E-039

CLAP ベクトル特徴量と音響特徴量による感性語の対応分析

Correspondence analysis of sensible words using CLAP features and acoustic features.

飯島日菜‡　　市川治‡
IijimaHina　　IchikawaOsamu

1. はじめに

音楽の印象を「明るい」「暗い」といった感性語で語ることは日常的であるが,「明るいけど暗い」というような複雑な印象を持つ楽曲に対する分析は困難である.本研究では,このような曖昧で複雑な感性語が,テンポやモードといった具体的な音楽特徴とどのような関係があるかを対応分析する.楽曲がどのような感性語に対応するかは,テキストと音響特徴の共通の潜在表現である CLAP（ContrastiveLanguage-AudioPretraining）のベクトル空間表現を用いる.

また,「明るいけど暗い」という複合的な感性語を CLAP が表現できているかを調べるため,,VAE（VariationalAutoencoder）,MIRToolbox,CLAP の 3 手法で楽曲推薦についての性能比較を行う.

2. CLAP による感性語の対応分析

本研究では,CLAP の音声・テキスト共通ベクトル空間において,感性語と音響特徴の関係を分析するために,楽曲ごとに,テンポやモードといった外形的な音響特徴と,その曲がどの感性語に対応しているかというクロス集計を行った.モードとは,楽曲が長調（メジャー）または短調（マイナー）のいずれに基づいて構成されているかを示す音楽的特徴であり,楽曲の明るさや感情的傾向に関係する重要な指標である.

使用する感性語は,SongDescriberDataset のキャプションに頻出する上位 75 の形容詞（adjective）を抽出し,それぞれの単語に対応する CLAP の特徴ベクトルと,その楽曲の音響特徴である CLAP の特徴ベクトルの類似度スコアを算出した.スコアの算出にはコサイン類似度を用いており,負の値はゼロにフロアリングしている.したがって,スコアの値は 0 から 1 の範囲の連続値をとる.クロス集計にて感性語のカウントを行う際には,これをそのまま,例えば,0.7 曲分といった寄与度で集計する.

2.1 分類基準

MIRToolbox により,外形的な音響特徴を算出する.このライブラリは,RMSenergy（音量平均値）, Tempo（テンポ平均値）, Brightness（1500Hz 以上の音が占めるパワーの比率）, Spectral irregularity（音質の変化の大きさ）, Inharmonicity（根音に従っていない音の量）, Mode（major と minor の音量の差）といった多種の特徴を算出することができるが,今回は,テンポとモードのみに着目した.

† 滋賀大学　Graduate School of Data Science
‡ 滋賀大学　Faculy of Data Science

対応分析を行うために,以下のような分類を行った.

テンポ：
- スローテンポ（90-120BPM）
- ミドルテンポ（120-140BPM）
- ハイテンポ（140-180BPM）

モード：
- マイナー傾向（mode＜−0.1）
- メジャー傾向（mode＞0.1）
- 中性的（−0.1≦mode≦0.1）

2.2 クロス集計と対応分析

各分類カテゴリ（テンポ 3 種・モード 3 種）ごとに,そこに属する楽曲と各感性語の類似度スコアを算出し,外形的な音響特徴（テンポ,モード）×感性語のクロス集計表を作成した.これらの集計表に対して対応分析（CorrespondenceAnalysis）を適用し,カテゴリ（テンポまたはモード）と感性語の関係性を 2 次元空間にマッピングした（図 1）.

特に「明るい」「暗い」といった基本的な感性語と意味的に近い形容詞に注目し,それらを重点的に可視化することで,複雑な印象を持つ楽曲の位置づけや傾向の把握を行った.

「明るい」語群：high,bright,positive,uplifting,joyful,lively
「暗い」語群：sad,melancholic,dark

対応分析結果を 3 ページ目下部の図 1 に示す.
図から,テンポやモードのカテゴリと感性語の意味的分布には多少の傾向が認められた.

特にメジャー調やハイテンポは,positive,joyful,bright といったポジティブな語と近く,

一方でミドルテンポやマイナー調は melancholic や dark といったネガティブ傾向に関連していた.

中性的なカテゴリは感性語空間の中心付近に位置しており,強い印象の偏りを示さないことが分かる.

3. CLAP による「明るいけど暗い曲」を自動的に抽出・分類する性能の従来手法との比較

次に,「明るいけど暗い」という複合的な感性語が,CLAP などの潜在表現で捕捉することができるのかを調べるために,複数の手法を用いて類似度ベースの楽曲推薦の実験を行った.

3.1 使用するデータ

各フリーBGM 投稿サイト[1][2][3]からダウンロードした 580 曲を用いた.1 曲の長さは平均 162 秒である.これらから

「明るいけど暗い」に該当する曲を 13 曲選択した.これらは被験者 1 名により主観的に選ばれた正解曲群となる.

3.2 比較のための従来法

新たに提案する CLAP を用いた楽曲特徴抽出手法の有効性を検証するために,2 つの従来手法を比較対象として用いた.1 つは MIRtoolbox を用いた音響特徴量ベースの手法,

もう 1 つは VAE（VariationalAutoencoder）を用いた潜在表現ベースの手法である.

3.2.1 MIRtoolbox を用いた音響特徴量ベースの手法

従来研究においては,MIRtoolbox[4]というライブラリを用いて,音響的な特徴量を抽出する例が多い.文献[5]においては,以下の 6 つの特徴量が使用されている：RMSenergy（音量の平均値）,Tempo（テンポの平均値）,Brightness（1500Hz 以上の周波数が占めるパワー比率）,Spectralirregularity（音質の変化の大きさ）,Inharmonicity（基音からずれた成分の量）,および Mode（メジャーとマイナーの音量差）である.本研究では,この手法を再現し,これらの特徴量を用いて楽曲を 5 次元のベクトルに変換し,キー曲と候補曲のペア間のコサイン類似度を算出することで,楽曲推薦を行った.

3.2.2 VAE（VariationalAutoencoder）を用いた教師なし学習

2 つ目の従来手法として,教師なしの深層学習モデルである VAE を用いた楽曲特徴抽出を行った.この手法では,まず手持ちの楽曲データを用いて VAE を学習する.モデルの入力および出力は,4 小節分のピアノロールデータであり,鍵盤方向に 48 次元,時間方向に 64 次元（16 分音符単位）を持つ.ピアノロールデータは 0 と 1 で構成されるバイナリ画像と同様の構造を持つため,CNN（畳み込みニューラルネットワーク）を用いて潜在空間にマッピングされる.

潜在空間は 16 次元の多次元正規分布に近づくように設計されており,類似する楽曲が近接した表現を持つように学習される.なお,波形データからピアノロールへの変換には,piano_transcription_inference というライブラリを用いて MIDI データへの変換を行い,長調は C,短調は Am へと移調した上で,曲の先頭から 4 小節ずつを 2 小節ごとにスライドしながら抽出した.この単位を「素片」と呼ぶ.

学習済みの VAE エンコーダを用いることで,各素片を 16 次元の潜在特徴量に変換する.類似度の算出においては,キー曲と候補曲の素片それぞれについてコサイン類似度を計算し,候補曲中で最も高いスコアを採用する.これを全ての素片について繰り返し,最終的な類似度スコアはその平均値とした.

4. CLAP を用いた提案法

提案法として,対象楽曲と複数の比較対象楽曲との音響的類似度を評価するために,CLAP（ContrastiveLanguage-AudioPretraining）モデルを用いた埋め込みベースの手法を実装した.CLAP は音声およびテキストを同一の意味空間にマッピング可能なマルチモーダルな深層学習モデルであり,音楽・環境音などを含む多様な音声信号に対して,高次元の意味的特徴ベクトル（埋め込み）を生成する能力を持つ.

具体的には,まず対象となる 1 曲の音声ファイルを入力とし,librosa を用いて 48kHz にリサンプリングした後,CLAP の音声プロセッサおよびエンコーダを用いてその音響特徴をベクトルとして抽出した.次に,比較対象とする複数のフォルダ内に含まれる音声ファイル群（計 580 曲）に対しても同様の処理を行い,それぞれの埋め込みベクトルを取得した.これらのベクトルと対象曲のベクトルとの間でコサイン類似度を算出することで,楽曲推薦を行った.

5. 評価実験

キー曲との類似度の計算をすべての候補曲について実行し，類似度の上位のものから並べる．正解の曲とされる別の「明るいけど暗い」曲が，何位に入っているかをみることで，推薦性能を評価することができる．評価の尺度として，MRR(MeanReciprocalRank)と Recall@k を用いる MRR は上位のものから眺めていって正解が表れたときに，その順位の逆数で重みづけした合計値である．Recall@k は，上位 k 個の予測に含まれる正解数を,総正解数で割った値である.今回は k=100 を使用した．評価実験では提案法をキー曲を変えて 4 回行った.

実験結果を表 1 ,2 に示す.

MRR	ハロウィンがやって来る!!	Peritune_Nook_Nights	PerituneMaterial_Spook3	n48	平均
MATLAB	0.042	0.042	0.042	0.042	0.0423
VAE	0	0.015	0.027	0.016	0.0145
CLAP	0.006	0.111	0.088	0.017	0.0561

表 1　MRR での評価結果

Recall@k	ハロウィンがやって来る!!	Peritune_Nook_Nights	PerituneMaterial_Spook3	n48	平均
MATLAB	0.153	0.461	0.307	0.461	0.346
VAE	0.076	0.384	0.076	0.153	0.173
CLAP	0.006	0.538	0.692	0.538	0.480

表 2　Recall@k での評価結果

表 1 および表 2 に示されるように,各手法は平均値において一定の性能を示したが,個々の楽曲におけるスコアには大きなばらつきが見られた.

表 1 の MRR 評価結果によれば,MATLAB および VAE は比較的安定した性能を示したのに対し,CLAP は一部の楽曲で高いスコアを記録しつつも,他の楽曲では低いスコアを示した.

表 2 の Recall@k の評価結果においては,「ハロウィンがやって来る!!」のみ精度が低かった.

キー曲の MIRtoolbox での抽出結果を表 3 に示す.

	RMS	Tempo	brightnes	regularity	Inharmonicity	mode
PerituneMaterial_Spook3	0.202	141.426	0.529	0.744	0.498	-0.221
Peritune_Nook_Nights	0.306	117.463	0.339	0.873	0.463	-0.227
ハロウィンがやって来る!!	0.094	138.582	0.325	0.713	0.470	-0.345
n48	0.137	133.429	0.334	0.592	0.438	0.137

表 3　キー曲の MIRtoolbox での抽出結果

精度が低かった「ハロウィンがやって来る!!」「n48」の 2 曲は RMSenergy（音量の平均値）が低かった.曲を聴いてもイントロが静かに始まっていたり,全体的に大きな曲調の変化の少ない曲であったりした.そのことから,音量の平均を統一することで精度が上がる可能性がある.

6. おわりに

CLAP を用いて,楽曲と感性語の関係性を分析し,音響的・意味的な特徴を捉える手法の可能性を探った.テンポ・モードと感性語の対応分析,そして「明るいけど暗い曲」の自動抽出に取り組んだ. 対応分析の結果,ハイテンポ・メジャー調はポジティブな感性語と,ミドルテンポ・マイナー調はネガティブな感性語との関連が示唆された.また,CLAP による推薦システムでは,両評価指標において従来手法を上回る精度を記録する場面が見られ,意味的な楽曲分類における CLAP の有効性が一部確認された.しかし,どちらもすべての楽曲には有用ではないという課題が見受けられる.そのため,精度が低かった楽曲の特徴を調べ,感性に基づく楽曲推薦の精度向上と汎用性の拡大を目指したい.

謝辞

本研究は JSPS 科研費 25K15402 の助成を受けた.

参考文献

[1] http://www.hmix.net
[2] https://hiiragimusic.com/
[3] https://peritune.com/
[4] O.Lartillot,et.al“AMATLABTOOLBOXFORMUSICALFEATUREEXTRACTIONFROMAUDIO”,DAFx-07,2007.
[5] 上原等,“コード進行・メタ情報・楽曲特徴量に基づく音楽可視化”,WISS2015.

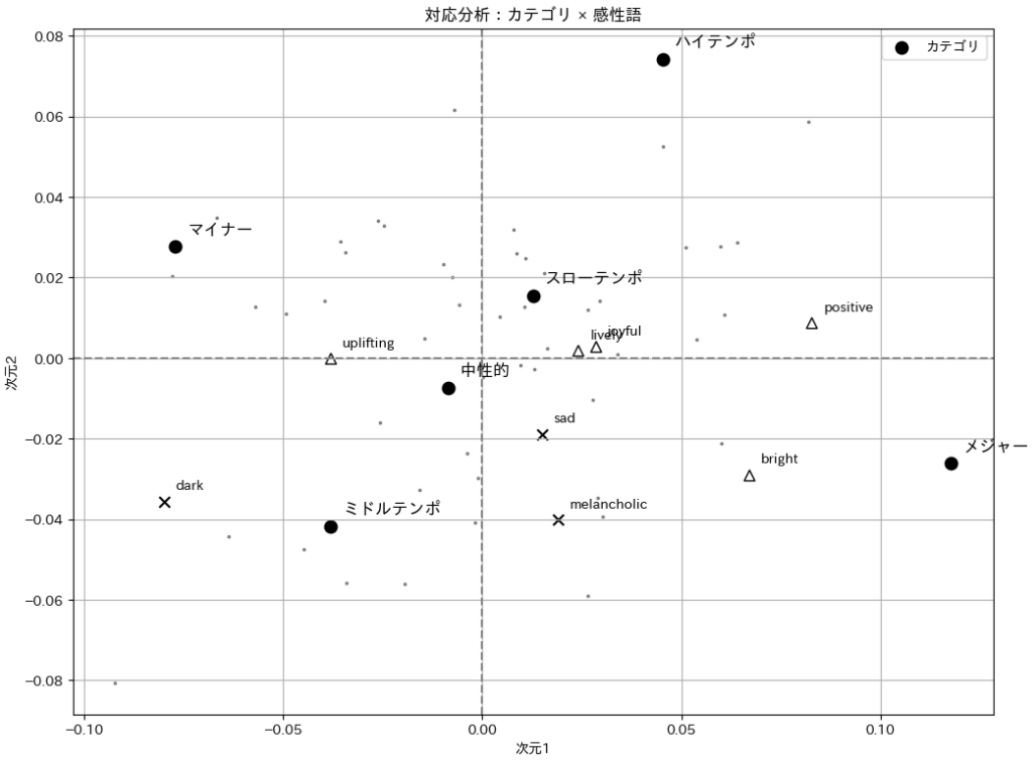

図 1　対応分析結果