基準を学習する．ここで，ポジティブサンプルはネガティブサンプルよりもアンカーに類似している必要がある．

Lee ら [8] はトリプレット損失を用いた距離学習により楽曲間類似度を計算するトラック正規化手法を提案している．この手法では，トラック情報のみを用いて抽出されたトリプレットによって，楽曲そのもののトラックベースの類似度を学習する．すなわち，アンカーと同じ楽曲から抽出したセグメントをポジティブサンプル，アンカーと異なる楽曲から抽出したセグメントをネガティブサンプルと定義する．

i 番目のアンカー，ポジティブサンプル，ネガティブサンプルをそれぞれ $x_i^{(a)}, x_i^{(p)}, x_i^{(n)}$ とすると，トリプレット t_i は $\{x_i^{(a)}, x_i^{(p)}, x_i^{(n)}\}$ の集合として構成される．ここで，$i = 1, \ldots, I$ は学習サンプルのインデックスを示す．トリプレット損失は次のように定義される．

$$\mathcal{L}(t_i) = \max\{d(x_i^{(a)}, x_i^{(p)}) - d(x_i^{(a)}, x_i^{(n)}) + \Delta, 0\}, \quad (1)$$

ここで，d はユークリッド距離やコサイン距離など 2 サンプル間の距離を測定するための距離関数，Δ はポジティブ，ネガティブサンプル間の最小距離を定義するマージンである．

3. 提案手法

楽曲そのものに着目した従来法に比べ，より多角的に楽曲間類似度を扱うことができ，柔軟性の高い楽曲推薦・検索システムを実現するために，楽曲の一部分である各楽器音に着目した楽曲間類似度算出手法を提案する．提案手法の概要を図 1 に示す．

提案手法では，ドラム，ピアノなどの個別の楽器音に対して，トリプレット損失を用いた距離学習を適用する．ジャンルやアーティスト，ムードなどのタグと異なり，楽器音に着目して楽曲の類似度を表すタグをアノテーションするには人的コストを要する．そこで，2.2 節で述べたトラック情報を用いる手法を用い，ポジティブサンプルとネガティブサンプルを定義する．つまり，アンカーと同じ楽曲から抽出したセグメントをポジティブサンプル，異なる楽曲からのセグメントをネガティブサンプルと定義する．

距離学習における楽曲間類似度計算のための特徴表現を抽出するために，畳み込み層と完全連結層で構成されるネットワークを使用する．このネットワークアーキテクチャを図 2 に示す．また，各楽器音に対して異なるネットワークを個別に学習させる．アンカーとポジティブサンプル間の距離 $d(x_i^{(a)}, x_i^{(p)})$，アンカーとネガティブサンプル間の距離 $d(x_i^{(a)}, x_i^{(n)})$ はコサイン距離により測定する．また，実際には楽器ごとに録音した音源が必ずしも利用できるわけではないことを想定し，楽器音源分離法を用いてミックスされた元の音楽信号から抽出した分離楽器信号に対しても提案手法を適用する．

4. 実験的評価

4.1 実験条件

使用するデータセットは slakh[13] で，様々な楽器音をミックスした音源（以下，混合音）だけでなく，各楽器の音源（以下，オリジナル楽器音）も含まれている．なお，このデータセットに含まれる楽曲にはボーカルは含まれていない．提案手法では，ドラム，ベース，ピアノ，ギターのオリジナル楽器音を用いて，各楽器音に着目した楽曲間類似度計算を行う．また，音楽信号をドラム，ベース，ピアノ，ボーカル，その他に分離できる Python ライブラリの楽器音分離手法“spleeter”[14] を用いて混合音からドラム，ベース，ピアノの音源を抽出し，それら分離音（以下，分離楽器音）を用いた場合の楽曲間類似度計算への影響を検討する．また，参考として混合音を用いた楽曲間類似度計算も行う．

データセットから 180 曲を使用し、各楽曲を 50 %オーバーラップを含む 3 秒のセグメントに分割し，無音区間を除いた最初の 40 セグメントからなる合計 7200 個のセグメントを学習用セグメントとして使用する．テストでは，20 曲を使用し，同様にセグメントに分割し，各楽曲の無音部分を除いたすべてのセグメントを使用する．これらのセグメントをメルスペクトログラムに変換し入力として使用する．メルスペクトログラムに変換する際の分析窓長は 2048，フレームシフトは 512 である．学習時には，メルスペクトログラムから任意のアンカーを選択し，トリプレット損失に基づく距離学習を行う．図 2 の CNN を用いて，楽曲間類似度を測るための特徴表現として，メルスペクトログラムから 128 次元の埋め込みベクトルを抽出する．損失関数のマージンは 0.2，バッチサイズは 64，エポック数は 150 とする．CNN の初期設定を変えて 5 回学習し，それぞれで後述する評価を行い，その結果を平均する．

4.2 評価方法

4.2.1 客観評価

一般に，推薦システムに適した特徴表現は，(1) 類似項目（同じ楽曲からのセグメント）の特徴表現は互いに近く，(2) 非類似項目（異なる楽曲からのセグメント）の特徴表現は非類似度に応じて互いに離れている，といった 2 つの性質を満たす必要がある．そこで，学習した特徴表現を評価するために，特徴表現の類似度を用いて推論した楽曲 ID の正解率を用いる．具体的には，K-nearest neighbor (kNN) 法を用いて，テストセグメントの楽曲 ID を推論する．推論するセグメント以外のテストセグメントの楽曲 ID は既知であると仮定し，全てのテストセグメントを学習した特徴表現空間に埋め込み，上位 5 つの最近傍テストセグメントの ID を用いた多数決により，各テストセグメントの楽曲 ID を予測する．混合音，オリジナル楽器音，分離楽器音を用いて学習した各特徴表現空間について，テストデータセット全体を用いた 5 回の試行の結果を平均した正解率を算出する．

提案手法の目的は，各楽器音に着目した場合に，異なる類似度基準を構築することである．そこで，20 曲のテスト楽曲の重心特徴表現に対する距離行列を用いた評価を行う．重心特徴表現は，同じ楽曲内の全セグメントに対する特徴表現の平均化により求める．平均距離行列を 5 回の試行の平均により求め，混合音と各オリジナル楽器音を用いて学習した楽曲間類似度基準間で比較し，どの程度異なるかを調査する．距離行列の可視化に加え，相関係数と Spearman の順位相関係数 [15] を用いて楽曲間類似度基準の違いを定量化する．相関係数は，各オリジナル楽器音と混合音のそれぞれの平均距離行列の対角部を除く上三角部の要素をベクトル化し，各組の 2 つのベクトル間の相関係数を算出する．Spearman の順位相関係数については，各テスト楽曲ごとに，自身以外の楽曲の重心との距離値，すなわち平均距離行列の各列を用いて類似する楽曲の順位をつけ，各組の 2 つの順位の間の Spearman の順位相関係数を算出する．この係数をテスト楽曲 20 曲について計算し，その平均を求める．

また，各楽器音を用いて学習した類似度基準において似ているとされた楽曲が，混合音を用いて学習した類似度基準において類似しているかを定量化するために，MRR（Mean Reciprocal Rank）[16] を用いる．前述の各テスト

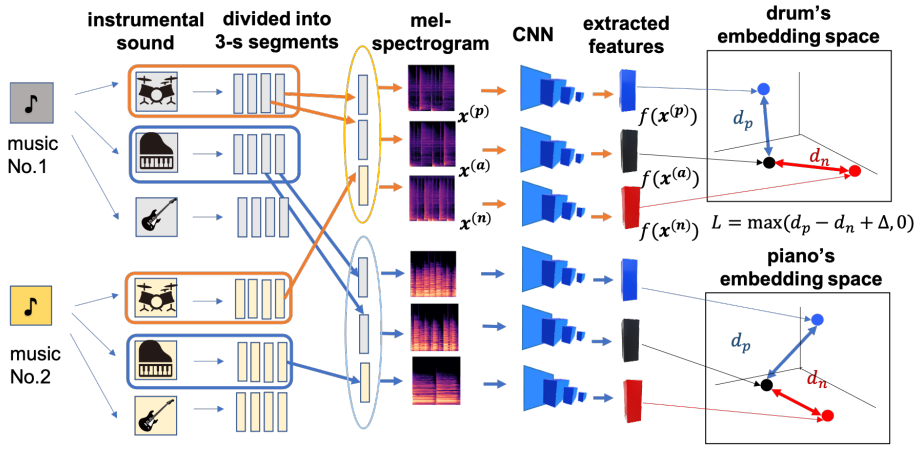

図 1: 提案手法の概要 トリプレット損失を用いた距離学習により各楽器音の特徴表現を個別に抽出する．$x^{(a)}$, $x^{(p)}$, $x^{(n)}$ はそれぞれアンカー，ポジティブ，ネガティブサンプルである．

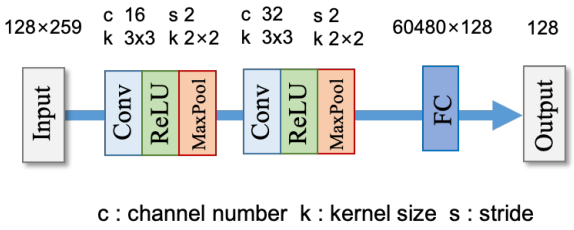

図 2: CNN のネットワーク構成．‘c’, ‘k’, ‘s’ はそれぞれチャンネル数，カーネルサイズ，ストライドを表す．“Conv”は畳み込み層，“FC”は完全連結層をそれぞれ表す．入力と出力の上の数字はそれぞれのデータサイズで，完全連結層の上の数字は層の入力と出力のサイズである．

楽曲ごとの類似曲の順位を用い，各楽器音の類似度基準においての類似曲上位 3 曲に関して，それぞれその楽曲の混合音の類似度基準における順位の値を逆数にしたものが MRR スコアである．このスコアをテスト楽曲 20 曲について計算し，その平均を求める．

4.2.2 主観評価

各楽器音や楽曲そのものに着目して学習した類似度基準が，他種類の音源で学習した類似度基準よりも，着目した観点において知覚的に類似したセグメントを発見できるかどうかを評価するために，主観評価実験を実施する．実験に参加した 13 名の被験者に，アンカーと 2 つの候補を含む 3 つの音サンプルを聞いてもらい，どちらがアンカーに似ているか，また回答に自信があるかどうかを選択してもらう．なお，各楽器音に着目して聞いてもらう際には，被験者が各楽器音に着目しやすいように，音サンプルは混合音ではなく，オリジナル楽器音を使用する．各被験者に 40 セットの音を提供し，以下の要領で実験を実施する．

楽曲からランダムに切り取った音サンプルを実験に用いる．音サンプルの長さが短い場合，被験者が類似度を判断するのが困難となるため，各音サンプルの長さは 10 秒間とする．ドラム，ベース，ピアノ，ギターのオリジナル楽器音と，混合音で学習した 5 種類の楽曲間類似度基準を評価するために，それぞれのケースにおいて，次の手順により計 8 つの音サンプルセットを作成する．以下，具体例を挙げて手順を説明する．

ドラムに注目した楽曲間類似度基準を評価する場合，(1) 各音サンプルセットについて，まずアンカーとなるセグメントをランダムに選択する．(2) ドラム音で学習した特徴表現空間において，アンカーに 1 番目または 2 番目に近いセグメントから，ポジティブサンプルをランダムにどちらか選択する．(3) 混合音を用いて学習した別の特徴表現空間において，アンカーに 1 番目と 2 番目に近いセグメントから，ネガティブサンプルをランダムにどちらか選択する．ただし，ポジティブサンプルの候補 2 つのうちいずれかが，ネガティブサンプルの候補に含まれる場合は，音サンプルセットを無効とする．(4) 選択されたドラム音のセットを被験者に提示する．

混合音に着目した楽曲間類似度基準を評価する場合，混合音で学習した特徴表現空間において，ポジティブサンプルを選択し，オリジナル楽器音 4 種からランダムに選んだ楽器音を用いて学習した特徴表現空間からネガティブサンプルを選択する．提示するのは混合音のセットである．

以上の手順より，各音サンプルセットにおいて，ポジティブサンプルに対応する候補が選択されれば，学習された類似度基準による選択と聴感上の類似度による選択が一致することを示す．

4.3 結果

4.3.1 客観評価

予測された楽曲 ID の正解率を表 1 に示す．混合音とオリジナル楽器音の場合，約 90%〜95%の高い精度が得られていることが分かる．ドラムやピアノのオリジナル楽器音は混合音よりも精度が高く，一部の楽器音を用いて学習した類似度基準は，混合音を用いて学習した類似度基準よりも精度が高いことが示唆される．一方，音源分離を用いた

表 1: 各特徴表現を用いた kNN による正解率．“オリジナル”と “分離”の列は，それぞれオリジナル楽器音と分離楽器音を示す．また，“混合” は混合音を示す．また，“分散”の列は，5 回の試行における分散を示す．

オリジナル	正解率 [%]	分散	分離	正解率 [%]	分散
混合	93.24	2.8e-4			
ドラム	95.13	2.3e-5	ドラム	79.35	1.1e-3
ベース	87.07	4.7e-4	ベース	41.59	1.4e-4
ピアノ	94.98	1.6e-5	ピアノ	70.12	6.2e-5
ギター	91.15	6.6e-5			

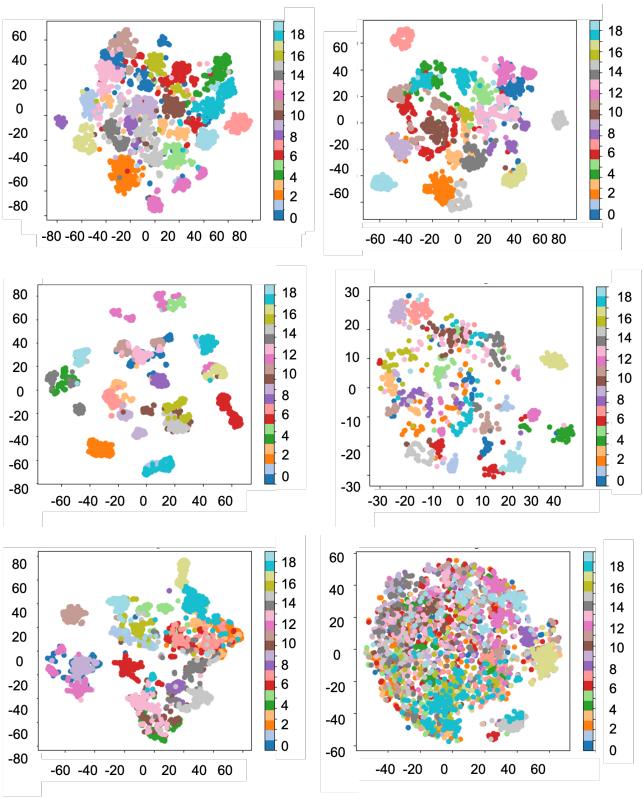

図 3: それぞれ，混合音（上段左），オリジナルギター音（上段右），オリジナルドラム音（中段左），分離ドラム音（中段右），オリジナルベース音（下段左），分離ベース音（下段右）を用いて学習し可視化した特徴表現の一例．カラーバーの右側の数字は，20 曲のテスト楽曲の楽曲 ID を示す．両図とも，同じ楽曲からのセグメントは同じ色でプロットされている．

場合には，精度が 80%未満まで低下する．この原因として，他の楽器のアーチファクトや残留成分によって，分離された音の音質が元の音よりも低下していることが考えられる．なお，全てのケースで有意に低い分散値が達成されたことは，トリプレット損失を用いて異なる初期ネットワークパラメータで類似度基準を安定的に学習できることを示している．また，t-SNE[17] を用いて，128 次元の特徴表現を 2 次元表現に圧縮し可視化した特徴表現空間の一例を図 3 に示す．オリジナル楽器音を用いた場合は，同じ楽曲の特徴表現が集中し，クラスタを形成していることが分

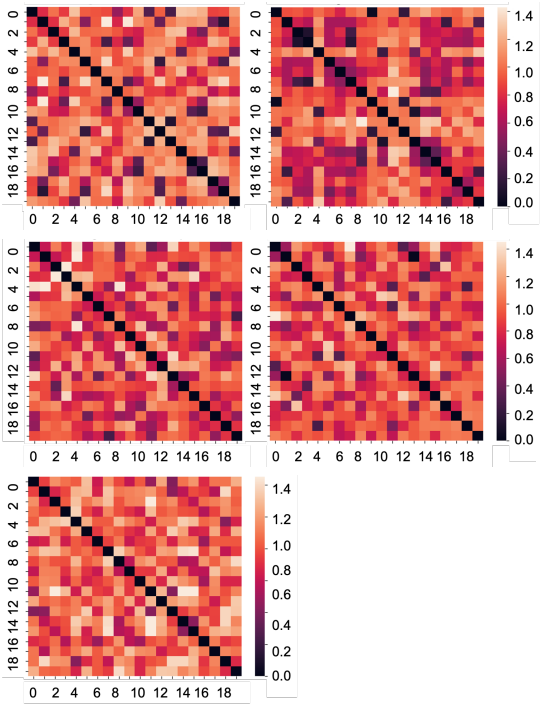

図 4: それぞれ，ドラム音（上段左），ベース音（上段右），ピアノ音（中段左），ギター音（中段右），混合音（下段）で学習した楽曲間類似度基準を用いて計算した 20 曲のテスト楽曲の平均距離行列．

かる．一方，分離楽器音を用いた場合は，分離精度の影響を受け，クラスタを形成していない部分があり，特に表 1 における精度の低いベース音では，精度の高いドラム音に比べてクラスタを形成していない部分が大きい．

図 4 に各平均距離行列を可視化したものを示す．これらの図から，異なる楽器音で学習した特徴表現空間の違いを視覚的に確認することができる．表 2 に，相関係数とスピアマンの順位相関係数を示す．いずれも 0 に近い値を示しており，各楽器音を用いて異なる類似度基準をうまく学習できていることがわかる．また，表 3 に，MRR スコアを示す．全ての値が 1/3 付近もしくはそれより小さい値をとっていることから，各楽器音に着目した際に類似している楽曲は，混合音に着目した際は似ているとは限らないということがわかる．

4.3.2 主観評価

図 5 に主観評価実験の結果を示す．ポジティブ，ネガティブサンプルが選択された場合をそれぞれ“True”と“False”で表している．また，“+”と“-” はそれぞれ，被験者がその選択に自信があった場合と自信がなかった場合を示す．表 4 に示すように，自信のない場合も含めると，被験者はすべての場合で 50%以上ポジティブサンプルを選択し，混合，ベース，ギターの場合では 70%より高い割合が得られている．これらの結果から，提案手法で学習した特定の音の種類に着目した楽曲間類似度基準は，着目した観点において知覚的に類似したセグメントを見つけることができ，人間の感覚の知覚とよく対応することが分かる．

表 2: 20 曲のテスト楽曲の平均距離に対する，異なる楽曲間類似度基準間の相関．

(a) 相関係数

楽器	混合	ドラム	ベース	ピアノ	ギター
混合	1	0.23	0.11	0.12	0.32
ドラム		1	-0.014	0.14	0.026
ベース			1	0.028	0.015
ピアノ				1	0.13
ギター					1

(b) Spearman の順位相関

楽器	混合	ドラム	ベース	ピアノ	ギター
混合	1	-0.013	0.12	-0.078	0.060
ドラム		1	0.17	-0.014	0.11
ベース			1	0.063	0.047
ピアノ				1	0.15
ギター					1

表 3: 各楽曲間類似度基準の MRR スコア

楽器	1 位	2 位	3 位
ドラム	0.31	0.35	0.22
ベース	0.20	0.25	0.12
ピアノ	0.28	0.25	0.20
ギター	0.32	0.22	0.14

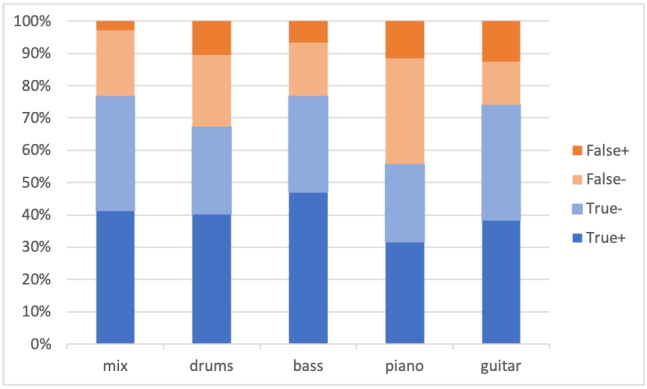

図 5: 楽曲間類似度基準と知覚される類似度の対応についての主観評価結果．

表 4: 主観評価による True の割合と 95%信頼区間．

楽器	True の割合 [%]
混合	76.92 ± 0.27
ドラム	67.31 ± 0.25
ベース	76.92 ± 0.32
ピアノ	55.77 ± 0.28
ギター	74.04 ± 0.24

5. 結論

本報告では，各楽器音に着目した楽曲間類似度算出手法を提案した．提案手法は，トリプレット損失を用いた深層距離学習により類似度基準を学習する．実験により，異なる楽器音に着目した類似度基準の学習が可能であることが示された．また，一部の楽器音を用いて学習した類似度基準は，楽曲そのものを用いて学習した類似度基準よりも精度の高い結果を導くことがわかった．しかし，分離した楽器音を用いて学習すると性能が低下することがわかった．主観評価実験の結果，提案した楽曲間類似度基準は知覚的な類似度によく対応することが明らかになった．

謝辞 本研究の一部は JST，CREST，JPMJCR19A3 の支援を受けたものである．

参考文献

[1] Y. Song, S. Dixon, M. Pearce, “A survey of music recommendation systems and future perspectives,” in *Proc. of CMMR*, pp. 395–410, 2012.
[2] M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, “Content-based music information retrieval: Current directions and future challenges,” *Proc. of the IEEE*, 96(4), pp. 668–696, 2008.
[3] P. Hamel and D. Eck, “Learning features from music audio with deep belief networks,” in *Proc. of ISMIR*, pp. 339–344, 2010.
[4] A. Elbir, N. Aydin, “Music genre classification and music recommendation by using deep learning,” *Electronics Letters*, 56(12), pp. 627–629, 2020.
[5] M.S. Fathollahi, F. Razzazi, “Music similarity measurement and recommendation system using convolutional neural networks,” *International Journal of Multimedia Information Retrieval*, 10(1), pp. 43–53, 2021.
[6] R. Lu, K. Wu, Z. Duan, C. Zhang, “Deep ranking: Triplet matchnet for music metric learning,” in *Proc. of IEEE ICASSP*, pp .121–125, 2017.
[7] J. Cleveland, D. Cheng, M. Zhou, T. Joachims, D. Turnbull, “Content-based music similarity with triplet networks,” *arXiv:2008.04938*, 2020.
[8] J. Lee, N.J. Bryan, J. Salamon, Z. Jin, J. Nam, “Disentangled multidimensional metric learning for music similarity,” in *Proc. of IEEE ICASSP*, pp. 6–10, 2020.
[9] J. Choi, J. Lee, J. Park, J. Nam, “Zero-shot learning for audio-based music classification and tagging,” in *Proc. of ISMIR*, pp. 67–74, 2019.
[10] E. Hoffer, N. Ailon, “Deep metric learning using triplet network,” in *Proc. of SIMBAD*, pp. 84–92, 2015.
[11] T.L.H. Li, A.B. Chan, A.H.W. Chun, “Automatic musical pattern feature extraction using convolutional neural network,” *Proc. of IMECS*, 1, pp. 546–550, 2010.
[12] M. Kaya, H.Ş. Bilge, “Deep metric learning: a survey,” *Symmetry*, 11(9), p.1066, 2019.
[13] E. Manilow, G. Wichern, P. Seetharaman, J. Le Roux, “Cutting music source separation some slakh: a dataset to study the impact of training data quality and quantity,” in *Proc. of IEEE WASPAA*, pp. 45–49, 2019.
[14] R. Hennequin, A. Khlif, F. Voituret, M. Moussallam, “Spleeter: a fast and efficient music source separation tool with pre-trained models,” *The Journal of Open Source Software*, 5(50), p. 2154, 2020.
[15] C. Spearman, “The proof and measurement of association Between Two Things,” *American Journal of Psychology*, 15, pp. 72–101, 1904.
[16] D. R. Radev, H. Qi, H. Wu, W. Fan, “Evaluating Web-based Question Answering Systems,” in *Proc. of ELRA*, pp. 1153–1156, 2002.
[17] L. van der Maaten, G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, 9, pp. 2579–2605, 2008.