知覚感情の不整合：人間が作曲した音楽，音楽を記述したテキスト，テキスト楽音合成による音楽の比較

阪井 瞭介[1,a)]　福田 航希[1]　松下 嶺佑[1]　高道 慎之介[1,2,b)]　植村あい子[3]

概要：本研究では，人間が作曲した音楽，音楽を記述したテキスト，T2M（text-to-music）が生成した音楽を対象に知覚感情における差異を分析する．聴取者に対してラッセルの円環モデルを用いて感情座標を評価させ，前述した 3 種のデータから聴取者が享受する感情の特徴を比較した．さらに聴取者の性別，年齢，楽器経験などの属性情報を取得し，これらが感情評価に与える影響についても調査した．本研究の結果は T2M モデルの入出力であるテキストと音楽間の知覚感情一致に向けた知見を提供する．

1. はじめに

深層学習の進展により，生成 AI は音楽生成の分野で注目を集めている．音楽生成 AI は入力情報に従った音楽を生成する深層学習モデルであり，MusicGen [1], MusicLM [2], Stable Audio [3]，Jukebox [4] などが開発されている．これらは，入力テキストに従った音楽を生成することから text-to-music（T2M）モデルと呼ばれており，T2M モデルで生成された音楽は商業施設や教育分野など幅広い用途で使用されている [5].

T2M を普及させるには T2M で合成された音楽（T2M 合成音楽）を評価すべきである．その評価軸として，人間の作曲した音楽（人間作曲音楽），あるいは T2M モデルに与えたテキスト（テキストプロンプト）との整合を図るものがある．例えば，Fréchet audio distance (FAD) [6] による評価は，人間作曲音楽と T2M 合成音楽間で音響特徴量分布の差異を測ることで，両音楽が音響特徴量レベルで整合することを目指している [1]．また，contrastive language-audio pretraining (CLAP) [7] による評価は，テキストプロンプトと T2M 合成音楽間で埋め込みベクトル距離を測ることで，テキストプロンプトと T2M 合成音楽の内容（例えば，楽器構成，リズム，テンポ）が整合することを目指している．これらは，特徴量レベル，あるいはテキスト表層レベル[*1]においてデータ間の整合を図るものである．

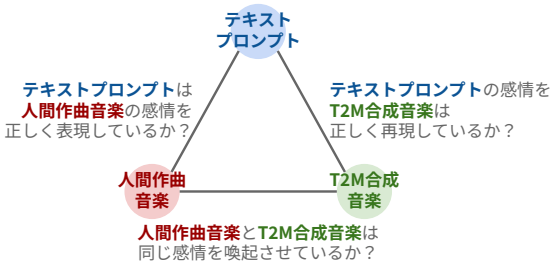

図 1　本研究の調査内容．3 種データの間で感情が整合するかを調査する．

一方で，音楽を聴取する目的の一つが感情の喚起であること [8] を鑑みれば，感情レベルにおいても整合を測るべきである．具体的には，図 1 で示すようにテキストプロンプトからその T2M 合成音楽に期待する感情，人間作曲音楽から聴取者が知覚する感情，および T2M 合成音楽から聴取者が知覚する感情の整合を測る．この整合を測り向上させることで，感情レベルで整合する T2M モデルの学習が可能

[1] 慶應義塾大学
Keio University
[2] 東京大学
University of Tokyo
[3] 津田塾大学
Tsuda University
[a)] r73ryo.s@keio.jp
[b)] shinnosuke_takamichi@keio.jp

[*1] テキストに表記されており，推論を伴わない内容

になると思われる．また，音楽から知覚される感情が，聴取者の文化的背景，年齢，楽器経験などに大きく依存すること [9], [10], [11], [12] を踏まえると，整合は聴取者属性に応じて測られるべきである．

そこで本研究では，テキストプロンプト，人間作曲音楽，T2M 合成音楽に対して，大規模な感情知覚評価を実施した結果を報告する．本論文ではこの対応のある 3 種のデータを “メディアデータ対” と呼ぶことにする．3 種のデータそれぞれについてラッセルの円環モデル [13] 上で感情を回答させ，3 種の間で知覚感情が一致するかを調査する．評価時には聴取者属性も取集し，聴取者属性が知覚感情の一致に与える影響を調査する．本研究で得た回答内容はプロジェクトページ [*2] にて公開されている．

2. 関連研究

2.1 T2M モデル

T2M モデルとは，テキストプロンプトを入力として音楽を生成する深層学習モデルである [1], [2], [3], [4]．このモデルはテキストプロンプトをエンコードし，その内容に基づいてメロディー，ハーモニー，リズムなどを構成する音楽を自動合成することを目的としている．具体的には「リラックスした雰囲気のピアノ音楽」や「エネルギッシュなギターロック」といったプロンプトを与えることで，それらに応じた音楽が生成される．

T2M モデルの例が MusicGen [1] である．MusicGen のアーキテクチャは自己回帰型の Transformer デコーダ [14] を基盤としている．音楽データは EnCodec [15] を用いて量子化され，コードブック表現として圧縮される．Transformer はテキストエンコーダの埋め込み表現と過去に生成した音楽トークンの系列を入力として受け取り，次のトークンを予測する．さらに，複数のストリームを効率的に処理する設計により，高次元の音楽データを低次元の表現に変換しつつ，高品質な音楽生成を実現している．

T2M モデルの評価は，T2M 合成音楽単体に対する評価と，T2M 合成音楽に対応のあるデータとの評価に大別される．前者は，例えば音楽らしさ [16]，好感度 [17]，快適さ [18]，一貫性 [19]，合成音楽に感じるか [20] に対するリッカート尺度である．後者の例は，FAD [6] と CLAP [7] である．FAD と CLAP はそれぞれ，人間作曲音楽と T2M 合成音楽，テキストプロンプトと T2M 合成音楽の間の整合を定量化するものである．これらの整合は，音響特徴量あるいはテキスト表層レベルで各データが整合することのみを目指しており，各データから人間が知覚する感情の整合については言及されていない．

[*2] https://github.com/takamichi-lab/sakai25_music_dataset

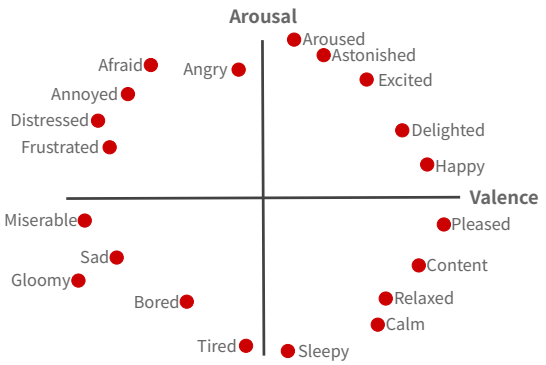

図 2 ラッセルの円環モデル ([13], [21] を参考にして作成)．横軸は感情価（快/不快），縦軸は覚醒度を表す．

2.2 感情知覚と音楽生成

T2M モデルにおける感情知覚の評価は図 1 で示している 3 点に着目して評価されるべきである．

(1) テキストプロンプトの感情を T2M 合成音楽は正しく再現しているか？
(2) テキストプロンプトは人間作曲音楽の感情を正しく表現しているか？
(3) 人間作曲音楽と T2M 合成音楽は同じ感情を喚起させているか？

(1)，(2)，(3) それぞれに対して 2.1 節で挙げた評価方法が存在するが，音楽感情予測および，自動音楽生成の研究における感情の評価で広く使われている方法 [9], [22], [23], [24], [25] としてラッセルの円環モデル [13] がある．ラッセルの円環モデルとは，図 2 のように感情を arousal（覚醒–沈静）で定義される縦軸と valence（快–不快）で定義される横軸で構成される 2 次元平面上に配置する．例えば，Excited は縦軸，横軸ともに正の象限である第一象限に配置され，Sad は縦軸，横軸ともに負の象限である第三象限に配置される．感情の境界線は曖昧であり，「幸せ」「嬉しい」や「悲しい」「辛い」といった離散的な感情タグでの評価は人による影響が大きくなる [26] ため，ラッセルの円環モデルが使われることが多い．

(1) に着目した研究では，感情に基づく音楽生成アプリケーションの有効性とユーザー満足度を評価する研究がある [27]．この研究では感情を含めたテキストを入力とした際に合成された音楽がテキストに含めた感情を持つものであるかを調査している．ここでの感情の評価方法は，「この音楽はテキストが持つ感情を持つか」に対する「はい・いいえ」での回答である．しかし，別の先行研究 [26] でも述べら

れているように，感情の境界線は曖昧であるため感情タグのように感情を離散値として扱うのは難しい．以上を踏まえると感情評価は感情の連続値を扱えるラッセルの円環モデルを用いて測られるべきである．

（3）に着目した研究では，感情に基づく音楽生成モデルを提案し，合成された音楽の感情知覚を調査した研究がある [10], [24]．前者の研究はラッセルの円環モデルを用いた感情知覚評価を，後者の研究はリッカート尺度による感情知覚評価を実施している．前者の研究ではモデルを構築するにあたって EMOPIA [25] を学習データセットとして使用しているため，合成された音楽と対応する EMOPIA の音楽との間での感情知覚の比較を行っている．しかし，ここでの感情知覚の比較はラッセルの円環モデルにおける象限の一致度という粒度の荒い方法であり，人間作曲音楽と T2M 合成音楽が同じ感情を喚起させているかを測るためには，象限だけでなく座標での関係性を調査するべきである．更に，EMOPIA の使用も適切ではない．何故ならば，EMOPIA の全キャプションは特定の数名の作成者によってつけられたものであり，キャプションに含まれる作成者バイアスを排除できないためである．

2.3 聴取者属性が知覚感情に与える影響

音楽を聴取して知覚される感情は，聴取者の属性によって異なることが先行研究で示されている [9], [10], [11], [12], [28]．具体的には以下のようなことが確認されている．

- 年齢：若者層は高齢層と比較して音楽に対する感情知覚の範囲が広いことが確認されている [12]．
- 性別：女性は音楽を通じた感情表現に対して，より敏感であり，感情的な反応が強いことが確認されている [28]．
- 楽器演奏経験：楽器演奏経験の無い方が気分変化を起こしやすいことが確認されている [29]．

本研究においても感情知覚は聴取者属性に依存すると予想されるため，知覚感情に加え聴取者属性を収集する．

3. 提案手法

3.1 概要

本研究ではテキスト，人間作曲音楽，T2M 合成音楽をそれぞれ聴取者に提示し，ラッセルの円環モデルを用いた感情評価を行う．聴取者に図 3 で示すようにテキスト，人間作曲音楽，T2M 合成音楽それぞれと，ラッセルの円環モデルの軸が書かれた二次元感情平面が含まれた評価ページを提示する．テキストの場合は，当該プロンプトから合成される T2M 合成音楽に期待する感情[*3]を，人間作曲音楽と T2M 合成音楽の場合には，当該音楽から知覚する感情を，聴取者に回答させる．語の印象による誘導を避けるため，感情表現語（例えば，図 2 における “Excited”）は平面に表示しない．この回答より各データに対応する感情平面上の座標（以降，感情座標）を得る．音楽の場合には更に，当該音楽が人間作曲音楽と T2M 合成音楽のどちらに聴こえるかを回答させる．すべてのテキストと音楽について感情座標を回答させた後，感情座標と聴取者属性を対応させるため，自身の属性を聴取者に回答させる．

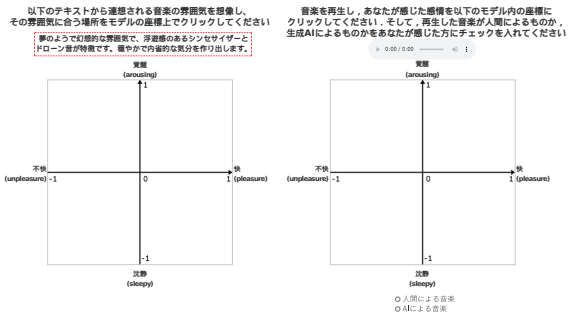

図 3 評価ページ．掲載にあたり，些末な説明文を削除するなどして整形している．この図ではテキストと音楽の評価を並べて表示しているが，実際のページはテキストと各音楽で別ページになっている．

3.2 刺激音・刺激テキストの作成

本研究では，以下の 3 つを対とする刺激を聴取者に提示する．

(1) 感情表現語を含むテキストプロンプト
(2) テキストプロンプトに対応する人間作曲音楽
(3) テキストプロンプトに対応する T2M 合成音楽

まず，人間作曲音楽データベースから，(1) と (2) を収集する．このとき，(1) は音楽を説明するキャプションである．次に，(1) を学習済み T2M モデルに与えることで (3) を得る．

意図しないバイアスが聴取者間で発生することを避けるため，各聴取者に提示する刺激を，ラッセルの円環モデル上でバランスよく配する．具体的には，(1) に含まれる感情表現語（感情ラベル）が，円環モデルの各象限に同数だけ含まれるよう，各聴取者の刺激を決定する[*4]．

[*3] テキスト自体から知覚される感情ではなく，そこから合成される音楽に期待する感情であることに注意する．
[*4] 第一象限 (Excited, Aroused など)，第二象限 (Angry, Annoyed など)，… の感情表現語を持つ (1)，および対応する (2)，(3) が，各象限について各聴取者の刺激に同数だけ含まれることを意味する．

4. 実験的評価

4.1 実験条件

本実験はクラウドソーシングプラットフォームであるランサーズ[*5] にて聴取者を募集した．各聴取者には 250 円の謝金を支払った．

4.1.1 テキストプロンプトと人間作曲音楽の収集

テキストプロンプトと人間作曲音楽は，音楽素材提供サービスである Shutterstock[*6]からダウンロードした．ダウンロードするプロンプトと音楽は，感情ラベルを検索クエリとして決定した．具体的には，ラッセルの円環モデルの各象限から 3 個ずつ感情ラベルを抽出し，各感情ラベルを Shutterstock の検索クエリとした．具体的な感情ラベルは，Astonished, Aroused, Excited, Angry, Annoyed, Frustrated, Miserable, Tired, Sad, Sleepy, Pleased, Calm の 12 個とした．各検索クエリについて人気順で 10 曲をダウンロードした．すなわち，人間作曲音楽の総数は 120 曲である．本研究では，器楽曲 (ボーカルなし曲) のみを対象とし，声楽曲 (ボーカルあり曲) が含まれないよう，声楽曲を手作業で排除した．各音楽の時間長は 15 秒，サンプリング周波数は 32 kHz，チャネル数は 1（モノラル），フォーマットは mp3 とした．Shutterstock は時間長を指定して音楽をダウンロードできるため，時間長を 15 秒[*7]に指定した．ただし，ダウンロードされたファイルには 15 秒を超えるファイルも存在した．そのため冒頭の 15 秒あるいは末尾の 15 秒のうち有音区間を多く含む方を切り出した．音楽に付随する英語キャプション（テキストプロンプト）は，各楽曲で異なる楽曲配布者（作曲者）が記載したものである[*8]．

4.1.2 T2M 合成音楽の収集

種々の T2M モデルが提案されている [1], [2], [3], [4] が，本研究では，オープンソースであることと学習データ構成が部分的に明記されていることから，MusicGen [1] を採用した．このモデルは，Shutterstock の音楽データを学習データとし，英語テキストプロンプトを入力とする．複数モデルによる評価は今後の予定とする．上記で収集したプロンプトを当該モデルに入力し，T2M 合成音楽を得た．MusicGen の学習データはすべて器楽曲であり，得られる T2M 合成音楽もすべて器楽曲である．データ総数，各音楽の時間長，サンプリング周波数，チャネル数，およびフォーマットは，それぞれ人間作曲音楽と同じとした．MusicGen は確率モデルであり，音楽の確率的サンプリングに関連するハイパーパラメータ（例えば，サンプリング温度 [30]，Top-k [31], Top-p [32]）を有する．これらの値は，すべてオープンソース[*9]の既定値とした．また，各プロンプトに対する合成回数は 1 回とした．

4.1.3 刺激音・刺激テキスト

各聴取者に提示する刺激は，各象限からランダムに選ばれた，重複のない 2 つの感情ラベルのデータを含むよう構成された．すなわち，各聴取者に提示する刺激数は，2 (感情ラベル数/象限) ×4 (象限) ×3 (メディアデータ対) = 24 である．24 個の刺激からなるセットを 30 セットを用意し，聴取者には 30 セットからランダムに選んだ 1 セットを提示して評価させた．テキストプロンプトは英語表記である一方，聴取者は日本語話者とした．そのため聴取者に提示する前に，ChatGPT-4o[*10]を用いてテキストプロンプトを日本語に翻訳した．その際，“I will now send you some English sentences, so please translate them into Japanese. When translating, please try to avoid using katakana as much as possible and translate to a level of Japanese that is appropriate for a junior high school student.”[*11] のプロンプトを与えた．これは，プロンプトを与えずに翻訳した際に，音楽を表す英語句の多くがカタカナで翻字され，翻訳文の意味を聴取者が捉えられないと考えたためである．

4.1.4 評価ページ

評価内容とラッセルの円環モデルに関する説明を，評価開始前に聴取者に提示した．テキストに対する評価では「テキストから連想される音楽の雰囲気を想像」の説明文を，音楽に対する評価では「音楽を再生し，あなたが感じた感情」の説明文を提示し，ラッセルの円環モデルの座標 1 点を選択させた．音楽に対する評価ではさらに，「音楽が人間によるものか，生成 AI によるものか」の説明文を提示し，そのどちらかを強制選択させた（図 3）．評価ページにおける感情平面のフォーマットは既存研究 [24] を参考にした．データの提示順は，対となるテキストプロンプト，音楽，音楽の順であり，人間作曲音楽と T2M 合成音楽の順はランダムとした．

[*5] https://www.lancers.jp
[*6] https://www.shutterstock.com/ja/music
[*7] shuttestock の配布音楽の最短時間が 15 秒であったため，15 秒に設定した．
[*8] すべての音楽について作曲者が異なるわけではなく，同一の作曲者（キャプション作成者）による複数の音楽を含む
[*9] https://github.com/facebookresearch/audiocraft
[*10] https://openai.com/index/hello-gpt-4o/
[*11] 日本語訳：“英語のテキストを送るので日本語に翻訳してください．翻訳の際，カタカナの使用をできるだけ避けて，中学生でも内容を把握可能な表記にしてください．”

表 1 質問 01：被験者の性別

Male	Female	N/A	Total
133	72	1	206

表 2 質問 02：被験者の年齢

10s	20s	30s	40s	50s	60s	70s<	Total
0	9	45	94	41	13	4	206

表 3 質問 04：被験者の好きなジャンル（左からポップ/シンセポップ，ロック，クラシック/オーケストラ）

Pop	Rock	Classic	Others	Total
87	52	21	46	206

表 4 質問 08：被験者の楽器経験の有無

No	Yes	Total
104	102	206

表 5 質問 09：被験者の経験のある楽器数（回答から計算）

None	1	2	3	4	5	Total
104	63	19	16	3	1	206

表 6 質問 10：被験者の楽器演奏の経験年数

None	<1	1-2	2-3	3-4	4<	Total
104	9	12	13	13	55	206

表 7 各被験者に対するアンケート内容

ID	Content
01 (単一選択)	性別を教えてください
02 (単一選択)	年齢を教えてください
03 (自由記述)	過ごした時間が一番長い国を教えてください
04 (単一選択)	一番好きな音楽のジャンルを教えてください
05 (単一選択)	今の調子を教えてください
06 (単一選択)	何時ごろに実験を行ったかを教えてください
07 (単一選択)	どこで実験を行ったかを教えてください
08 (単一選択)	楽器演奏経験の有無を教えてください
09 (複数選択)	演奏経験のある楽器の種類を教えてください
10 (単一選択)	楽器演奏の経験年数 (合計) を教えてください

4.1.5 聴取者属性

表 7 に示す聴取者属性アンケートを実施した．表 1 から表 6 に，各属性の聴取者数を示す[*12]．ただし，回答数が少数であった要素（例えば質問 04 の映画音楽，ブルース，EDM，チルアウト，カントリー，ヒップホップ，ジャズ）は，Others として集計した．

4.1.6 調査内容

本節では，以下の内容について調査する．

- **メディアデータ対の間で感情平面の象限を跨ぐ混同が生じるか？**：ラッセルの円環モデルの最も大きい区分は象限である．メディアデータ対の間で象限を跨ぐほどに大きな不整合があるかを調査する．例えば，テキストプロンプトは第一象限の感情座標を示すにもかかわらず，T2M 合成音楽は第三象限の感情座標を示す場合を，ここでは不整合と呼んでいる．
- **メディアデータ対の間で感情座標の分布は異なるか？**：より細かい不整合として，メディアデータ対の間で感情座標分布が異なるかを調査する．感情座標を感情価軸と覚醒軸に分け，それぞれの軸で分布間に有意な差があるかを調査する．
- **各メディアデータについて，聴取者属性は感情座標分布に影響を与えるか？**：聴取者属性によって，感情座標の感情価軸と覚醒軸に有意な差が生まれるかを調査する．

[*12] これ以外の質問 ID については，ほぼ全員から同じ回答が得られたので記載していない．詳細はプロジェクトページを参照されたい．

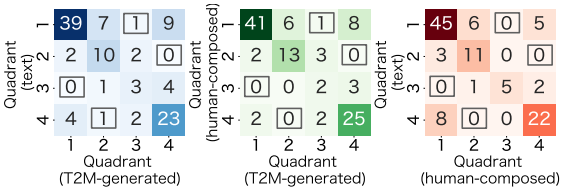

図 4 感情平面における象限の混同

4.2 感情平面における象限の混同

象限を超える混同（感情の整合不整合）を調査するため，テキスト，人間作曲音楽，T2M 合成音楽の間の象限混同数を計算した．実験では，各刺激（音，テキスト）について複数聴取者による座標が回答されている．そのため，複数聴取者の回答座標を平均したものを各刺激の代表座標とした．

テキスト，人間作曲音楽，T2M 合成音楽の間の象限混合の結果を図 4 に示す[*13]．多くの場合で，テキスト，人間作曲音楽，T2M 合成音楽の間で象限が一致（各図の対角成分）している．また，隣接しない象限への混同（各図の四角）がほとんど存在していない．そのため，テキスト，人間作曲音楽，T2M 合成音楽の間で知覚感情はある程度整合すると考えられる．

4.3 対データ間での感情座標分布の差異

座標での感情の整合を調査するため，テキスト，人間作曲音楽，T2M 合成音楽の間の感情座標分布を比較した．感情価軸と覚醒度軸における座標分布パラメータ（中央値，分散）について，データ間での有意差を調査する．3 群間の等分散の検定には Brown-Forsythe 検定 [33] を，Brown-Forsythe 検定の事後検定であるペアワイズ検定には Brown-Forsythe 検定を用いた．また，3 群間の中央値の検定には Friedman 検定 [34] を，Friedman 検定の事後検定であるペア

[*13] 例えば，対応するテキストプロンプトと T2M 合成音楽の象限がともに第一象限である場合，図 4 左の左上隅のセルにカウントされる．

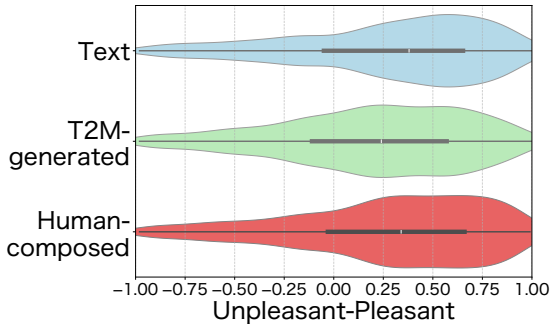

図 5 快–不快軸（横軸）についてのバイオリンプロット（上からテキスト，T2M 合成音楽，人間作曲音楽）

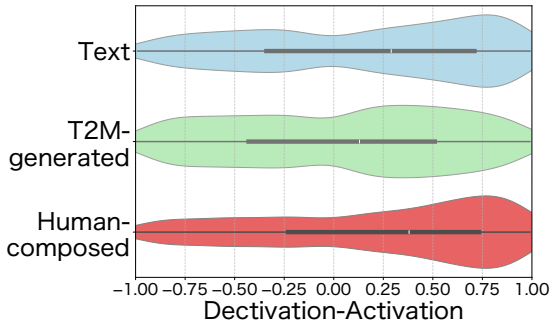

図 6 覚醒–沈静軸（縦軸）についてのバイオリンプロット（上からテキスト，T2M 合成音楽，人間作曲音楽）

表 8 3 種のデータの統計量 (横軸：快–不快, 縦軸：覚醒–沈静)．Med. は中央値，Var. は分散を表す．

	横軸：快–不快			縦軸：覚醒–沈静		
	Mean	Med.	Var.	Mean	Med.	Var.
Text	0.277	0.380	0.232	0.171	0.290	0.341
T2M-generated	0.195	0.240	0.221	0.053	0.130	0.308
Human-composed	0.271	0.340	0.233	0.232	0.380	0.336

ワイズ検定には Wilcoxon 検定 [35] を用いた．

4.3.1 快–不快軸における分散

テキスト–人間作曲音楽–T2M 合成音楽のそれぞれについて得た感情座標を分析する．図 5 は感情座標の快–不快軸（横軸）についてのバイオリンプロットである．この 3 群の等分散性を検定した結果，$p = 0.868$ で帰無仮説（3 群の分散が等しい）が棄却できなかった．この結果から，テキストから音楽に期待される感情，人間作曲音楽から知覚される感情，そして T2M 合成音楽から知覚される感情は，快–不快軸において同程度の広がりを持って分布することがわかる．換言すれば，T2M モデルは，快–不快軸において，テキストによる表現あるいは人間作曲音楽と同程度に多様な音楽を生成できると言える．

4.3.2 快–不快軸における中央値

3 群の中央値を検定した結果，$p < 10^{-9}$ で有意差があると判断された．次にペアワイズ検定を実施した結果，テキスト–人間作曲音楽間では有意差が認められなかった ($p = 0.650$) 一方で，テキスト–T2M 合成音楽間 ($p < 10^{-9}$)，人間作曲音楽–T2M 合成音楽間 ($p < 10^{-9}$) で有意差が認められた[*14]．表 8 の中央値を参照すると，T2M 合成音楽の中央値は他の中央値よりも原点（ニュートラル，平静）に接近していることがわかる．以上のことから T2M モデルは，快–不快軸において，モデル入力のテキストあるいは人間作曲音楽よりも平静寄りの感情を喚起することがわかる．

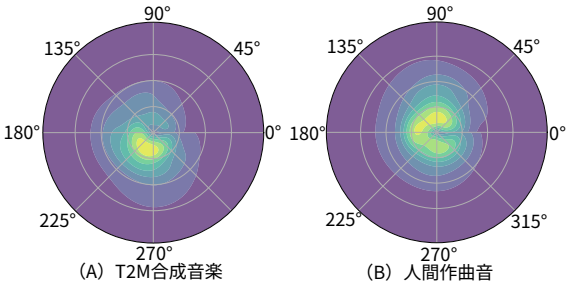

図 7 （A）T2M 合成音楽の座標からテキストの座標を引いた値の密度分布．（B）人間作曲音楽の座標からテキストの座標を引いた値の密度分布

4.3.3 覚醒–沈静軸における分散

一方，縦軸の覚醒–沈静軸についてのバイオリンプロットを図 6 に示す．この 3 群の等分散性の帰無仮説は，$p = 0.024$ で棄却された．その後のペアワイズ検定では，テキスト–人間作曲音楽間 ($p = 0.108$)，T2M 合成音楽–人間作曲音楽間 ($p = 0.298$) で有意差が認められなかった一方で，テキスト–T2M 合成音楽間で有意差が認められた ($p = 0.004$)．表 8 を参照すると，テキストについて得た感情座標分布の分散が 0.341，T2M 合成音楽について得た感情座標分布の分散が 0.308 である．これらの結果より T2M モデルで合成される音楽には，覚醒–沈静軸においてテキストから音楽に期待される感情ほどの表現の多様さがないことがわかる．

4.3.4 覚醒–沈静軸における中央値

3 群の中央値を検定した結果，$p < 10^{-9}$ で有意差があると判断され，ペアワイズ検定でもすべての組み合わせてで有意差があると判断された ($p < 10^{-4}$)[*15]．表 8 より，中央値は T2M 合成音楽 < テキスト < 人間作曲音楽であることから，人間作曲音楽はテキストに比べ過剰に，T2M 合成音楽は逆に過少に覚醒–沈静を表現していると言える．

[*14] ペアワイズ検定で有意差のあった組み合わせにおける Cohen's d はそれぞれ，0.172 , 0.160 であった．
[*15] Cohen's d はテキスト–T2M 合成音楽間，テキスト–人間作曲音楽間，人間作曲音楽–T2M 合成音楽間でそれぞれ，0.207 , 0.105 , 0.316 であった．

4.3.5 座標の差の分析

これを詳細に分析するために，メディアデータ対間で座標の距離を計算した．具体的には，各対について，T2M 合成音楽の座標からテキストの座標を引いた値と，人間作曲音楽の座標からテキストの座標を引いた値を計算した．すなわち，テキストの座標を基準として，T2M 合成音楽と人間作曲音楽の座標がそれぞれどのように変化したかを計算した．仮にこの変化が一様であれば，この変化はテキストと音楽の対に依存せず別の要因に依存すると言える．

T2M 合成音楽の座標からテキストの座標を引いた値と，人間作曲音楽の座標からテキストの座標を引いた値について，その密度分布をそれぞれ図 7 に示す．それぞれについて，角度の一様性を Hodges-Ajne 検定 [36] した結果，T2M 合成音楽，人間作曲音楽ともに $p < 10^{-5}$ で有意差が認められた．またそれぞれの角度の中央値を Common Median Test[*16]した結果，$p < 10^{-5}$ で有意差が認められた．具体的な中央値は T2M 合成音楽が 242.43 °，人間作曲音楽が 181.07 °であった．これよりテキストを基準とした音楽の知覚感情の評価の差異は一様ではなく，テキストと音楽の対に依存すると言える．T2M 音楽の場合は左下（不快，沈静向き）に，人間作曲音楽の場合は左上（不快，覚醒向き）に強く分布する傾向にあり，両者で変化が異なることも興味深い．

4.4 聴取者属性が感情分布に与える影響

表 7 にあるアンケートで集めた聴取者属性が座標で表される感情分布に影響を与えるのかを調査するために，テキスト，人間作曲音楽，T2M 合成音楽それぞれに対する各属性要素の分布と，全属性要素の分布を比較した[*17]．表 1 から表 6 にある聴取者属性の各属性要素分布の比較を実施した[*18]．感情価軸と覚醒度軸における座標分布パラメータ（中央値，分散）について，聴取者の各属性要素間での有意差を調査する．属性間の等分散の検定には Brown-Forsythe 検定を，Brown-Forsythe 検定の事後検定であるペアワイズ検定には Brown-Forsythe 検定ペアワイズ検定を用いた．また，属性間の中央値の検定には Kruskal-wallis 検定 [37] を，Friedman 検定の事後検定であるペアワイズ検定には Dunn 検定 [38] を用いた．

[*16] https://github.com/circstat/pycircstat
[*17] 性別の場合は，男性，女性，全体 (両性別+回答無し) の 3 群．
[*18] それぞれの属性要素を持つ人数が 20 人に満たない属性内要素は結合している．例えば年齢は 40 歳未満（54 人），40 代（94 人），50 歳以上（58 人）のように結合した．

表 9 音楽座標における，横軸：快–不快の統計量．Med. は中央値，Var. は分散を表す．

	人間作曲音楽座標			T2M 合成音楽座標		
	Mean	Med.	Var.	Mean	Med.	Var.
All	0.271	0.340	0.233	0.195	0.240	0.221
Q09: None	0.271	0.340	0.239	0.210	0.255	0.218
Q09: 1	0.266	0.390	0.256	0.169	0.230	0.246
Q09: 2<	0.280	0.320	0.179	0.198	0.220	0.188

4.4.1 テキスト座標における快–不快軸の分析

テキスト座標における快–不快軸の属性要素の分布比較を行った．

等分散性の比較．全属性要素の分布と各属性要素の分布で等分散性を検定した結果, $p = 0.007$ で性別属性で帰無仮説が棄却され，他の属性では棄却されなかった（表 1 から表 6 の属性順に $p = 0.007$，$p = 0.073$, $p = 0.280$，$p = 0.948$, $p = 0.092$, $p = 0.507$ であった）．次にペアワイズ検定を実施したところ女性と男性の間で $p = 0.002$ で，女性と全体の間で $p = 0.027$ で有意差が認められた．女性データの分散は 0.266 , 男性データの分散は 0.215 であったことから女性の方が男性よりも快–不快軸において知覚する感情の広がりが広いと言える．これは 2.3 で触れたように女性の方が男性より感情的な反応が強いこと [28] からも納得できる内容である．

中央値の比較． 全属性要素の分布と各属性要素の分布間で中央値を検定した結果, 年齢属性（$p = 0.009$）と好きなジャンル属性（$p = 0.033$）で有意差が認められた．年齢属性に関しては，50 歳以上の分布と他の全ての分布との組み合わせの比較でそれぞれ有意差が認められた．具体的な中央値は 40 歳未満，40 代ではともに 0.350 であり，50 歳以上が 0.445 であった．このことから年齢が上がるにつれてより音楽をポジティブに知覚することがわかった．好きなジャンル属性に関しては，クラシック/オーケストラを好む集合の分布と他の全ての分布との組み合わせの比較でそれぞれ有意差が認められた．具体的な中央値は，クラシック/オーケストラが 0.268，ポップ/シンセポップが 0.231，ロックが 0.223，全属性要素が 0.232 であった．これより，クラシック/オーケストラのような器楽曲を主に好む聴取者は，全体傾向よりも，快の感情をテキストからの音楽に期待しやすい可能性がある．

4.4.2 テキスト座標における覚醒–沈静軸の分析

テキスト座標における覚醒–沈静軸の属性要素の等分散性と中央値を検定した結果，全ての属性で有意差が認められなかった．すなわち，テキストから期待される音楽の覚醒–沈静は，聴取者属性によらず等しい分布であると言える．

表 10 音楽座標における，縦軸：覚醒–沈静の統計量．Med. は中央値，Var. は分散を表す．

	人間作曲音楽座標			T2M 合成音楽座標		
	Mean	Med.	Var.	Mean	Med.	Var.
All	0.232	0.380	0.336	0.053	0.130	0.308
Q04: Pop	0.203	0.360	0.366	0.032	0.130	0.325
Q04: Rock	0.259	0.41	0.303	0.065	0.140	0.279
Q04: Classic	0.288	0.440	0.310	0.108	0.185	0.287
Q08: Yes	0.252	0.395	0.305	0.077	0.160	0.289
Q08: No (Q09: None)	0.211	0.365	0.366	0.029	0.110	0.326
Q09: 1	0.269	0.430	0.317	0.078	0.150	0.312
Q09: 2<	0.224	0.340	0.286	0.077	0.180	0.252

4.4.3 音楽座標における快–不快軸の分析

人間作曲音楽座標と T2M 合成音楽座標のそれぞれについて属性要素の分布比較を行った．

等分散性の比較．全属性要素の分布と各属性要素の分布で等分散性を検定した結果，経験楽器数属性に対して，人間作曲音楽座標で $p = 0.006$，T2M 合成音楽座標で $p = 0.013$ となり，有意差が認められた．その後のペアワイズ検定では，人間作曲音楽と T2M 合成音楽のそれぞれで，2 種類以上楽器経験がある集合の分布と他の全ての分布との組み合わせでそれぞれ有意差が認められた．表 9 を参照すると具体的な分散値から楽器経験が豊富である方が快–不快軸において知覚する感情の広がりが狭くなると言える．

中央値の比較．一方，経験楽器数属性での中央値の検定では有意差が認められなかった．また人間作曲音楽に対する好きな音楽ジャンルの中央値（$p = 0.023$）を除き，他の属性の検定では等分散性と中央値ともにどれも有意差が認められなかった．

4.4.4 音楽座標における覚醒–沈静軸の分析

人間作曲音楽座標と T2M 合成音楽座標のそれぞれについて属性要素の分布比較を行った．

人間作曲音楽について． まず，人間作曲音楽座標の等分散性については，好きなジャンル属性（$p = 0.012$），楽器経験属性（$p = 0.010$），経験楽器数属性（$p = 0.013$）の 3 属性で有意差が認められた．ペアワイズ検定をした結果，以下の間で有意差が認められた．

- 好きなジャンル属性：ポップ/シンセポップ–クラシック/オーケストラ（$p = 0.044$），ポップ/シンセポップ–ロック（$p = 0.002$）
- 楽器演奏経験属性：経験なし–経験あり（$p = 0.003$）
- 経験楽器数属性：なし–1 種類（$p = 0.034$），なし–2 種類以上（$p = 0.002$），全属性要素–2 種類以上（$p = 0.036$）

一方で，中央値の検定ではすべてのケースで有意差が認められなかった．

T2M 合成音楽について．次に T2M 合成音楽座標の等分散性についても，好きなジャンル属性（$p = 0.017$），楽器経験属性（$p = 0.031$），経験有り楽器数属性（$p = 0.003$）の 3 属性で有意差が認められた．ペアワイズ検定をした結果，以下の間で有意差が認められた．

- 好きなジャンル属性：ポップ/シンセポップ–ロック（$p = 0.004$），全属性要素–ロック（$p = 0.032$）
- 楽器演奏経験属性：経験なし–経験あり（$p = 0.009$）
- 経験楽器数属性：なし–2 種類以上（$p = 0.002$），1 種類–2 種類以上（$p = 0.009$），全属性要素–2 種類以上（$p = 0.003$）

中央値については，人間作曲音楽と同様に，すべてのケースで有意差が認められなかった．

以上を総括すると，人間作曲音楽と T2M 合成音楽の両方で，同じ聴取者属性で有意差があることが言える． 両音楽でジャンル，楽器演奏経験，演奏経験楽器数の有意差が認められており，また，ペアワイズ検定においても同様の対に有意差が認められている．すなわち，これらの聴取者属性に起因する，人間作曲音楽と T2M 合成音楽で共通する覚醒–沈静知覚要素が存在すると期待される．この期待を強める要素として，分散値の大小が，人間作曲音楽と T2M 合成音楽の間で一致していることがあげられる（表 10）．例えば，ジャンルにおいてポップ/シンセポップの分散は音楽を問わず常に他ジャンル（ロック，全体）より大きく，同様に，楽器経験のない集合（経験楽器数が 0 の集合）は他要素に比べ常に分散が大きい．

5. まとめ

本研究では，人間が作曲した音楽，音楽を表現したテキスト，テキスト楽音合成 (text-to-music; T2M) で合成した音楽について，その知覚感情に差があるかを調査した．その結果，以下の知見を得た．

- T2M 合成音楽は，快–不快軸においてはテキストと同程度の広がりを持って分布するが，快–不快軸と覚醒–沈静軸の両方でその中央値が原点方向に縮退する．
- テキストを基準としたとき，人間作曲音楽と T2M 合成音楽は，それぞれ別の傾向を以って知覚感情が変化する．
- 性別や楽器演奏経験などの聴取者属性が，知覚感情に影響を与える．

これらに基づいて今後は，感情レベルで整合する

T2M モデルの開発が期待される．

謝辞： 本研究は，JSPS 科研費 23H03418，21H04900，23K18474，JST 創発的研究支援事業 JPMJFR226V の支援を受けて実施した．

参考文献

[1] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Advances in Neural Information Processing Systems*, 2023.

[2] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "MusicLM: Generating music from text," *arXiv arXiv:2301.11325*, 2023.

[3] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," in *ICML*, 2024.

[4] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[5] 朝日新聞, "音楽授業にＡＩ、みんな作曲家 埼玉の小学校、歌詞に対応したメロディーを自動生成," 2024 年 6 月 23 日. [Online]. Available: https://www.asahi.com/articles/DA3S15965163.html

[6] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[7] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning Audio Concepts From Natural Language Supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[8] 森数馬 and 岩永誠, "音楽と感情に関する研究の展開―心理反応, 末梢神経系活動, 音楽および音響特徴―," **心理学評論**, vol. 57, no. 2, pp. 215–234, 2014.

[9] E. Y. Koh, K. W. Cheuk, K. Y. Heung, K. R. Agres, and D. Herremans, "MERP: A music dataset with emotion ratings and raters' profile information," *Sensors*, vol. 23, no. 1, p. 382, 2022.

[10] X. Shi and C. Vear, "Generating Emotional Music Based on Improved C-RNN-GAN," in *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*. Springer, 2024, pp. 357–372.

[11] H. Lee, F. Hoeger, M. Schoenwiesner, M. Park, and N. Jacoby, "Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms," in *ISMIR*, 2022.

[12] M. T. Pearce and A. R. Halpern, "Age-related patterns in emotions evoked by music." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 9, no. 3, p. 248, 2015.

[13] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, 2017.

[15] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," in *ICLR*, 2024.

[16] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A. M. Dai, M. D. Hoffman, and D. Eck, "Music Transformer: Generating Music with Long-Term Structure," *arXiv preprint arXiv:1809.04281*, 2018.

[17] K. Déguernel, B. L. Sturm, and H. Maruri-Aguilar, "Investigating the relationship between liking and belief in AI authorship in the context of Irish traditional music," in *CREAI 2022 Workshop on Artificial Intelligence and Creativity*, 2022.

[18] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "MuseGan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[19] Y.-K. Wu, C.-Y. Chiu, and Y.-H. Yang, "JukeDrummer: Conditional Beat-aware Audio-domain Drum Accompaniment Generation via Transformer VQ-VAE," in *ISMIR*, 2022.

[20] P. Sarmento, J. Loth, and M. Barthet, "Between the AI and Me: Analysing Listeners' Perspectives on AI-and Human-Composed Progressive Metal Music," in *ISMIR*, 2024.

[21] Y.-S. Seo and J.-H. Huh, "Automatic emotion-based music classification for supporting intelligent iot applications," *Electronics*, vol. 8, no. 2, p. 164, 2019.

[22] J. Kang and D. Herremans, "Are We There Yet? A Brief Survey of Music Emotion Prediction Datasets, Models and Outstanding Challenges," *arXiv preprint arXiv:2406.08809*, 2024.

[23] A. Dash and K. Agres, "AI-Based Affective Music Generation Systems: A Review of Methods, and Challenges," *ACM Computing Surveys*, vol. 56, no. 11, pp. 1–34, 2024.

[24] N. Imasato, K. Miyazawa, C. Duncan, and T. Nagai, "Using a language model to generate music in its symbolic domain while controlling its perceived emotion," *IEEE Access*, vol. 11, pp. 52 412–52 428, 2023.

[25] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," in *ISMIR*, 2021.

[26] E.-G. Han, T.-K. Kang, and M.-T. Lim, "Physiological signal-based real-time emotion recognition based on exploiting mutual information with physiologically common features," *Electronics*, vol. 12, no. 13, p. 2933, 2023.

[27] C. Chong *et al.*, "Emotional music generation: An analysis of effectiveness and user satisfaction by using python and dart," in *CS & IT Conference Proceedings*, vol. 13, no. 5, 2023.

[28] S. B. Kamenetsky, D. S. Hill, and S. E. Trehub, "Effect of tempo and dynamics on the perception of emotion in music," *Psychology of Music*, vol. 25, no. 2, pp. 149–160, 1997.

[29] 渡辺恭子, “音楽経験とパーソナリティーに関する一考察,” **金城学院大学論集. 人文科学編**, vol. 10, no. 2, pp. 103–113, 2014.
[30] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi, “Optimization by simulated annealing,” *science*, vol. 220, no. 4598, pp. 671–680, 1983.
[31] A. Fan, M. Lewis, and Y. Dauphin, “Hierarchical Neural Story Generation,” in *ACL*, 2018.
[32] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The Curious Case of Neural Text Degeneration,” in *ICLR*, 2020.
[33] M. B. Brown and A. B. Forsythe, “Robust tests for the equality of variances,” *Journal of the American statistical association*, vol. 69, no. 346, pp. 364–367, 1974.
[34] M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.
[35] F. Wilcoxon, “Individual comparisons by ranking methods,” in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 196–202.
[36] E. Batschelet, “Recent statistical methods for orientation data,” *NASA, Washington Animal Orientation and Navigation*, 1972.
[37] W. H. Kruskal and W. A. Wallis, “Use of ranks in one-criterion variance analysis,” *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.
[38] O. J. Dunn, “Multiple comparisons using rank sums,” *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.