GMM に基づいた楽曲特徴と感性情報の対応関係のモデル化

西尾 圭一郎 †，酒向 慎司 †，北村 正 †
† 名古屋工業大学大学院工学研究科

1 はじめに

ある楽曲の持つ特徴と，人がそれを聞いたときに抱く印象には何らかの関係がある．この関係を表現したモデル（感性モデル）の利用例として，個人の好みや感性による音楽検索が挙げられる．本研究では，音響信号の物理的な特徴（物理量）からそれに対する印象（心理量）を推定することを主な目的とし，この両者の対応関係をモデル化する．

先行研究には，ニューラルネットワークによる感性モデル [1] が提案され，物理量と心理量の複雑な対応関係のモデル化を可能としていた．本研究では，類似した特徴を持つ曲を聞いた際に抱く印象の差異が確率的な振る舞いをすると仮定し，その対応関係から印象を推定する．その手法として GMM によるパラメータ変換に基づいて楽曲の特徴と印象の対応関係のモデル化を試みた [2]．しかし，未知データの推定においてまだ十分な結果が得られていない．原因として，従来の特徴量は音高や強度のみを表現しており，音楽の速度に関する印象を考慮できていないため，印象に関係する特徴を表現しきれていないことが考えられる．そこで本論文では，新規に速度に関する特徴であるリズムの特徴を追加した物理量でのモデル化を行い，実験により有効性を確認する．

2 感性モデル

一般的に楽曲が異なればそれに対して抱く印象には差異が生じるが，印象が類似していればそれらの楽曲は何らかの音響的な特徴が類似していると考えられる．その音響的な特徴として，楽器構成やリズムといったものが挙げられ，統計的にみることで音響的な特徴と印象の間に大きな相関があると考えられる．そこで，楽曲間の印象の差異が確率的に振る舞うと仮定し，双方の相関を確率的に評価することで，楽曲の特徴と印象の対応関係を表現する．

特徴と印象を数値的に扱うために，音響信号の短時間分析により物理量を抽出し，評価実験により心理量を測定する．そして，これらの対応関係を確率モデルによって学習的に獲得できる GMM に基づいてモデル化し，それによるパラメータ変換を印象推定に適用する．

2.1 GMM に基づく特徴量間関係のモデル化

楽曲の特徴と印象には相関があり，それらをベクトル表現した物理特徴と心理特徴にも相関があると言えるため，この相関を印象推定に使用することを考える．しかし，両者の特徴の性質や，多次元データであることから，明確に対応付けることは困難かつ非効率である．そこで高精度な声質変換手法の一つとして注目されている GMM に基づくパラメータ変換に着目した [3]．この手法は，相関のある特徴量間の変換モデルとして一般化することができ，多量のデータから特徴量間の対応関係を最尤基準で学習し，それに基づくパラメータ変換を可能とする．本研究ではこれを物理量と心理量の変換モデルに応用する．

以下に基本的な GMM の学習，パラメータ変換手順を示す．楽曲の物理量，心理量をそれぞれ $\boldsymbol{X}_n$，$\boldsymbol{Y}_n$ とし，これらを結合した $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_N]$，$\boldsymbol{Z}_n = [\boldsymbol{X}_n, \boldsymbol{Y}_n]$ を学習データとして GMM でモデル化をする．図 1 に GMM に基づくパラメータ変換の概略図を示す．

$$P(\boldsymbol{Z}|\boldsymbol{\lambda}) = \prod_{n=1}^{N} \sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{Z}_n|\boldsymbol{\mu}_i^{(Z)}, \boldsymbol{\Sigma}_i^{(Z)})$$

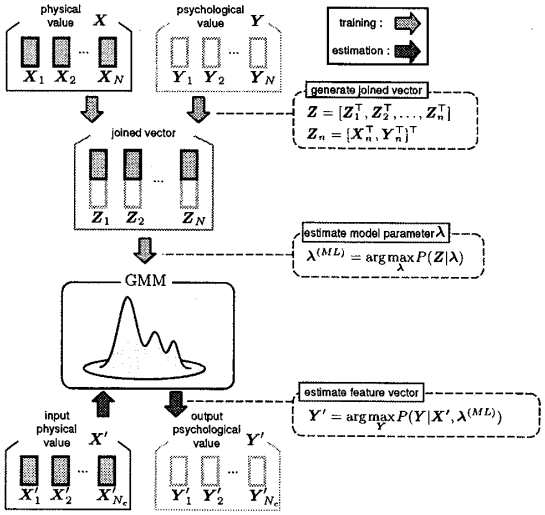

図 1: パラメータ変換の概略図

ただし，w_i は混合重み，$\boldsymbol{\mu}_i^{(Z)}$ は平均ベクトル，$\boldsymbol{\Sigma}_i^{(Z)}$ は共分散行列，M は混合数である．これにより楽曲と印象の関係が完全データの確率密度分布として表現される．

モデルパラメータの推定を ML 基準に基づいて行い，学習データ $\boldsymbol{Z}$ の尤度を最大にするモデルパラメータを $\boldsymbol{\lambda}^{(ML)}$ とする．このようにして推定された心理モデル $\boldsymbol{\lambda}^{(ML)}$ を用いて，ある楽曲の物理量 $\boldsymbol{X}'$ に対応する最尤の心理特徴ベクトル $\boldsymbol{Y}'$ は以下のように表される．

$$\boldsymbol{Y}' = \arg\max_{\boldsymbol{Y}} P(\boldsymbol{Y}|\boldsymbol{X}', \boldsymbol{\lambda}^{(ML)})$$

このようにして，様々な楽曲データに対応した物理量と心理量の結合データから学習された心理モデルに基づいて，ある楽曲の物理量から未知の心理量を推定する．

2.2 物理量抽出

先行研究 [2] で，曲の特徴を決める主な要素として，メロディ，曲の盛り上がり，楽器の種類などがあり，それらを表す特徴量として，表 1 の条件で全体エネルギー，中心周波数，周波数帯域幅，低周波数成分の割合を求めた．次に，楽曲全体で共通の特徴空間を定めるために，フレーム毎に求めた各特徴量の値の分布状況をヒストグラムで表現した．特徴量毎に分割領域を定め，各領域に属するフレームの数を算出する．特徴量毎に求めたヒストグラムを結合することで 64 次元の物理量とした [2]．さらに新規に速度の印象に対応する特徴としてリズムの特徴量を追加する．

今回は，楽曲のパワーの周期性がリズムに対応すると考え特徴を抽出する [4]．図 2 に抽出手順と以下にその説明を示す．

1. 音響信号を離散ウェーブレット変換し複数の周波数帯に分割
2. 各周波数帯でパワー包絡線抽出
3. 全ての包絡線を合計し楽曲の包絡線を求める
4. 包絡線の自己相関関数 (R) を求めピークを検出

Modeling of the Relationship between Feature of Audio Signal and Emotional Information Based on GMM
Keiichiro Nishio †, Shinji Sako †, Tadashi Kitamura †, † Graduate School of Engineering, Nagoya Institute of Technology

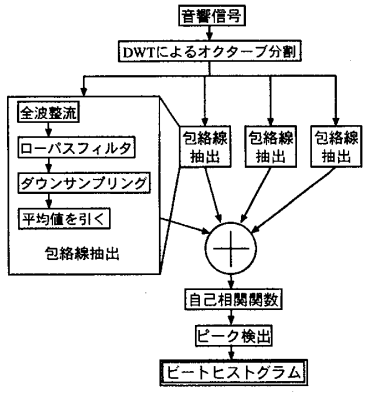

図 2: ビートヒストグラム抽出手順

5. 階級を BPM，度数を R のピークの値としたヒストグラムを作成

以上の手順で抽出したビートヒストグラムをリズム特徴量とする．この際，DWT はフレーム長を 3 秒，シフト長を 1.5 秒の条件で行う．そして各フレーム毎に自己相関関数を求め，検出された複数のピークの値をビートヒストグラムの対応する BPM の階級に足し込むことで楽曲全体のリズムの特徴を表現する．また，ピークは 60〜200BPM に対応する部分に対してのみ抽出する．また，階級を 10BPM 刻みとすることで，リズム特徴量を 14 次元で表す．これを従来の 64 次元の特徴量と結合することで物理量は 78 次元となる．

2.3 心理量測定

楽曲の印象を数値的に得るために聴取実験を行った．今回は複数の形容詞対を尺度とした 7 段階 SD 法を用い，その評価値を心理量として使用した．

実験に用いる評価語として，先行研究 [1]，音色の評価・測定法，音楽と楽器の音響測定を参考に，本研究の対象音響であるクラシック音楽の評価に適していると考えられる形容詞対を 14 種類選出した．それらを表 2 に示す．被験者 20 名それぞれに 60 曲の楽曲を聞かせ，その楽曲に当てはまる尺度の度合を形容詞毎に選択させ，1 曲につき 10 名分の評価データを得た．最後に，評価値を形容詞対毎に平均することで，各楽曲の心理量を 14 次元のベクトルとして表現する．

3 実験

3.1 モデル作成及び評価方法

1 曲を 92 次元 (物理量部: 78 次元，心理量部: 14 次元) のベクトルで表現し，120 曲の内 100 曲を学習データとしてモデルを作成した．この際，混合数と学習データの組合せを変更し，複数のモデルを作成した．また，残りの 20 曲を未知データとして使用し推定実験を行った．

評価方法として，全ての推定値に対して実際の心理量とのユークリッド距離を調べ，推定値が本来の心理量にどれだけ近い値を推定できているかについて評価を行った．

まず，ある推定値に対して全曲の心理量とのユークリッド距離を計算し，距離の小さい順に順位付けを行う．そして，本来の心理量との距離の順位を評価する．つまり順位が高ければ，近い値を推定できていることとなる．今回は本来の心理量との距離の近さが 1 位に評価された曲数の割合，また 3 位以内に評価された曲数の割合を調査した．

3.2 実験結果

図 3 に各推定値が本来の心理量と最も近い値で推定された曲数の割合を示し，図 4 に心理量との距離が 3 位以内に評価された曲数の割合を示す．また，学習データの推定はほぼ 100%の推定結果を得られているので省略する．先行研究 [2] の結果を赤色で示し，今回の結果を青色で示す．

この結果から，オープンデータの推定において 1 位に評価された曲数の割合が向上していることが確認できる．特に混合数 10 においては，23%の曲が 1 位に評価され，先行研究 [2] の最高値を上回る結果を得られた．また 3 位以内に評価された曲数の割合に関しては，最高値は下回るが前回とほぼ同程度の約 4 割の精度となった．

4 むすび

本研究では，音響信号と印象の対応関係を GMM によるパラメータ変換に基づいてモデル化を行った．その際，音響信号の物理量として，新規にリズムの特徴を用いることで速さに関する印象に対応する特徴も考慮した物理量を使用した．

実験結果から，オープンデータに関しては 1 位に評価された曲数の割合は，先行研究 [2] の精度の最高値を 1.5%上回る結果が得られたが，3 位以内に評価された曲数の割合では改善は見られなかった．今回の結果からは大きな有効性を確認できず，十分な結果を得られたとは言えない．原因として，データ不足が考えられる．また，心理量として評価値の平均データを用いているため，感性の個人性を考慮できていないことも原因として挙げられる．

今後の課題として，より多くの学習データを収集する事，感性の個人性を考慮したモデル化手法の検討が挙げられる．

表 1: 物理量の分析条件

データベース	RWC-MDB
データ形式	RIFF-WAVE(ステレオ)
サンプリング周波数	44.1 kHz
データ長	10 sec
フレーム長・フレーム周期	44.66 msec・20 msec
窓関数	ブラックマン窓

表 2: 使用した感性語一覧

軽い	–	重い	明るい	–	暗い
しんみりした	–	うきうきした	迫力のある	–	静かな
穏やかな	–	激しい	陰気な	–	陽気な
のびやかな	–	抑えたような	速い	–	遅い
優雅な	–	荒々しい	寂しい	–	賑やかな
慌しい	–	のんびりとした	重厚な	–	軽快な
安らぐ	–	緊張した	華やかな	–	素朴な

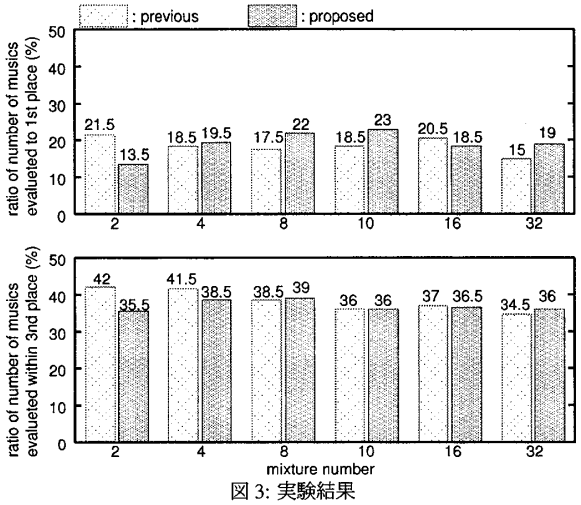

図 3: 実験結果

参考文献

[1] 平江 遼，西 隆司：「感性に基づくクラシック音楽の分類」，日本音響学会誌 64 巻 10 号，pp.607–615 (2008–8)

[2] 西尾 圭一郎，酒向 慎司，北村 正：「クラシック音楽を対象とした GMM に基づく感性モデルに関する研究」，電子情報通信学会 2009 年総合大会講演論文集，A-15-2 (2009–3)

[3] 戸田 智基：「最尤特徴量間変換法とその応用」，電子情報通信学会技術研究報告，105，571，SP2005–147，pp.49–54 (2006–1)

[4] G. Tzanetakis, P. Cook：「Musical Genre Classification of Audio Signals」，IEEE Transactions on Speech and Audio Processing, vol.10, No.5, (2002–7)