Vol.2024-MUS-139 No.15
2024/3/9

複数人による音楽アノテーション結果の項目反応理論に基づく統合と機械学習への応用

中野 倫靖[1,a)]　後藤 真孝[1,b)]

概要：ある楽曲やその一部分に対して、人間によるアノテーション（ラベル・タグ付け、評価）を行うことは、音楽情報処理研究において重要なタスクの一つである。例えば、セマンティックタグを付与したり、歌唱力の評価結果を付与したりすることで、それらを自動推定する機械学習モデルの学習データに用いることができる。しかし、「音楽に対する人間によるアノテーション」という点においては、アノテータの能力や主観に影響され、判断に曖昧性があることから、複数人でアノテーションした結果を統合して扱うことが多い。本稿では、そのような複数人によるアノテーションとしてセマンティックタグの付与と、リッカート尺度による段階的評価のような離散的なアノテーションを対象とし、それらの新しい統合方法と機械学習への応用を議論する。このような離散ラベルの統合においては、従来、多数決による決定が多く用いられてきたが、我々は項目反応理論 (IRT) 及び段階反応モデル (GRM) を用いた統合を提案する。提案手法は、個々のアノテータの特性を考慮しながらラベルを統合できる点で新しく、アノテータ数が偶数であったり評価値が順序尺度であったりしても統合を行える特長がある。

1. はじめに

音楽情報科学研究において、音楽に対するアノテーションは重要である。例えば、楽曲構造、ビート時刻、感情、音楽ジャンル、音素系列、テンポ、F0、歌唱力、好み、等がある。これらアノテーションの結果は、深層学習等の機械学習における学習データや、構築した音楽情報処理モデル・システム評価のために用いられるだけではなく、心理実験などにおける楽曲の特性分析でも重要な役割を果たす。

しかし、音楽や歌声へのアノテーション自体の多様性・曖昧性や、アノテータの主観・経験・能力・状況の違いなどが原因となり、そのアノテーション結果は一つに定まらないことが多い。実際、音楽アノテーション結果の合意の度合いを表す Krippendorff's α は、ほとんどの場合、それが 1（完全な合意）となることはない。例えば、Kim *et al.* [1] による、歌声を含む楽曲に対する歌声セマンティックタグの付与においては、アノテータ 3 名による結果、その多くのタグ付けの Krippendorff's α が 0.4 から 0.75 の範囲となった。また、Bogdanov *et al.* [2] は、アノテータ 3 名が同一の楽曲セットに対して valence（感情価）と arousal（覚醒度）の相対アノテーションを実施し、Krippendorff's α は valence で 0.39、arousal で 0.48 であった。つまりこのような、人によって異なるアノテーション結果をどのように扱うかは、研究を進める上で重要な課題である。

従来、同一楽曲に対する複数人によるアノテーションとしては、前述した歌声タグ [1] と相対的な感情値 [2] の他にも、楽曲構造 [3, 4]、ビート [5]、音楽セマンティックタグ [6]、歌唱力 [7]、コンセプト [8]、valence と arousal（連続値）[9]、等がある。これらは、複数人のアノテーション結果を統合して扱う方針 [1, 2, 6, 7, 9] 以外にも、異なる複数の結果が正解となりうることも議論されてきた [10, 11] 背景から、複数人の結果をそのまま保持しておく方針 [4] もあった。

本研究では、機械学習への応用を前提として、複数のアノテーションの結果を、一つの学習（正解）データとして統合することを考える。このような統合方法としては、従来、多数決に基づく方法 [1, 2, 6]、タグ付けしたアノテータ数 [1]、平均値 [9]、Best-Worst Scaling [7] が用いられてきた。また、ゲーム化による二人以上の同意によるラベル付け [12–14] も提案されてきた。しかし、我々の知る限り、これらではアノテータの特性やアノテーション結果の揺らぎを考慮した統合は検討されてこなかった。例えば、アノテータによってタグ付けの基準が異なる場合、それを考慮できるとよりよい統合となる可能性がある。また、多数決

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
[a)] t.nakano@aist.go.jp
[b)] m.goto@aist.go.jp

による統合では 2 値化により情報が失われるため、連続値として扱ったり、その値の揺らぎを考慮して扱ったりできる統合手法の方が望ましい可能性がある。

そこで本稿では、個々のアノテータの特性を考慮でき、かつ離散的な値のアノテーションを潜在的な連続値として推定できる項目反応理論（Item Response Theory, IRT）[15] の応用を検討する。また、リッカート尺度による音楽の評価は応用可能性が高いと考えられるが、アノテーションにおいて使われたり、それを統合する方法が示されたりすることはなかった[*1]。項目反応理論では、それを順序関係のある離散値を扱えるように発展させた段階反応モデル（Graded Response Model, GRM）[16] によりリッカート尺度に基づく値を扱うことができ、新たなアノテーションの統合方法として検討することは意義がある。

2. 関連研究

音楽アノテーション結果の統合と IRT の応用に関する事例として、本稿と特に関連する研究に絞って説明する。

2.1 音楽アノテーションにおける統合

前章で述べたように、これまで音楽アノテーションにおいて、同一曲に複数人がアノテーションした事例は多い。音楽ジャンル分類 [17,18]、音楽感情認識 [2,19,20]、楽曲間類似度 [21]、コード認識 [22]、セマンティックタグ [1,6]、に関して、アノテータの合意に関する研究がなされており、その度合いを表現するために Krippendorff's α が使われることが多い [1, 2, 19, 20, 22]。しかし、Krippendorff's α はアノテーション結果の分析に使われるのみであり、統合においては多数決に基づく方法 [1,2,6] が用いられてきた。また、実際に学習データとしては用いられなかったものの、タグをつけたアノテータ数を分析のために示した例がある [1]。

Gupta *et al.* [7] は、クラウドワーカによって歌唱力を順位付けさせ、その最上位と最下位の結果を同一尺度上に配置する Best-Worst Scaling [23] を用いて統合した。ここで音楽に限らず、クラウドソーシングに基づくアノテーションにおいては、クラウドワーカの信頼性を推定しながら集約する方針が取られる [24] ことがあり、Learning from crowds と呼ばれることもある [25]。

以上、学習データの統合に対して、IRT に類似した手法が取られることはあったが [24]、音楽アノテーションにおいて用いられたことはなかった。

2.2 項目反応理論を用いた応用

川島ら [26] は、IRT を短歌の評価に応用し、その能力値 θ を 1 次元ではなく 2 次元と多次元化することで、単純な優劣以外の分析を可能とする手法を提案した。大谷ら [27] は、IRT の拡張として、Samejima *et al.* [16] による段階反応モデル（GRM）を活用し、翻訳システムの優劣を評価する枠組みを提案した。ベースラインとなるシステムの翻訳結果とそれ以外のシステムとの結果を相対比較したものを 3 段階の順序データとして扱って、翻訳システムの能力値を推定して比較した。また、Sharma *et al.* [28] は、機械学習モデルの性能分析に IRT を用いた。

以上、IRT の結果を分析に用いられたことはあったが、その結果を機械学習の学習データとして用いる検討がなされたことはなかった。

表 1 テストにおける IRT の例。ここでは行が受験者 s、列が問題 q を意味し、行列の値 1 は正答、0 は誤答を意味する。a, b, θ は IRT モデルパラメータの推定値。

	q_1	q_2	q_3	q_4	θ
s_1	1	0	1	0	0.031
s_2	1	1	1	1	0.905
s_3	1	0	0	1	0.056
s_4	0	0	0	0	−0.976
s_5	1	1	0	0	0.041
s_6	1	0	1	0	0.019
a	0.982	0.762	0.681	0.76	
b	−0.791	0.33	0	0.365	

3. 項目反応理論 (IRT) と段階反応モデル (GRM) の音楽アノテーションへの適用

項目反応理論（Item Response Theory, IRT）[15] は、テストや評価のための数理モデリング手法であり、複数の「項目」に対する複数の「反応」をモデル化する。例えば、受験やテストにおいては、項目はテスト問題、反応は受験者の回答に対応する（**表 1**）。受験者の能力を表す潜在変数 θ と、項目の特性を表すパラメータ a, b（識別力 a と難易度 b）の関係を確率モデルで定義することで、テストの信頼性を向上させたり、受験者の能力を正確に推定させたりすることが可能となる。例えば、同じ点数を取った受験者であっても、回答した問題の難易度によって能力に差をつけることができる。

3.1 音楽タグ付けの項目反応モデル（二値反応データ）

二値反応データに対する項目反応モデルは、受験者 s_i に潜在的な能力 θ_i を導入し、能力 θ_i の受験者が問題 j に正解する確率を以下のような項目反応関数で表現する。

$$p_{ij}(\theta_i) = \frac{1}{1 + \exp(-a_j(\theta_i - b_j))} \qquad (1)$$

ここでは、項目反応関数をロジスティック関数で表現する 2 パラメータ・ロジスティックモデル (2PLM) とした。式 (1) において b_j は、その値よりも能力 θ_i が高い場合に正

[*1] ただし、順序関係のある数値を、離散的な選択肢 (−1.0〜1.0 の範囲で 11 段階の値) の中から複数人が選択し、それを平均して用いた事例 [9] はある。

答できることから難易度と呼ばれる。一方、a_j はロジスティック関数の傾きであり、これが高いと b_j 前後での回答の正誤が区別しやすくなることから識別力と呼ばれる。

表 1 の例では、すべての問題に正答した受験者 s_2 の能力値 θ_2 の値が最も高いことや、逆にすべての問題に誤答した受験者 s_4 の能力値 θ_4 の値が最も低いことが分かる。また、難易度 b について、6 名中 5 名が正答した q_1 の値が低く、正答者数が少ない q_2, q_4 の値が高い。また s_1 は、自身の能力値 $\theta_1 = 0.031$ を超える難易度の q_2, q_4 に正解していない。

本稿ではこのような IRT モデルを、受験者 s を楽曲、問題 q をアノテータとするタグ付けタスクに適用することを考える。多数決と異なり、アノテータ数が奇数でなくとも、そのタグが付与される潜在的な度合い θ を表現できる。つまりこの例では、θ_2 が高い楽曲 s_2 にはこのタグが付与される可能性が高い。また、β_1 が低いアノテータ q_1 は多くの楽曲にこのタグをつける、つまり、場合によってはあまり信頼できない（精度の低い）可能性があることが分かる。

3.2 Likert 評価の段階反応モデル（段階反応データ）

順序尺度を持つ多値型の反応モデルとして、段階反応モデル（Graded Response Model, GRM）[16] を用いる。GRM は、一つの項目に対して 2 つ以上の順序のある反応を扱うことができる点で、2PLM を一般化したモデルと見なすことができる。つまり、7 段階 Likert 尺度や、「A が高い」「どちらも同じ」「B が高い」のような相対的な比較評価などの、順序関係を持った反応データをモデル化できる。前者では反応カテゴリ数 $K = 7$、後者では $K = 3$ として、ある項目 u_{ij} がカテゴリ $k \in 1, ..., K$ と反応する確率 $p_{ij}(u_{ij} = k|\theta_i)$ を、$k-1$ と k とにそれぞれ反応する確率の差として、以下のようにモデル化する。

$$p^*_k(\theta_i) = \frac{1}{1 + \exp(-a_j(\theta_i - b_{j,k}))} \tag{2}$$

$$p_{ij}(u_{ij} = k|\theta_i) = p^*_{k-1}(\theta_i) - p^*_k(\theta_i) \tag{3}$$

ここで、k はカテゴリの順番を意味する。$b_{j,k}$ は項目 j における k より大きいカテゴリに反応する難易度を表し、その個数は $K-1$ となる。ただし、$p^*_0(\theta_i) = 1$、$p^*_K(\theta_i) = 0$ である。式 (3) を、$\eta = a_j\theta_i$, $c_k = a_j b_{j,k}$ として、ロジスティック関数を逆 logit 関数で記述すると以下となる。

$$p_{ij}(u_{ij} = k|\theta_i) = \begin{cases} 1 - \text{logit}^{-1}(\eta - c_{k+1}), & \text{if } k = 1 \\ \text{logit}^{-1}(\eta - c_k) - \text{logit}^{-1}(\eta - c_{k+1}), & \text{if } 1 < k < K \\ \text{logit}^{-1}(\eta - c_k) - 0, & \text{if } k = K \end{cases} \tag{4}$$

本稿ではこのような GRM を、受験者 s を楽曲、問題 q をアノテータとする Likert 評価タスクに適用することを考える。従来、順序尺度を持つ多値型のアノテーションを複数人で実施した事例としては、Bogdanov *et al.* [2] による、1 曲につき 3 名の相対アノテーション（$C = 3$）があり、その結果が多数決に基づいて分類された。また、Yang *et al.* [9] は、VA 値を −1.0 から 1.0 の 11 段階の順序尺度でラベル付けし（$K = 11$）、1 曲につき 10 名以上の回答を平均して扱った。本手法は、タグ付けタスクと同様に、多数決と異なってアノテータ数が奇数であっても適用可能であり、単純な平均よりもアノテータの特性を考慮してより正確な評価値を得ることが期待される。

表 2 音楽ジャンル 15 種

英語	日本語
Rock	ロック
Electronic	エレクトロニック
Pop	ポップ
World & Country*	ワールド&カントリー*
Jazz	ジャズ
Classical	クラシック
Hip hop	ヒップホップ
Funk/Soul	ファンク/ソウル
Stage & Screen	ステージ&スクリーン
Blues	ブルース
Reggae	レゲエ
Latin	ラテン
Non-music	ノンミュージック
Children's	キッズ
Brass & Military	ブラス&ミリタリー

*Discogs と異なる

4. 実験

本章では、前章までで説明した IRT と GRM について、音楽アノテーションの実例に適用して統合した結果を報告し、機械学習への応用可能性を議論する。音楽タグ付けにおいては楽曲ジャンルとセマンティックタグ、Likert 評価においては 7 段階の歌唱力評価を対象とする。

4.1 IRT による音楽タグ付け結果の統合

IRT を用いた音楽アノテーションの統合の実例として、AIST Music Database (AIST-MDB) の日本語楽曲と、それに付与された音楽タグ付けを対象とする。

4.1.1 データ（楽曲及びアノテーション）

AIST-MDB は、産業技術総合研究所 (AIST) が学術目的で構築した非公開の音楽データベースである。2018 年前後のポピュラー音楽シーンを反映した多様な楽曲を、新規に作詞、作曲、編曲、レコーディング、ミックスダウン、マスタリングをして制作した。対象とする楽曲データはそのうちの日本語楽曲 120 曲とする。

アノテータは、母国語を日本語とする音楽のエキスパート（歌唱に関する知識、音楽的な知識及びそれらの評価に

表 3 音楽サブジャンル 38 種

Alternative	オルタナティブ
Ballad	バラード
Chill/Chillout	チル、チルアウト
Classic	クラシック
Country	カントリー
Dance	ダンス
Easy listening	イージーリスニング
Electronic	エレクトロニック
Electro	エレクトロ
Folk	フォーク
Funk	ファンク
Hard rock	ハードロック
Heavy metal	ヘビーメタル
Hip-Hop	ヒップホップ
House	ハウス
Indie pop	インディーポップ
Indie rock	インディーロック
Jazz	ジャズ
Lounge	ラウンジ
Mellow	メロウ
Metal	メタル
New age	ニューエイジ
Oldies	オールディーズ
Orchestral	オーケストレイション
Party	パーティー
Pop	ポップ
Popfolk	ポップフォーク
Poprock	ポップロック
Progressive rock	プログレッシブ・ロック
Punk	パンク
Reggae	レゲエ
R&B	R&B
Rock	ロック
Soul	ソウル
Soundtrack	サウンドトラック
Techno	テクノ
Trance	トランス
World	ワールド

関する経験が十分あり、曲を客観的に聴いて評価・タグ付けできる）の 6 名（M1〜M3 の男性 3 名、F1〜F3 の女性 3 名）で、性別が同一とならない（例：女性 2 名＋男性 1 名）ように、3 名ずつ 2 組「M1, F1, F3」「M2, M3, F2」に分けた。アノテーションとしては、1 曲毎に、いずれかの組のアノテータ 3 名が音楽ジャンル等のタグ付けをしたデータを用いる。各組は、120 曲の半分の 60 曲が割り当てられて、Discogs[*2]に基づく音楽ジャンル 15 種（**表 2**）を必ず 1 つ以上タグ付けした。Discogs は大規模な音楽ジャンルのオープンデータベースであり、メタデータ分析 [29] や音楽ジャンル埋め込みに関する研究 [30,31] でも対象とされている。

[*2] https://www.discogs.com/ja/

表 4 音楽セマンティック（感情・ムード・テーマ）タグ 28 種

英語	日本語
Angry/Aggressive	アングリー/アグレッシブ
Arousing/Awakening	覚醒させる/目覚めさせる
Beat	ビート的、ビート感
Beautiful	美しい
Bizarre/Weird	奇妙/奇怪
Calming/Soothing	落ち着く/心地よい、うっとりする
Carefree/Lighthearted	気楽な/気軽な
Catchy	受けそうな、人気を呼びそうな
Cheerful/Festive	陽気な/にぎやかな
Emotional/Passionate	情緒的な、感情に訴える/情熱的な
Exciting/Thrilling	わくわくする/ぞくぞくする、スリル満点の
Fast	速い
Happy	ハッピーな
Laid back/Mellow	のんびり/まったり
Light/Playful	軽い/遊び心がある
Loud	うるさい
Loving/Romantic	愛情にあふれた/ロマンチックな
Pleasant/Comfortable	楽しい/心地よい
Positive/Optimistic	前向き/楽観的
Powerful/Strong	力強い/強い
Quiet	静かな
Sad	悲しい
Sexy	セクシーな、色気のある
Slow	ゆっくり
Tender/Soft	やわらかい/ソフトな
Touching/Loving	感動/愛
Energetic	エネルギッシュな
Relaxing	リラックスさせる、くつろいだ気分にさせる

その後、先行研究に基づいて決定したサブジャンル 38 種（**表 3**）と、感情・ムード・テーマ等のセマンティックタグ 28 種（**表 4**）を、それぞれ必ず 1 つ以上、該当するタグを付与した。これらは、MagnaTagATune (MTAT)、Million Song Dataset (MSD)、MTG-Jamendo、CAL500exp を活用した研究 [6,32–36] で用いられるタグ（上位 50 のタグなど）を参考にした。

4.1.2 データの分布

図 1、**図 2**、**図 3** に、AIST-MDB 120 曲に付与されたジャンル、サブジャンル、及びセマンティックタグの頻度分布を示す。横軸はそれぞれのタグが付与された楽曲数、n は付与したアノテータ数である。ポピュラー音楽シーンを反映した楽曲群として構築したデータベースであるが、その通り Pop が最も多く付与されている。また、セマンティックタグは、楽曲ジャンルよりは偏りの少ない結果であった。

ここで、ジャンルとサブジャンルに同じタグがあるが（「Pop」や「Jazz」など）、これらの頻度は必ずしも一致しなかった。これは、タグ付けの順序からも、サブジャンルは補助的なタグとして機能しているためである。

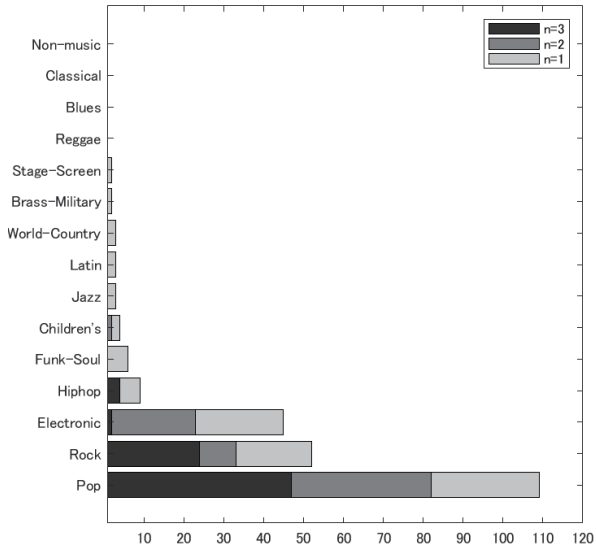

図 1 音楽ジャンルタグの頻度分布。横軸はそれぞれのタグが付与された楽曲数で、n は付与したアノテータ数。

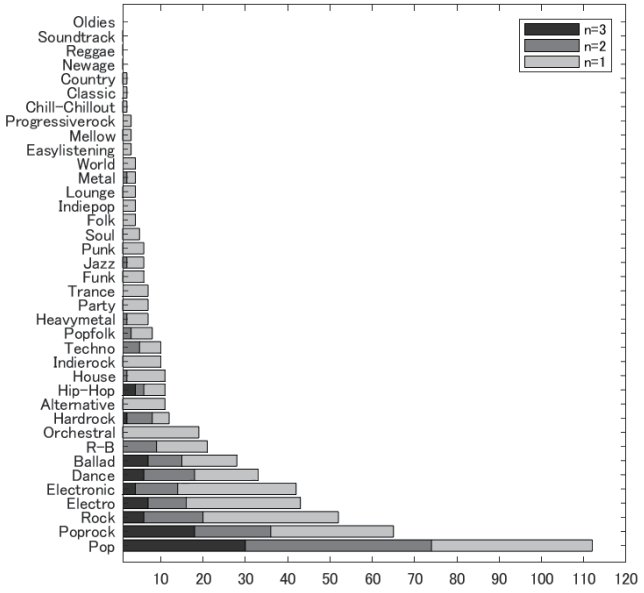

図 2 音楽サブジャンルタグの頻度分布。横軸はそれぞれのタグが付与された楽曲数で、n は付与したアノテータ数。

4.1.3 モデル

3.1 節で述べたように、音楽タグ付けのモデル化には式 (1) に示す IRT モデル（2PLM）を用いる。本研究では IRT を音楽タグ付けに適用するため、各楽曲 s_i 毎に、楽曲の潜在的なタグの度合い（能力）を表す θ_i を考え、**表 5** のような二値反応データ $X^{(m)} = \{x_{ij}^{(m)}\}(i = 1 \cdots N_m, j = 1 \cdots N_a)$ から θ_i を推定する。ここで N_m は楽曲数、N_a はアノテータ数である。つまり、タグごとに $(N_m \times N_a)$ のサイズの二値行列を用いて、そのタグのための θ_i を推定する。

本稿では、2PLM の各種パラメータに対して、以下のような事前分布を仮定する。

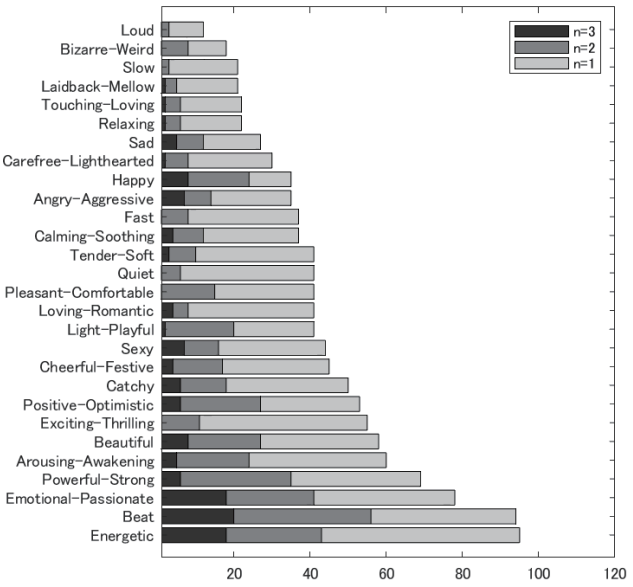

図 3 セマンティック（感情・ムード・テーマ）タグの頻度分布。横軸はそれぞれのタグが付与された楽曲数で、n は付与したアノテータ数。

表 5 Pop（音楽ジャンル）のアノテーション結果。*はアノテーションしていないこと（欠損値）を意味する。ただし本稿では、欠損値がなくなるように、アノテータ「M1, F1, F3」の 3 名と「M2, M3, F2」の 3 名の結果に対し、別々にパラメータ推定を行った。

曲番号	M1	M2	M3	F1	F2	F3
001	1	*	*	1	*	0
002	0	*	*	0	*	0
003	0	*	*	0	*	1
004	1	*	*	1	*	0
005	0	*	*	0	*	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
117	*	1	0	*	0	*
118	*	0	0	*	0	*
119	*	0	0	*	0	*
120	*	1	0	*	0	*

$$\theta_i \sim \text{Normal}(0.0, 1.0), \quad i = 1 \cdots N_m \tag{5}$$

$$a_j \sim \text{HalfNormal}(1.0), \quad j = 1 \cdots N_a \tag{6}$$

$$b_j \sim \text{Normal}(0.0, 1.0), \quad j = 1 \cdots N_a \tag{7}$$

表 5 に示すタグ付けにおいては、120 曲の楽曲に対してアノテータ「M1, F1, F3」の 3 名が 60 曲、残りの「M2, M3, F2」の 3 名が残りの 60 曲にタグ付けをした。本稿では、これら「M1, F1, F3」の 3 名と「M2, M3, F2」の 3 名の結果を別々に扱い、欠損値がなくなるように θ_i を推定した。したがって、楽曲数 $N_m = 60$、アノテータ数 $N_a = 3$ である。学習データの少なさと、上記の 3 名 2 組がそれぞ

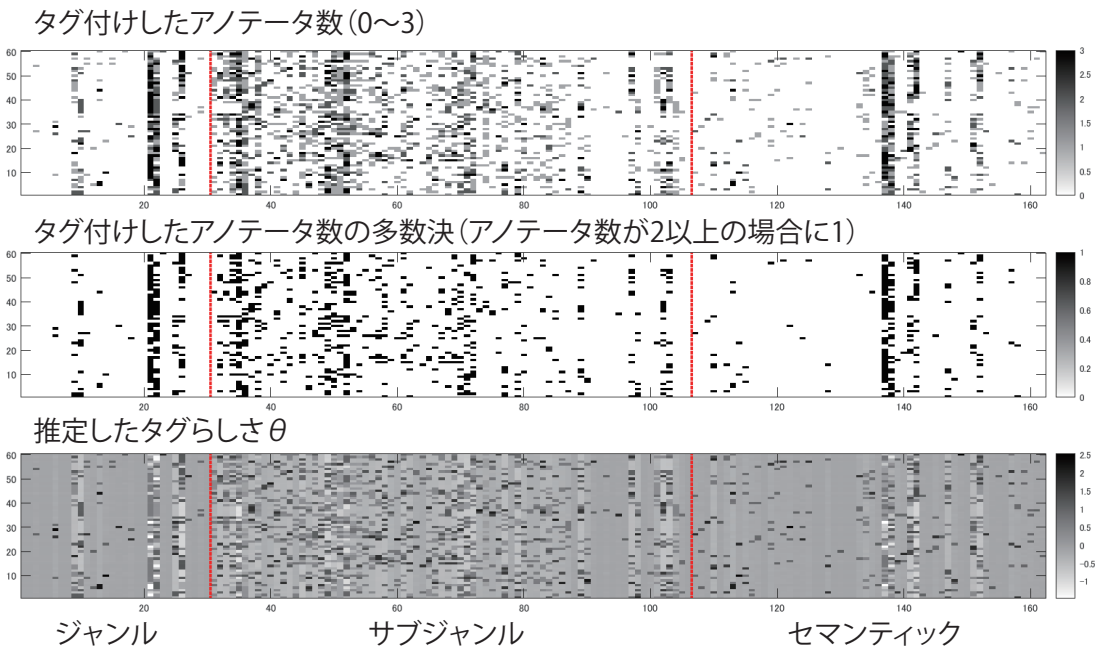

図 4 タグ付けしたアノテータ数とその多数決結果、そして IRM で推定された θ。横軸は各タグを意味するが、3 名のアノテータが 60 曲ずつアノテーションしたため、一つのタグにつき 2 列存在する。縦軸は楽曲（ただし、曲番号ではなく、各行が必ずしも同一の楽曲ではないことに注意）。

れ重複なく異なる 60 曲をタグ付けしたことから、120 曲 6 名のデータを同時に使ってしまうと、今回用いたモデルでは適切にパラメータ推定が行えなかったことが理由である。

モデルパラメータ θ, a, b は、マルコフ連鎖モンテカルロ法（MCMC）の一種である No-U-Turn Sampler (NUTS) [37] を用いて直接推定した[*3]。Burn-in 数は 1000、draw 数は 1000、chain 数は 3 とした。つまり、事後サンプル 3000 を用いて、その事後平均を推定結果として用いた。

4.1.4 結果

図 4 に、タグをつけたアノテータ数とその多数決結果、そして推定された θ をそれぞれ示す。横軸が各タグを意味し、縦軸は楽曲数を意味する。ただし、60 曲ずつタグの θ を推定したため、横軸はタグの数（$81 = 15 + 38 + 28$）の 2 倍（$= 192$）である。

θ は、アノテータ数と高い相関を持ちながら、連続値として活用できることが分かる。実際、アノテータ数との相関の平均は 0.991、多数決結果との相関の平均は 0.814 であった。ただし、誰もタグをつけなかった場合は、相関係数が計算できない（分母が 0 になる）ので、上記の平均計算からは除いた。

4.2 GRM による Likert 評価評価結果の統合

GRM を用いた音楽アノテーションの統合の実例として、AIST Singing Experience Database (AIST-SEDB-120JP) の日本語歌唱と、それに付与された 7 段階の歌唱力評価結果を対象とする。

表 6 歌唱力の 7 段階評価基準

評価	説明
7	プロアーティスト
6	セミプロ（少額でも報酬を受け取れる）
5	プロを目指しレッスンを受けているアマチュア
4	カラオケが上手い
3	可もなく不可もない
2	カラオケは行くが下手
1	音痴でカラオケも行かない

4.2.1 データ（楽曲及びアノテーション）

AIST-SEDB-120JP は、産業技術総合研究所 (AIST) が学術目的で構築した非公開の歌声データベースである。歌唱経験・声質・歌い方のバラエティに富んだ歌唱者（ボーカリスト）40 名（歌手番号 1〜20 の男性 20 名、歌手番号 21〜40 の女性 20 名）が、それぞれ 3 曲ずつ歌った 120 歌唱と、それらを 10 名（M4〜M8 の男性 5 名、F4〜F8 の女性 5 名）の評価者が詳細に歌唱力評価をしたアノテーションを含む。歌唱者及び評価者はすべて、日本語が母国語である。評価者は、ポピュラー音楽の歌唱を客観的に聴いて評価できるエキスパート (ボイストレーナー、コンテスト審査経験者等) に依頼した。

対象楽曲は RWC 研究用音楽データベース [38] から、（歌唱者が複数人ではなく）単独歌唱の 20 曲（男女 10 曲ずつ）を選曲[*4]した。これらの 20 曲から 40 名が 3 曲ずつ歌った

[*3] PyMC5（`https://www.pymc.io/welcome.html`）を用いた。

[*4] 楽曲番号 RWC-MDB-P-2001 No. 7, 12, 13, 16, 24, 27, 28, 35, 37, 38, 46, 47, 54, 56, 62, 65, 76, 77, 78, 80。

表 7 女性歌唱（原曲の楽曲番号 RWC-MDB-P-2001 No.7）と男性歌唱（RWC-MDB-P-2001 No.12）の歌唱力評価の結果（評価観点：総合力）。ここで歌手番号「−」は、原曲の歌手を意味する。

楽曲番号（RWC-MDB-P）	歌手番号	M4	M5	M6	M7	M8	F4	F5	F6	F7	F8	平均	標準偏差
7	−	6	4	7	5	6	5	6	5	5	6	5.5	0.850
7	23	6	5	6	6	7	6	6	6	6	7	6.1	0.568
7	26	4	4	5	4	5	3	4	4	5	3	4.1	0.738
7	31	3	3	4	4	4	3	4	4	3	3	3.5	0.527
7	34	4	3	3	3	3	3	3	3	3	3	3.1	0.316
7	37	2	2	2	2	2	2	2	2	2	2	2.0	0.000
7	40	1	1	1	1	1	1	1	1	1	1	1.0	0.000
12	−	4	3	6	5	5	4	4	6	4	5	4.6	0.966
12	3	5	5	7	6	5	4	5	7	6	7	5.7	1.059
12	6	3	4	5	4	3	3	3	4	3	3	3.5	0.707
12	10	3	3	4	3	3	3	3	3	3	3	3.1	0.316
12	14	3	3	3	3	3	3	3	3	3	3	3.0	0.000
12	17	2	2	2	2	3	2	3	2	3	2	2.3	0.483
12	20	2	2	1	2	2	1	2	2	2	1	1.7	0.483

120 歌唱に加え、RWC 研究用音楽データベースの原曲の 20 歌唱も含めて歌唱力評価することとし、計 140 歌唱に対する歌唱力評価アノテーションを得た。

歌唱者 40 名に対し、対象楽曲のカラオケ音源を用いて、冒頭から A メロ＋B メロ＋サビの単位（ただしサビ始まりの曲の場合、最初のサビは除く）の歌唱を収録した。曲のイメージや声量によってマイクとの距離を調整したり、また歌唱者によって、自分の歌いやすい距離感を持っている場合は、それに合わせた。歌唱者は事前に練習を行うなどして、曲の内容について十分に記憶して収録を行い、歌として完成するように収録した。つまり、正しいメロディを忘れて歌う等、歌唱力以外が原因で低く評価されることがないようにした。

評価者 10 名による歌唱力評価は、通常の楽曲中の歌声を聴取しているときと同じ状況になるように、カラオケ音源とミックスされた歌声に対して行い、最終的なマスタリングが施された音響信号データを用いた（44.1 kHz サンプリング / 16 bit 量子化 / 2ch）。収録及びマスタリングにおいて、ボーカルトラック以外のエフェクトなどは、原則としてすべて同一とした。ボーカルは必要に応じてリバーブをかけるなどのエフェクトを施し、全体の音量を一括で調整した。ピッチ補正は行わないが、歌手の声域を考慮して楽曲のキー（調）を変更した方が品質向上につながる場合には、キーを歌手ごとに変えてもよいものとした。ただし、RWC 研究用音楽データベースの原曲の 20 歌唱には、ピッチ補正などのエフェクトが施されていた可能性がある。

評価者は、音高、リズム、発音・滑舌、表現、発声、総合力の 6 種類の評価観点から 7 段階評価を実施し、それとは別に、評価の根拠を自由記述した。評価者毎の評価基準を統制するために、事前に 7 段階評価それぞれについて、**表 6** に示した基準及び歌唱の実例を教示した。これにより、評価値の意味がなるべく同一となり、異なる評価者間での評価を可能とすることを狙った。

4.2.2 データの分布

表 7 に、女性歌唱（原曲の楽曲番号 RWC-MDB-P-2001 No.7）と男性歌唱（RWC-MDB-P-2001 No.12）の歌唱力評価について、総合力の 1〜7 の 7 段階評価結果を示す。ここでは 2 つの歌唱の 1 つの評価観点の結果しか示していないが、実際にはこうした評価結果が、140 歌唱のそれぞれに対して、6 種類のそれぞれの評価観点で求まっている。

この表 7 から、アノテータによって評価値が異なり、No.7 の原曲（歌手番号「−」）や、No.12 の原曲、No.12 の歌手番号 3 のように、評価値がアノテータによって 7 段階中の 3 段階も異なることがあったことがわかる。一方、No.7 の歌手番号 40 のように、すべてのアノテータが同一の評価値 1 とした場合もある。

図 5 に、歌唱力評価の 6 種の評価観点について、評価者 10 名のアノテーション結果を歌唱者毎に平均した散布図行列を示す。この図からは、高評価 (7) の歌唱と低評価 (1) の歌唱数が少ないこと、それ以外の評価結果がおよそ同等であること、6 種の観点の評価結果に正の相関があることが分かる。

4.2.3 モデル

3.2 節で述べたように、歌唱力評価における Likert 尺度の段階反応データのモデル化には、式 (4) に示す GRM を用いる。本研究では、4.1.3 節同様、GRM を Likert 尺度評価に適用するため、各歌唱 s_i 毎に、歌唱の潜在的な評価値タグの度合い（能力）を表す θ_i を考え、表 7 のような 7 段階反応データ $X^{(s)} = \{x_{ij}^{(s)}\}(i = 1 \cdots N_s, j = 1 \cdots N_e)$ から θ_i を推定する。ここで $N_s = 140$ は歌唱数、$N_e = 10$ は評価者数である。つまり、評価の観点ごとに $(N_s \times N_e)$ のサイズの 7 段階反応データが格納された行列を用いて、評

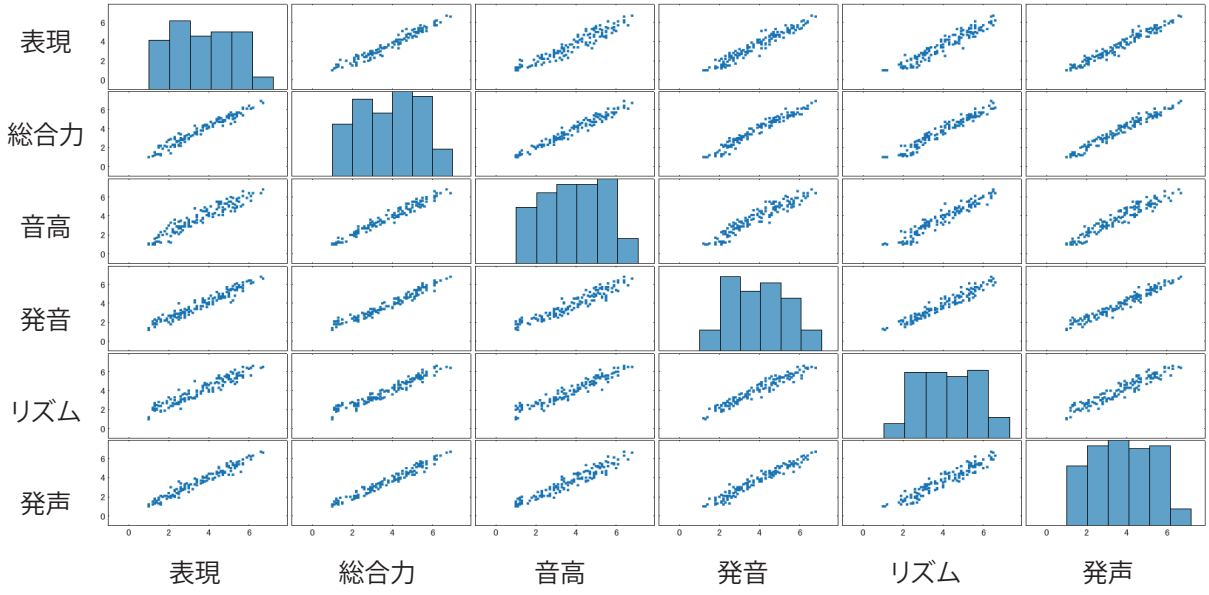

図 5 歌唱力評価結果の散布図行列。6 種の各評価軸において、評価者 10 名のアノテーション結果を歌唱毎に平均し、それらを散布図行列として示した。対角に沿った小座標軸には、対応する列のデータのヒストグラム（140 歌唱に対して、それぞれ 10 人分の評価結果の平均のヒストグラムで、自動決定された bin 数で表示）が示されている。

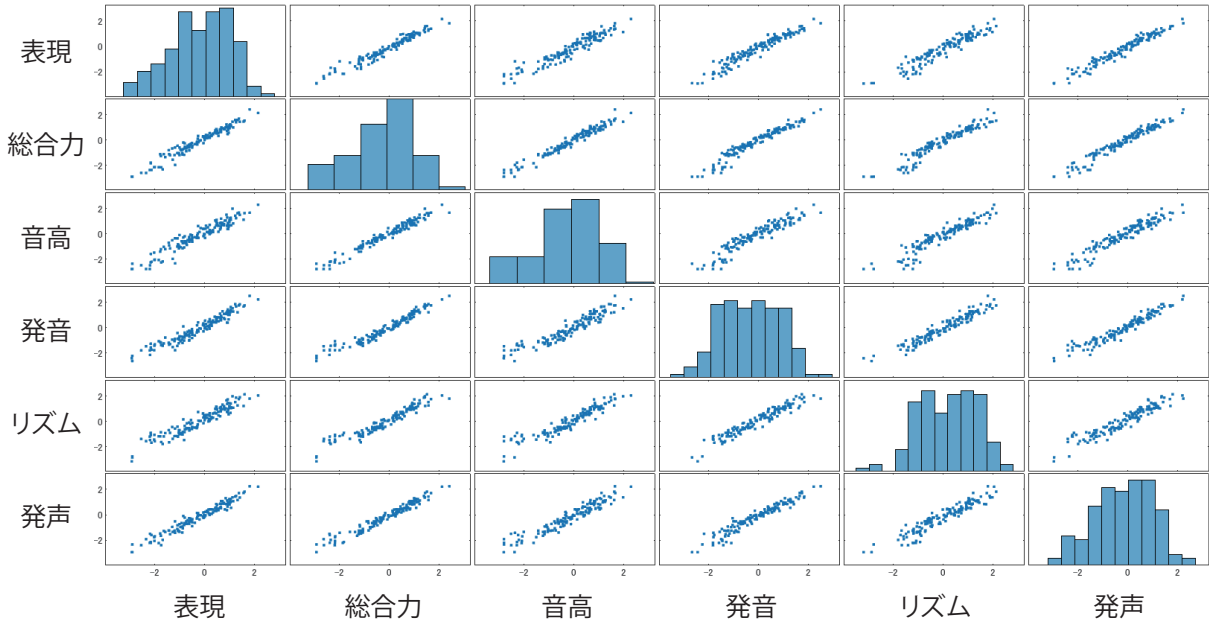

図 6 GRM で推定した歌唱力 θ の散布図行列。対角に沿った小座標軸には、対応する列のデータのヒストグラム（140 歌唱に対する θ のヒストグラムで、自動決定された bin 数で表示）が示されている。

価観点毎の θ_i を推定する。

本稿では、GRM の各種パラメータに対して、以下のような事前分布を仮定する。

$$\theta_i \sim \text{Normal}(0.0, 1.0), \quad i = 1 \cdots N_s \tag{8}$$

$$a_j \sim \text{HalfNormal}(1.0), \quad j = 1 \cdots N_e \tag{9}$$

$$b_{j,k} \sim \text{Normal}(\mu_k, 1.0), \quad k = 1 \cdots K-1 \tag{10}$$

ここで μ_k は、$\mu_1 = -1.0$ から $\mu_{K-1} = 1.0$ まで、等間隔で分割した値とした。7 段階反応データなので $K = 7$ となる。

モデルパラメータ θ, a, b は、4.1.3 節と同様に NUTS (MCMC) [37] でサンプリングし、Burn-in 数は 1000、draw 数は 1000、chain 数は 3 として、事後平均として推定した。

4.2.4 結果

図 6 に GRM で推定した歌唱力 θ の散布図行列を、**図 7** に歌唱力評価アノテーション結果の歌唱毎の平均と θ の散

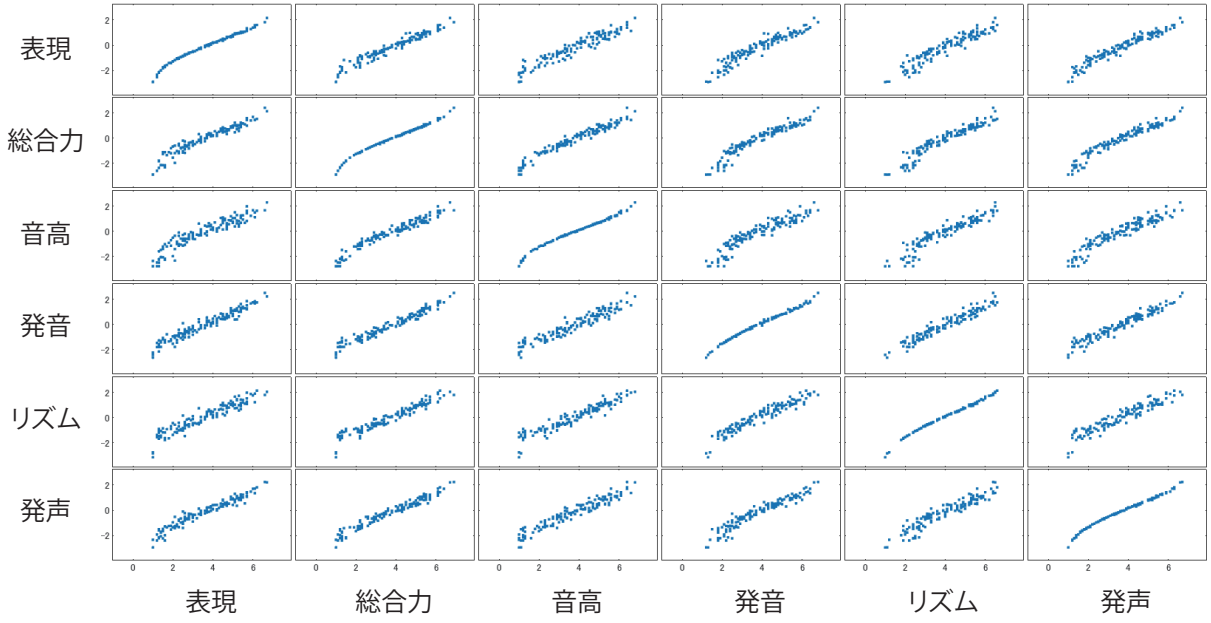

図 7 歌唱力評価アノテーション結果の平均（横軸）と、GRM で推定した歌唱力 θ（縦軸）の散布図行列。

布図行列を示す。

図 6 の GRM で推定した歌唱力 θ を、図 5 の歌唱力評価アノテーション結果の歌唱毎の平均と比較すると、6 種類の評価観点毎の値の頻度分布が正規分布に近づいた形状をしていることが分かる。また、異なる観点間にも正の相関がある。

さらに、図 7 からは、歌唱力評価アノテーション結果の平均と θ においても正の相関があることが分かり、実際、ピアソンの線形相関係数の平均は 0.99 であった。

5. 機械学習への応用に関する考察

前章までで、実際の音楽アノテーション結果を用いて、提案手法に基づいて IRT と GRM のそれぞれによるパラメータ推定を行った。その結果、本稿では機械学習への応用の第一段階として、θ に関して、4.1 節の IRT ではタグ付けしたアノテータ数、4.2 節の GRM ではアノテーション結果の平均と相関があることを示すことができた。したがって、このようにして得られた θ は、回帰モデル等の機械学習データとして十分利用可能なはずである。また、もし判断の曖昧性が高いアノテータが含まれていても、そのようなアノテータの影響を受けにくい可能性 [27] があるため、今後、実際に機械学習データとして扱う実験により検証していく予定である。

深層学習を用いた音楽のタグ推定タスクにおいては、そのタグが付与されている（1）かいない（0）かの二値ラベルが用いられ、Binary Cross Entropy ロスによって学習される [39]。したがって予測時には 0〜1 の連続値となるが、学習データとしてはそのような連続性を持っていなかった。しかし、実際の音楽タグ付けにおいて、アノテーション結果が一つに定まらないということは、それぞれのタグ付けの段階でその度合いも含めることが有効である可能性がある。実際、楽曲中の構造の区間境界のアノテーションにおいて、複数人のアノテーション結果から、そのような度合いの強さを分析した事例がある [10]。従来の「タグ付けしたアノテータ数」は、このような度合いの強さを反映させることが可能な一つの方法であるが、解像度の低い整数値となってしまい、アノテータ数が同じ楽曲における度合いの違いを反映できない（同点になってしまう）。それに対して提案手法では、4.1 節の IRT で得られる θ のように、アノテータ毎の特性を反映して推定することから、同点にならない特性を持っていることが知られていて [40]、より解像度高く推定できる利点がある。

また、Likert 尺度に基づいた評価を機械学習データとする場合、それを間隔尺度として考えれば平均を取って連続値を得ることができる。しかし、Likert 尺度に基づいた評価結果を間隔尺度として扱えるか、それとも順序尺度として扱うべきかは議論があり [41]、直感的にも、この間隔が人によって異なる可能性は十分ありうる。そこで、4.2 節の GRM によって推定された結果から項目反応曲線（式 (4)）を求めたところ、**図 8** のようになった。この図から、評価者によって判断基準が異なっていたことが分かる。例えば、M5 はそれぞれの点数をつける範囲が広く、判断基準が比較的曖昧である。このように、従来のように Likert 尺度から中央値や平均値を算出する代わりに、その間隔に相当する $b_{j,k}$ を推定しながら統合できる提案手法には意義がある。

深層学習において、Likert 尺度のように順序関係のある離散カテゴリを出力する方法としては、順序回帰問題を

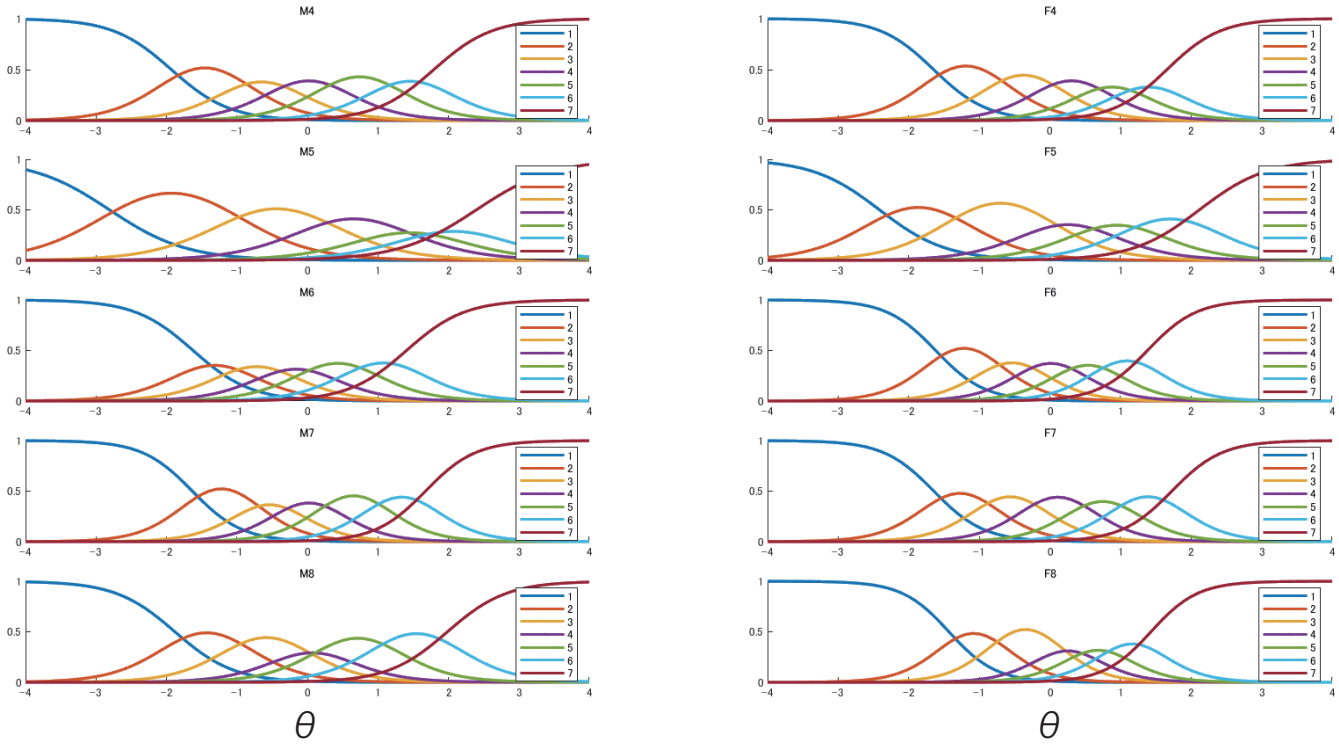

図 8 GRM による歌唱力評価者の項目反応曲線（式 (4)）

バイナリ分類の部分問題に置き換えてそれを統合する方法 [42,43] が知られていて、多クラス分類アプローチと比較して推定誤差が小さいことが報告されている [43]。また、分類問題として解いた後に、ソフトマックス期待値で重み付き平均して推定する方法が知られている [44]。したがって、提案手法を用いることで、順序回帰を回帰問題に置き換えて連続値で扱えることは、より機械学習性能を上げることができる可能性がある。

6. おわりに

本稿では、項目反応理論を用いた音楽アノテーション結果の統合手法を検討した。音楽のアノテーションは多様であるが、その中でも、実用可能性が高いと考えられるタグ付けと、Likert 尺度に基づく評価を対象とした。具体的には、音楽ジャンル・セマンティックタグと、歌唱力評価を対象として、それぞれ IRT の 2PLM と GRM でモデル化する方法を提案した。

実験で用いるデータとして、アノテーション自体の品質・信頼性が十分高くなるよう、音楽及び歌唱に関するエキスパートがアノテーションしたデータベースを独自に構築して用意した。これにより、アノテーション自体の精度・信頼性が低いことが原因となるような考察の困難さを、未然に防げると考えている。

今後、今回得られた θ の機械学習における有効性を、従来のタグ付けにおけるアノテータ数や、Likert 評価結果の平均値と比較評価して検証するために、実際の機械学習実験の結果を報告する予定である。

謝辞 本研究の一部は JST CREST JPMJCR20D4 と JSPS 科研費 JP21H04917 の支援を受けた。

参考文献

[1] Kim, K. L., Lee, J., Kum, S., Park, C. L. and Nam, J.: Semantic Tagging of Singing Voices in Popular Music Recordings, *IEEE/ACM TASLP*, Vol. 28, pp. 1656––1668 (2020).

[2] Bogdanov, D., Lizarraga-Seijas, X., Alonso-Jiménez, P. and Serra, X.: MusAV: A Dataset of Relative Arousal-Valence Annotations for Validation of Audio Models, *Proc. ISMIR 2022*, pp. 650–658 (2022).

[3] Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., Roure, D. D. and Downie, J. S.: Design and Creation of a Large-Scale Database of Structural Annotations, *Proc. ISMIR 2011*, pp. 555–560 (2011).

[4] Nieto, O. and Bello, J. P.: Systematic Exploration Of Computational Music Structure Research, *Proc. ISMIR 2016*, pp. 547–553 (2016).

[5] Nieto, O., McCallum, M. C., Davies, M. E. P., Robertson, A., Stark, A. M. and Egozy, E.: The Harmonix Set: Beats, Downbeats, and Functional Segment Annotations of Western Popular Music, *Proc. ISMIR 2019*, pp. 565–572 (2019).

[6] Turnbull, D., Barrington, L., Torres, D. A. and Lanckriet, G. R. G.: Semantic Annotation and Retrieval of Music and Sound Effects, *IEEE Trans. Speech Audio Process.*, Vol. 16, No. 2, pp. 467–476 (2008).

[7] Gupta, C., Li, H. and Wang, Y.: Automatic Leaderboard: Evaluation of Singing Quality Without a Standard Reference, *IEEE/ACM TASLP*, Vol. 28, pp. 13–26 (2020).

[8] Yang, Y.-H., Lin, Y.-C., Lee, A. and Chen, H. H.: Im-

proving Musical Concept Detection by Ordinal Regression and Context Fusion, *Proc. ISMIR 2009*, pp. 147–152 (2009).
[9] Yang, Y.-H., Lin, Y.-C., Su, Y.-F. and Chen, H. H.: A Regression Approach to Music Emotion Recognition, *IEEE TASLP*, Vol. 16, No. 2, p. 448–457 (2008).
[10] Bruderer, M. J., McKinney, M. and Kohlrausch, A.: Structural boundary perception in popular music, *Proc. ISMIR 2006*, pp. 198–201 (2006).
[11] McFee, B., Nieto, O., Farbood, M. M. and Bello, J. P.: Evaluating Hierarchical Structure in Music Annotations, *Frontiers in Psychology*, Vol. 8, pp. 1–17 (2017).
[12] Law, E. L. M., von Ahn, L., Dannenberg, R. B. and Crawford, M.: TagATune: A Game for Music and Sound Annotation, *Proc. ISMIR 2007*, pp. 361–364 (2007).
[13] Libeks, J. and Turnbull, D.: You Can Judge an Artist by an Album Cover: Using Images for Music Annotation, *IEEE MultiMedia*, Vol. 18, No. 4, pp. 30–37 (2011).
[14] Turnbull, D., Liu, R., Barrington, L. and Lanckriet:, G. R. G.: A Game-Based Approach for Collecting Semantic Annotations of Music, *Proc. ISMIR 2007*, pp. 535–538 (2007).
[15] Lord, F. M.: *Applications of Item Response Theory to Practical Testing Problems*, L. Erlbaum Associates (1980).
[16] Samejima, F.: Estimation of Latent Ability Using a Response Pattern of Graded Scores, *Psychometrika monograph supplement* (1969).
[17] Lippens, S., Martens, J.-P. and Mulder, T. D.: A Comparison of Human and Automatic Musical Genre Classification, *Proc. IEEE ICASSP 2004*, pp. 233–236 (2004).
[18] Seyerlehner, K., Widmer, G. and Knees, P.: A Comparison of Human, Automatic and Collaborative Music Genre Classification and User Centric Evaluation of Genre Classification Systems, *Proc. Adaptive Multimedia Retrieval. Context, Exploration, Fusion*, pp. 118–131 (2010).
[19] Soleymani, M., Aljanaki, A., Yang, Y.-H., Caro, M. N., Eyben, F., Markov, K., Schuller, B. W., Veltkamp, R. C., Weninger, F. and Wiering, F.: Emotional Analysis of Music: A Comparison of Methods, *Proc. ACMMM 2014*, pp. 1161–1164 (2014).
[20] Fan, J., Tatar, K., Thorogood, M. and Pasquier, P.: Ranking-Based Emotion Recognition for Experimental Music, *Proc. ISMIR 2017*, pp. 368–375 (2017).
[21] Flexer, A. and Grill, T.: The Problem of Limited Inter-Rater Agreement in Modelling Music Similarity, *J. New Music Res.*, Vol. 45, No. 3, pp. 239–251 (2016).
[22] Koops, H. V., de Haas, W. B., Burgoyne, J. A., Bransen, J., Kent-Muller, A. and Volk, A.: Annotator Subjectivity in Harmony Annotations of Popular Music, *J. New Music Res.*, Vol. 48, No. 3, pp. 232–252 (2019).
[23] Louviere, J. J., Flynn, T. N. and Marley, A. A. J.: *Best-Worst Scaling: Theory, Methods and Applications*, Cambridge University Press (2015).
[24] O'Donovan, P., Libeks, J., Agarwala, A. and Hertzmann, A.: Exploratory Font Selection Using Crowdsourced Attributes, *ACM Transactions on Graphics (TOG)*, Vol. 33, No. 4, pp. 1–9 (2014).
[25] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L. and Moy, L.: Learning From Crowds, *Journal of Machine Learning Research*, Vol. 11, pp. 1297–1322 (2010).
[26] 川島寛乃，持橋大地：多次元項目反応理論による短歌の評価傾向の分析，情報処理学会研究報告，Vol. 2023-NL-256, pp. 1–15 (2023).
[27] 大谷直樹，中澤敏明，黒橋禎夫：機械翻訳評価のための項目反応理論に基づく一対比較結果の統合，*Japio year book*, pp. 284–289 (2017).
[28] Sharma, C., Pahari, A. K., Balasubramanian, V. N. and Vijaykeerthy, D.: Do Models See Corruption as We See? An Item Response Theory Based Study in Computer Vision, *Proc. ICLR 2023* (2023).
[29] Bogdanov, D. and Serra, X.: Quantifying Music Trends and Facts Using Editorial Metadata From the Discogs Database, *Proc. ISMIR 2017*, pp. 89–95 (2017).
[30] Hennequin, R., Royo-Letelier, J. and Moussallam, M.: Audio Based Disambiguation of Music Genre Tags, *Proc. ISMIR 2018*, pp. 645–652 (2018).
[31] Alonso-Jiménez, P., Serra, X. and Bogdanov, D.: Music Representation Learning Based on Editorial Metadata from Discogs, *Proc. ISMIR 2022*, pp. 825–833 (2022).
[32] Law, E., West, K., Mandel, M. I., Bay, M. and Downie, J. S.: Evaluation of Algorithms Using Games: The Case of Music Tagging, *Proc. ISMIR 2009*, pp. 387–392 (2009).
[33] Lee, J., Park, J., Kim, K. L. and Nam, J.: Sample-Level Deep Convolutional Neural Networks for Music Auto-Tagging Using Raw Waveforms, *Proc. SMC 2017*, pp. 220–226 (2017).
[34] Pons, J., Nieto, O., Prockup, M., Schmidt, E., Ehmann, A. and Serra, X.: End-to-End Learning for Music Audio Tagging at Scale, *Proc. ISMIR 2018*, pp. 637–644 (2018).
[35] Bogdanov, D., Won, M., Tovstogan, P., Porter, A. and Serra, X.: The MTG-Jamendo Dataset for Automatic Music Tagging, *Proc. ICML 2019* (2019).
[36] Wang, S.-Y., Wang, J.-C., Yang, Y.-H. and Wang, H.-M.: Towards Time-varying Music Auto-tagging Based on CAL500 Expansion, *Proc. IEEE ICME 2014*, pp. 1–6 (2014).
[37] Hoffman, M. D. and Gelman, A.: The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, *The Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1593–1623 (2014).
[38] 後藤真孝，橋口博樹，西村拓一，岡　隆一：RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol. 45, No. 3, pp. 728–738 (2004).
[39] Won, M., Ferraro, A., Bogdanov, D. and Serra, X.: Evaluation of CNN-based Automatic Music Tagging Models, *Proceedings of the Sound and Music Computing Conference* (2020).
[40] 独立行政法人日本学術振興会グローバル学術情報センター：CGSI レポート 第 7 号，技術報告 (2018).
[41] Wu, H. and Leung, S.: Can Likert scales be treated as interval scales?—A Simulation study，*Journal of social service research* (2017).
[42] Niu, Z., Zhou, M., Wang, L., Gao, X. and Hua, G.: Ordinal Regression with Multiple Output CNN for Age Estimation, *Proc. CVPR 2016*, pp. 4920–4928 (2016).
[43] Chen, S., Zhang, C., Dong, M., Le, J. and Rao, M.: Using Ranking-CNN for Age Estimation, *Proc. CVPR 2017*, pp. 742–751 (2017).
[44] Rothe, R., Timofte, R. and Van Gool, L.: Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks, *International Journal of Computer Vision*, Vol. 126, No. 2-4, pp. 144–157 (2018).