E-038

楽曲の感情価と覚せい度を予測するための歌詞と音響特徴に基づく統合手法
An Integrated Approach Using Lyrics and Acoustic Features to Predict Arousal and Valence in Music

渡邊 薪之助‡ Shinnosuke Watanabe
小俣 昌樹‡ Masaki Omata

1. はじめに

音楽が喚起する感情を予測・理解することは，音楽の構造や音響特徴といった多角的な観点から幅広く研究されてきた[1]．従来の音楽感情研究では，楽曲を「Happy」「Sad」「Angry」などの感情カテゴリに分類する手法が主流であり，歌詞や音響特徴を用いた多様な分類モデルが提案されてきた．特に近年では，Bidirectional Encoder Representations from Transformers（BERT）[10]や A Robustly Optimized BERT Pretraining Approach（RoBERTa）[11]といった Transformer 系の事前学習済み言語モデルの導入により，歌詞からの感情予測精度が向上している[3].

しかし，カテゴリ分類に基づく手法にはいくつかの課題がある．たとえば，「Happy」と分類された楽曲の中にも，活気に満ちたものと穏やかなものが存在するように，同一カテゴリ内での感情の強弱やニュアンスを表現することが難しい．このような課題を解決するために，覚せい度（Arousal）と感情価（Valence）の 2 次元で感情を表現するモデルが注目されている．これは心理学におけるラッセルの円環モデル[2]に基づくもので，感情の強弱を数値として定量的に扱うことが可能となる．

さらに，従来の研究では歌詞と音響特徴のどちらか一方のみを用いる手法が主流であった．しかし，歌詞は楽曲の「言語的側面」，音響特徴は「音響的側面」に基づく感情の手がかりを提供するため，両方を統合的に用いることでより言語と音響の両方から捉えた感情予測が可能になると考えられる．

本研究では，楽曲の感情予測において，歌詞からの予測に Transformer の回帰モデル，音響特徴からの予測に機械学習の回帰モデルをそれぞれ構築し，覚せい度および感情価の予測値を出力として得る．その上で，両モデルの予測結果に対して適切な重み（正の整数）を用いて加重平均を行い，最終的な楽曲の覚せい度および感情価を予測する手法を提案する．これにより，言語的・音響的手がかりを活用した楽曲感情予測を行う．

2. 関連研究

2.1 ラッセルの円環モデル

感情を定量的に扱う代表的なモデルとして，Russell が提案した覚せい度と感情価の 2 次元で感情を表現する円環モデル（図 1 参照）がある[2]．覚せい度が高い場合は「興奮」「緊張」，低い場合は「落ち着いた」「疲れ」などを表す．感情価は，正方向では「幸せ」「満足」などの快（ポジティブ）な感情，負方向では「怒り」「悲しみ」などの不快（ネガティブ）な感情を表す．

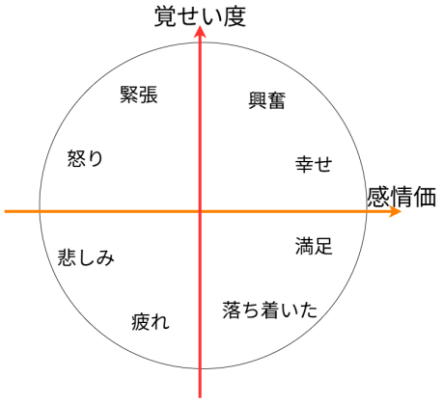

図 1　ラッセルの円環モデル

2.2 歌詞による感情予測

言語的側面に着目した研究として，近年では Transformer 系の大規模事前学習済み言語モデルを用いた歌詞感情分類の研究が進められている．Hung らは，歌詞を入力として Generalized Autoregressive Pretraining for Language Understanding（XLNet）を用いた楽曲感情認識モデルを提案した[3]．彼らの手法は，「Happy」「Sad」「Anger」などの感情カテゴリへの分類において従来手法よりの高い分類精度を示した．これらのモデルは文脈理解能力に優れており，高精度な感情特徴の抽出が可能であるとされている．

2.3 音響特徴による感情予測

音響的側面に着目した研究では，楽曲から得られるスペクトログラムや Mel-Frequency Cepstral Coefficients（MFCC），Zero-Crossing Rate（ゼロ交差率）などの特徴量を用いて，楽曲の感情を分類する手法が主に検討されている．Choi らは Convolutional Neural Network（CNN）を用いた楽曲感情分類手法を提案した[4]．彼らの手法は，明示的な特徴量の抽出を必要とせず，学習過程において分類に有効な特徴量を自動的に抽出できる点が特徴である．この手法は，従来の音響特徴に基づく手法と比較して，高い分類精度を示している．

3. 提案手法

提案手法の楽曲の歌詞と音響特徴から覚せい度（Arousal）と感情価（Valence）を予測する全体の流れを図 2 に示す．歌詞感情予測の入力は歌詞で，出力は歌詞の感情を表す $\mathrm{Arousal}_{\mathrm{lyrics}}$ および $\mathrm{Valence}_{\mathrm{lyrics}}$ である．そして，音響特徴感情予測の入力は音響特徴で，出力は音響特徴の感情を表す $\mathrm{Arousal}_{\mathrm{acoustic}}$ および $\mathrm{Valence}_{\mathrm{acoustic}}$ である．最後に，出

‡ 山梨大学　University of Yamanashi

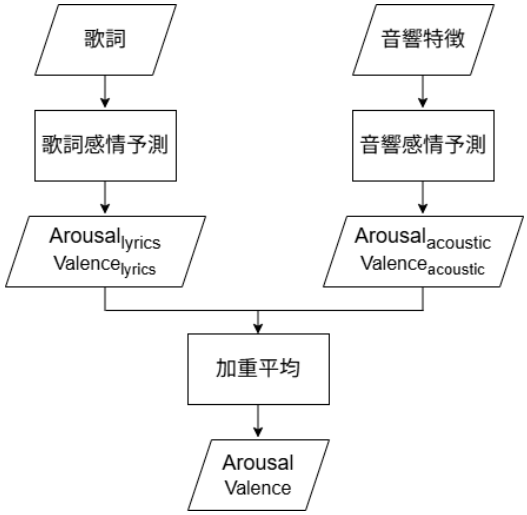

図 2　提案手法の全体の流れ

力された 0 から 1 の範囲の実数値を加重平均によって統合し，楽曲の最終的な感情を表す Arousal, Valence を出力する．

3.1 歌詞感情予測

歌詞から$\text{Arousal}_{\text{lyrics}}$ および$\text{Valence}_{\text{lyrics}}$の予測を行う．歌詞感情予測を行う歌詞モデルは，Transformer 系の事前学習済み言語モデルを用いて回帰モデルを構築した．具体的には，歌詞全体を入力とし，その意味的な情報を内部の文脈ベクトルとして抽出する．この抽出したベクトルを線形回帰層（全結合層）に入力して，$\text{Arousal}_{\text{lyrics}}$ および$\text{Valence}_{\text{lyrics}}$ に対応する 2 つの実数値を出力する．線形回帰層は，Transformer モデルが出力する文脈ベクトルに含まれる豊富な意味情報（語彙的特徴，語の前後関係，構文構造など）を，実数値へと効率的に変換する手段として有効である[5].

モデルは，各楽曲に付与された覚せい度および感情価の教師ラベル（0 から 1 の実数値）を正解値とし，出力との誤差を最小化するように損失関数を用いて最適化を行う．損失関数は平均二乗誤差（MSE）を用いた．

3.2 音響感情予測

音響特徴から$\text{Arousal}_{\text{acoustic}}$ および$\text{Valence}_{\text{acoustic}}$ の予測を行う．音響特徴感情予測を行う音響特徴モデルは，楽曲から抽出した音響特徴量を入力とし，$\text{Arousal}_{\text{acoustic}}$ および$\text{Valence}_{\text{acoustic}}$ の実数値を出力する回帰モデルを構築した．モデルは$\text{Arousal}_{\text{acoustic}}$ の予測と$\text{Valence}_{\text{acoustic}}$ 予測をそれぞれ独立した回帰問題として扱い，個別に学習を行っている．これにより，Arousal と Valence のそれぞれに特化した最適な予測モデルの構築を可能にしている．モデルは，音響特徴量を入力とし，各楽曲に付与された覚せい度および感情価の教師ラベル（0 から 1 の実数値）を正解値とし，出力との誤差を最小化するように損失関数（MSE）を用いて最適化を行う．

モデルの構築には，いくつかの機械学習アルゴリズムを候補として使用する．ここでいう機械学習アルゴリズムとは，入力となる音響特徴量と，出力として予測したい値（$\text{Arousal}_{\text{acoustic}}$ と $\text{Valence}_{\text{acoustic}}$）の関係を学習するための方法のことである．

3.3 加重平均

楽曲の最終的な Arousal および Valence は，歌詞モデルと音響特徴モデルの予測値に対してそれぞれ異なる重みを設定し，加重平均することで算出する．Arousal および Valence の算出する式を以下の式(1)(2)に示す．

$$A = \frac{a\,A_{lyrics} + b\,A_{acoustic}}{a+b} \tag{1}$$

$$V = \frac{c\,V_{lyrics} + d\,V_{acoustic}}{c+d} \tag{2}$$

ここで，AおよびVは楽曲の最終的な Arousal および Valence の値，A_{lyrics}およびV_{lyrics}は歌詞モデルによる予測値（$\text{Arousal}_{\text{lyrics}}$，$\text{Valence}_{\text{lyrics}}$），$A_{acoustic}$および$V_{acoustic}$は音響特徴モデルによる予測値（$\text{Arousal}_{\text{acoustic}}$，$\text{Valence}_{\text{acoustic}}$）を表す．また，重み係数 a，b，c，d はすべて正の整数であり，以下の式(3)の制約を満たす．

$$a+b=10, \quad c+d=10 \tag{3}$$

つまり，各重みは 0:10，1:9，2:8，...，10:0 のように 11 段階で設定される．

4. 実験

本実験の目的は，楽曲の感情予測において歌詞と音響特徴の両方を用いることで，それぞれを単独で用いた場合（歌詞モデル単独または音響特徴モデル単独）よりも高精度な感情予測が可能かを明らかにすることである．そのために，歌詞モデルおよび音響特徴モデルをそれぞれ構築・評価し，最良の予測精度を示したモデルを選定した．つぎに，選定したモデルの出力を加重平均によって統合し，各重みで予測精度を評価する統合評価をおこなった．そして，最も高い予測精度を示す最適な重みを特定し，その結果を単独モデルの予測精度と比較することで，統合による精度向上が見られるかどうかを検証した．

4.1 データセット

本研究では，歌詞モデルの構築・評価，音響特徴モデルの構築・評価，統合評価のそれぞれで，3 種類のデータセットを使用した．

4.1.1 歌詞データセット

歌詞データセットには，MoodyLyrics[6] を用いた．MoodyLyrics は，2,555 曲の英語ポップソングの歌詞を収録したデータセットであり，各楽曲には「Happy」「Sad」「Relaxed」「Angry」のいずれかのラベルが付与されている．これらのラベルは，Affective Norms for English Words（ANEW）レキシコンに基づき，各曲の覚せい度と感情価を算出し分類したものだが，覚せい度および感情価の数値ラベル自体は含まれていない．そこで本研究では，MoodyLyrics のラベル付けを参考にしつつ，独自に各楽曲の覚せい度および感情価の数値ラベルを生成した．具体的には，歌詞をトークン化し，ANEW に含まれる単語のスコア（覚せい度・感情価）を抽出して合計し，その単語数で

割ることで平均値を算出した．この数値ラベルは 0 から 1（小数第 2 位まで）の範囲に正規化している．

さらに，Transformer 系モデルは入力できるトークン数の制限（最大 512 または 1024 トークン）があるため，歌詞の前処理をおこなった．前処理は，タイトルやアーティスト名などのメタ情報の削除，不要な記号の削除，括弧内の削除，そして「oh oh oh」を「oh」に置き換えるなどの繰り返し表現の簡略化をおこなった．最後に，モデルの学習および評価のために，データセットを学習用 80%，テスト用 20%の割合で分割した．

4.1.2 音響特徴データセット

音響特徴データセットには，PMEmo[7]を使用した．PMEmo は，794 曲の楽曲に対して，覚せい度および感情価の数値ラベルが 0 から 1（小数第 2 位まで）の範囲で付与されている．

回帰モデルの構築には，数値ラベルと対応する音響特徴量を必要とするため音響特徴量の抽出をおこなった．音響特徴量の抽出には，librosa ライブラリ（ver.0.10.1）を使用し，PMEmo で提供されているコーラス部分の音源ファイルを対象とした．特徴量抽出は，およそ 23 ミリ秒（例：512 サンプル，サンプリングレート 22050Hz）の短時間フレームごとの時系列情報を基に計算されるが，本研究ではこれらを統計量（平均，最大値，最小値など）として集約した．この処理により，楽曲単位で固定長のベクトル表現が得られ，楽曲ごとに異なる長さでも，機械学習モデルが扱いやすい特徴量として利用できるようになる．抽出した音響特徴量は，MFCC，テンポおよび基本周波数（Tempo and Fundamental Frequency）などであり，これらを統計量として集約した結果，1 曲あたり合計で 40 次元の音響特徴量ベクトルが得られる．なお，PMEmo において実験に使用可能な音源は 767 曲であり，学習用を 80%，テスト用を 20%の割合で分割した．

4.1.3 統合評価用データセット

統合評価では，Database for Emotional Analysis in Music（DEAM）[8]を使用した．DEAM は 1,802 曲の楽曲を収録しており，各曲には覚せい度および感情価の数値スコアが 1 から 9 の実数値で付与されている．これらの値は，楽曲全体の情動的特性を示すラベルとして用い，0 から 1（小数第 2 位まで）の範囲に正規化した．統合評価には歌詞と音響特徴の両方を必要とするため，各楽曲のタイトルおよびアーティスト情報を基に，Genius API[9]を用いて対応する歌詞を収集した．なお，歌詞および音響特徴に対する前処理および特徴量の抽出は，それぞれ歌詞データセットおよび音響特徴データセットと同様の手順を適用した．また，本研究では英語の楽曲のみを対象とするため，非英語の歌詞が含まれる楽曲は除外した．その結果，楽曲が DEAM に収録されており，対応する英語歌詞が取得可能であった 790 曲を，統合評価用データセットとして使用した．

4.2 評価指標

評価指標には，平均二乗誤差（MSE）を用いた．MSE は，モデルが出力した予測値と実際の正解値との誤差を二乗し，それらをすべてのデータに対して平均した値である．MSE の数式を以下の式(4)に示す．

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (4)$$

ここで，n はデータ数，y_i は正解値，$\hat{y}_i$ はモデルによる予測値を表す．MSE の値が小さいほど，予測値と実際の値との誤差が小さく，予測精度が高いことを示す．

4.3 ハイパーパラメータの最適化

歌詞および音響特徴モデルでは，すべてのモデルにおいて，Optuna ライブラリを用いて最適化をおこなった．ハイパーパラメータは各モデルで 20 回の試行を通じて最適化され，MSE が最小となるハイパーパラメータが選択された．

4.4 比較モデル

本研究では，歌詞および音響特徴それぞれに対して複数の回帰モデルを構築し，予測精度を比較した．歌詞は，BERT[10]，RoBERTa[11]，Decoding-enhanced BERT with disentangled attention（DeBERTa）[12]，および XLNet[13]といった事前学習済み言語モデルを用いて特徴量を抽出し，それらを入力とする回帰モデルを作成した．音響特徴は，Light Gradient Boosting Machine （ LightGBM ） [14] ，Multilayer Perceptron（MLP）[15]，Support Vector Regression（SVR）[16]，eXtreme Gradient Boosting（XGBoost）[17]といった機械学習アルゴリズムを用いて回帰モデルを構築した．

5. 実験結果

5.1 歌詞モデルの予測精度の比較

各歌詞モデルの比較結果を表 1 に示す．実験の結果，BERT が最も優れた結果を示し，覚せい度の MSE が 0.0022，感情価の MSE が 0.0013 となった．この結果より，本研究の歌詞モデルに BERT を採用した．

5.2 音響特徴モデルの予測精度の比較

各音響特徴モデルの比較結果を表 2 に示す．XGBoost が最も優れた結果を示し，覚せい度の MSE が 0.0115，感情価の MSE が 0.0138 となった．この結果より，本研究の音響特徴モデルに XGBoost を採用した．

5.3 歌詞と音響特徴の統合における最適重みの探索

各重みにおける統合結果を表 3 に示す．実験の結果，覚せい度の予測においては歌詞と音響特徴の重みを 1:9 とした場合に，MSE が最も低い 0.0369 を示した．また，感情価の予測においては，重みを 6:4 とした場合に MSE が最も低

表 1　各歌詞モデルにおける MSE の比較結果

モデル	覚せい度	感情価
BERT	**0.0022**	**0.0013**
RoBERTa	0.0036	0.0020
DeBERTa	0.0120	0.0084
XLNet	0.0202	0.0067

表 2　各音響特徴モデルにおける MSE の比較結果

モデル	覚せい度	感情価
LightGBM	0.0116	0.0140
MLP	0.1195	0.0354
SVR	0.0424	0.0139
XGBoost	**0.0115**	**0.0138**

表 3　各重みにおける MSE の比較結果

歌詞:音響特徴	覚せい度	感情価
0:10	0.0376	0.0271
1:9	**0.0369**	0.0245
2:8	0.0370	0.0224
3:7	0.0378	0.0208
4:6	0.0392	0.0197
5:5	0.0413	0.0191
6:4	0.0441	**0.0189**
7:3	0.0476	0.0193
8:2	0.0518	0.0201
9:1	0.0567	0.0215
10:0	0.0623	0.0233

い 0.0189 を示した．これらの結果は，歌詞モデル単独（10:0）および音響特徴モデル単独（0:10）のいずれよりも低い誤差となっており，両者を統合することで単独モデルより高い精度の予測が可能となることを示した．

6. 考察

各歌詞モデルの比較実験では，歌詞に多く見られる比喩的・象徴的な表現や曖昧な言い回しに対し，各モデルがどのように解釈するかが精度に影響を与えた．BERT は，前後の文脈を同時に捉える構造をもつため，こうした複雑な表現の理解に適していると考えられる．また，BERT の覚せい度と感情価の予測誤差が非常に低かった点については，歌詞データセットのラベル付けに用いた手法が影響していると考えられる．数値ラベルを ANEW に基づいて付与しており，これにより数値ラベルの分布が高低いずれかに偏る傾向が生じた．このような偏りにより，モデルが極端な値をより正確に学習しやすくなり，MSE の低下につながったと考えられる．

各音響特徴モデルの比較実験では，統計量としてベクトル化された多様な音響特徴量をいかに効果的に扱えるかが予測精度に大きく影響した．本研究で用いた音響特徴量は，ピッチやリズム，テンポといった異なる性質の情報が混在しており，それらの特徴間に潜む有用なパターンを適切に捉える能力がモデルに求められた．XGBoost は，こうした異なる性質の特徴量に対しても安定して学習できたことから，他のモデルよりも優れた予測精度を発揮したと考えられる．

歌詞モデルと音響特徴モデルの出力を加重平均によって統合した際の各重みにおける比較実験では，覚せい度と感情価が異なる種類の情報に関係していることが，統合による予測精度の向上につながったと考えられる．表 3 の結果から，覚せい度が主に音響的側面に強く関連し，感情価は言語的側面に強く関連していることが示唆される．つまり，どちらか一方の情報のみを用いたモデルでは把握できる情報に偏りが生じ，正確な予測が難しくなる．これに対し，両者を組み合わせることで互いの弱点を補完し，より高精度の感情予測が実現できたと考えられる．また，感情価の MSE は覚せい度と比べて全体的に低く，重み比率に関係なく安定した精度を示した．この傾向は，感情価（快－不快）に関する語彙や文脈的な手がかりが歌詞中に多く含まれており，それらを比較的容易に捉えることができたためだと考える．一方で，覚せい度は音響的側面に強く関係しており，言語的な手がかりが少ないため，予測が相対的に難しかったと考える．さらに，楽曲は時間とともに雰囲気が変化する特性をもつが，本研究では楽曲全体を通じて抽出した音響特徴量の統計量を用いており，時間軸に沿った感情変化は直接的には考慮していない．楽曲内の感情の時間軸に沿った変化を反映させることで，より楽曲の構造的特性を活かした感情予測が可能になると考えられる．

7. おわりに

本研究では，従来の楽曲感情予測の問題点として，感情をあらかじめ定められたカテゴリに分類する手法では感情の強弱が捉えにくい点，歌詞または音響特徴のいずれか一方のみを用いている点の 2 点を挙げ，それらを解決することを目的とした．そのため本研究では，歌詞と音響特徴の両方を用いて，楽曲の覚せい度および感情価を数値で予測する手法を提案した．歌詞モデルには事前学習済みの Transformer 系モデルである BERT を，音響特徴モデルには XGBoost を用いて回帰モデルを構築した．また，両モデルの出力を加重平均で統合する際の，適切な重みの組み合わせを探索した．実験の結果，歌詞と音響特徴の重みにおいて，覚せい度の予測では歌詞と音響特徴の重みを 1:9 とした場合，感情価の予測では 6:4 とした場合に，それぞれ最も高い精度（MSE）を示した．これにより，歌詞と音響特徴を統合することでより高精度な楽曲の感情予測が可能となることを示した．今後は，楽曲の統計量でなく，時間軸に沿った感情変化を考慮した感情予測の検討が必要である．

参考文献

[1] 森 数馬，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-”，Japanese Psychological Review，Vol.57, No.2，215-234 (2014).
[2] Russell, J. A., “A Circumplex Model of Affect”, Journal of Personality and Social Psychology, Vol.39, No.6, 1161–1178 (1980).
[3] Agrawal, Y., et al., “Transformer-based approach towards music emotion recognition from lyrics”, arXiv: 2101.02051 (2021).
[4] Liu, X., et al., “CNN based music emotion classification”, arXiv preprint arXiv:1704.05665 (2017).
[5] Mendes, G. A., and Martins, B., “Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers”, arXiv preprint arXiv:2302.14021 (2023).
[6] Çano, E., and Morisio, M., “MoodyLyrics: A Sentiment Annotated Lyrics Dataset”, In Proc. the International Conference on Intelligent Systems, Metaheuristics and Swarm Intelligence (ISMSI 2017), pp.118–124, ACM, Hong Kong, March (2017).
[7] Zhang, K., et al., “The PMEmo Dataset for Music Emotion Recognition”, In Proc. the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18), pp.135–142, ACM(2018).
[8] Alajanki, A., et al., “Benchmarking music emotion recognition systems”, PLOS ONE, 11(3), e0150930 (2016).
[9] Genius Media Group Inc, 2022, Genius API, (2025 年 5 月 26 日取得, https://genius.com).
[10] Devlin, J., et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv preprint arXiv:1810.04805 (2018).
[11] Liu, Y., et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, arXiv preprint arXiv:1907.11692 (2019).
[12] He, P., et al., “DeBERTa: Decoding-enhanced BERT with Disentangled Attention”, arXiv preprint arXiv:2006.03654 (2020).
[13] Yang, Z., et al., “XLNet: Generalized Autoregressive Pretraining for Language Understanding”, arXiv preprint arXiv:1906.08237 (2019).
[14] Ke, G., et al., “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, In Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, pp. 3149–3157 (2017).
[15] Rumelhart, D. E., et al., “Learning representations by back-propagating errors”, Nature, Vol.323, No.6088, pp.533–536 (1986).
[16] Smola, A. J., and Schölkopf, B., “A Tutorial on Support Vector Regression”, Statistics and Computing, Vol.14, No.3, pp.199–222 (2004).
[17] Chen, T., and Guestrin, C., “XGBoost: A Scalable Tree Boosting System”, Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’16), pp. 785–794 (2016).