一般セッション | 一般セッション：感性に係る計測・評価技術とその活用に関する分野

2025年3月5日(水) 10:00 ~ 12:00 A会場(1号館 1階 0111講義室)

[1A01] 感性に係る計測・評価技術とその活用に関する分野（1）

座長:上條 正義(信州大学)、荒川 尚美(資生堂)

11:40 ~ 12:00

[1A01-06] データ品質と解析の客観性を考慮した大規模感性評価手法の構築 ーライブ感の構成要素の解明ー

*下薗 大樹[1]、篠井 暖[1]、塩澤 安生[1]、山崎 陽一[2]、橋本 翔[3]、長田 典子[4] (1. ヤマハ株式会社、2. 長崎県立大学、3. 西南学院大学、4. 関西学院大学)

データ品質と解析の客観性を考慮した大規模感性評価手法の構築

－ライブ感の構成要素の解明－

下薗　大樹*，篠井　暖*，塩澤　安生*，山崎　陽一**，橋本　翔***，長田　典子****

* ヤマハ株式会社，** 長崎県立大学，*** 西南学院大学，**** 関西学院大学

Development of a Large-Scale Kansei Evaluation Method Considering Data Quality and Objectivity of Analysis

– Unraveling the Components of Live Experience –

Taiki SHIMOZONO*, Dan SASAI*, Yasuo SHIOZAWA*, Yoichi YAMAZAKI**, Sho HASHIMOTO** and Noriko NAGATA**

taiki.shimozono@music.yamaha.com

Abstract: This study discusses effective data cleansing methods and an automated evaluation structure construction method for sensibility evaluation utilizing crowdsourcing. Focusing on the "live experience" in music concerts, an analysis of worker characteristics on major domestic crowdsourcing platforms revealed consistency in personality traits. Additionally, two data cleansing methods were examined to address effort-reducing behaviors in online experiments. Among them, the method using correlation values of reverse-coded items demonstrated superior balance in precision and recall. Using this method, sensibility data based on concert footage was collected from over 6,000 participants, achieving high data accuracy with 3,720 valid data points (precision: 0.96, recall: 1.00). Furthermore, an automated evaluation term construction system combining Sentence BERT, VGAE, and T5 models was developed, yielding results comparable to manual evaluation term construction.
Keywords: live performance, evaluation method, NLP

1. はじめに

感性評価では従来，データ収集に多くの時間とコストがかかる上，研究対象として設定した母集団の代表となるサンプルを確保することや、サンプルの多様性を確保することが課題であった[1]．この課題に対して，インターネットを通じて不特定多数の人々に業務を委託するクラウドソーシングの活用が，心理学研究においても長らく検討されている．一方で，クラウドソーシングによるデータ収集研究において，その品質や，収集した大量のデータの解析手法が問題視されることも少なくない．そこで本研究ではクラウドソーシングを活用した大規模感性評価にあたり，効果的なデータクレンジング手法と評価構造の自動構築手法を構築することを目指す．また，クラウドソーシングによるサンプルの多様性をするため，音楽ライブでの「ライブ感」を研究対象として設定した．

2. 先行研究

2.1 クラウドソーシングに関する先行研究

クラウドソーシングを活用した心理学研究はこれまでいくつか行われている．遠藤らの研究では，クラウドソーシングによる評価とベイズ最適化を組み合わせ，感性評価の効率的な収集方法を提案しており，その有効性が示唆されている[2]．また，クラウドソーシング内のワーカーの多様性は典型的な大学生サンプルよりも多様なサンプルであることも明らかになっている［3］．一方で，クラウドソーシングを心理学研究で活用する際の技術的あるいは倫理的配慮などについても議論されている．技術的な問題の１つとして，データの信頼性について挙げられるが，この信頼性の検証のため，従来の古典的な心理学上の知見のオンライン上での再現実験が国内外のプラットフォーム（Amazon Mechanical Turk，CrowdWorks）で行われており，どちらも従来の心理学的な傾向を再現することが示されている［4,5］．しかしながら，オンラインの協力者が注意資源を十分に割かず，必要最小限の手順で目的を達成しようとする行動，Satisfice が発生しうることが問題とされている［6］．Krosnick は，この Satisfice について，調査項目の内容を理解したうえで回答しようとしているが，選択可能な選択肢を部分的にしか検討しない「弱い省力化」と，調査項目の内容を理解するための認知的コストを払わず，誰にでも選択可能な選択肢を選んだり，あてずっぽうに選んだりする「強い省力化」に分類した[7]．

2.2 音楽ライブに関する先行研究

音楽における「ライブ感」とは、演奏者と観客が共有する独特な一体感のある体験を指す．中井らの研究では，「一体感」がライブの良さの重要な要素であるとして評価グリッド法を

用いた評価構造の解明を試みており，非日常であることと，観客やアーティストとの相乗効果重要であることを明らかにしている[8].

3. クラウドソーシングにおける回答品質の向上

本研究では，数千～数万オーダーでのデータ取得を目指し，国内の複数のプラットフォームでのクラウドソーシングによるデータ取得を検討した．また先行研究で議論されていた 2 種類の Satisfice の内「強い省力化」を対象に，省力化傾向のあるワーカーを判別するための「判別プロシージャ」の開発を行った.

3.1 国内クラウドソーシングプラットフォームのワーカーの傾向把握

本研究では，国内で規模の大きな以下の３つのクラウドソーシングプラットフォームを対象とした.

・Yahoo!クラウドソーシング
(https://crowdsourcing.yahoo.co.jp/)
・CrowdWorks(https://crowdworks.jp/)
・Lancers (https://www.lancers.jp/)

まずはプラットフォーム内のワーカーの特性を調べるために，性別，年齢，音楽経験年数，性格特性について調査を行った．性格特性については，小塩らの作成した TIPI-J(日本語版 Ten Item Personality Inventory)を用いた[9]．TIPI-J とは性格特性の１つである Big Five を測定するための指標で，10 項目の質問項目について「全く当てはまらない (1 点)」から「とてもよく当てはまる (7 点)」までの 7 段階で回答を行い，各性格特性の得点を算出する指標である．質問項目は表 1 に示す.

表 1 日本語版TIPI-Jの質問項目

外向性	活発で、 外交的だと思う	ひかえめで、 おとなしいと思う
協調性	他人に不満をもち、 もめごとを起こしやすいと思う	人に気をつかう、 やさしい人間だと思う
勤勉性	しっかりしていて、 自分に厳しいと思う	だらしなく、 うっかりしていると思う
神経症傾向	心配性で、 うろたえやすいと思う	冷静で、 気分が安定している
開放性	新しいことが好きで、 変わった考えをもつと思う	発想力に欠けた、 平凡な人間だと思う

小塩らの先行研究における大学生の性格特性の平均値とクラウドソーシングにより 3743 名より収集した性格特性の平均値と，Majima らの大学生とクラウドワーカーとの比較を表 2 に，それぞれの差を図 1 に示す．小塩らの研究の大学生実験参加者に対して，今回収集したクラウドワーカーの性格特性は，協調性・勤勉性が高く，外向性・神経症傾向が低いことが示唆された．この差の傾向は Majima らの先行研究と概ね類似した結果になっている[10]．またこの性格特性はプラットフォームごとに違いが見られなかった．性別，年齢，音楽経験年数は表 3 に記した通り，性別においては Yahoo!クラウドソーシング，Lancers，CrowdWorks の順に男性比率が高く，Yahoo!クラウドソーシングの年齢層が高めで，音楽経験年数は CrowdWorks と Lancers が比較的高い傾向にあった．この差はプラットフォームの性質の差として解釈でき，Yahoo!クラウドソーシングはその他 2 つのプラットフォームと比較して Yahoo ID を持っているだけで始めることができる敷居の低さから，より多様な人材がワーカーとして所属していることが推察される.

表 2 クラウドワーカーと大学生の性格特性の平均値

	外向性	協調性	勤勉性	神経症傾向	開放性
小塩ら，UNIV(n=9020)	7.83	9.48	6.14	9.21	8.03
本研究，CW(n=3743)	7.01	9.79	7.8	8.16	7.94
Majimaら，UNIV(n=131)	7.34	9.78	6.44	6.74	7.87
Majimaら，CW(n=295)	6.53	9.24	7.18	6.71	7.68

CW：クラウドワーカー，UNIV：大学生実験参加者

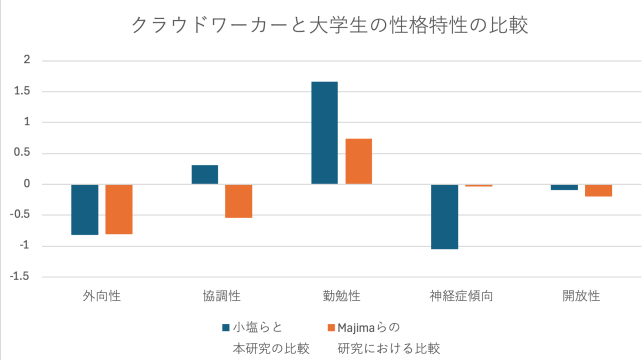

図 1 快適性と強度との関係

表 3 各クラウドソーシングプラットフォームのユーザー属性

	CrowdWorks	Lancers	Yahoo!クラウドソーシング
性別	男387, 女724 その他8	男551, 女556 その他5	男613, 女390 その他5
年齢	35歳±4歳	35歳±4歳	43歳±12歳
音楽経験年数	11年±7年	11年±7年	3年±7年

総じて，プラットフォームごとの性格特性を除く多少の属

性の違いは見られるが，性格特性におけるプラットフォーム間の同質性や，先行研究との傾向の類似性より，今回選定した 3 つのプラットフォームでのデータ収集は問題ないと考えられる．

3.2 判別プロシージャによるデータクレンジング

強い省力化傾向のあるワーカーを判別するために，2 種類の判別プロシージャの検討を行った．判別プロシージャ A は，実験前後で参加者属性の一部を重複回答させることで回答の一貫性を確認するという手続きで，判別プロシージャ B は前述の TIPI-J の性格特性項目の内，同一の指標に含まれている質問の逆転項目（例：「活発で，外交的だと思う」と「ひかえめで，おとなしいと思う」）の相関性を判定基準として採用した．

判別プロシージャの性能比較のために，134 名に対してデータ収集を行った．データ収集にあたって後述の 3.2 節のデータ収集と同様の項目を取得した．収集したデータの内，強い省力化を行っているか否かを研究者 3 名でラベル付けを行い平均値が 1 より小さい場合，不誠実回答者として判断した．このデータ真値として，全データのうち不誠実・誠実回答者を正確に判断できた割合を正解率，判別プロシージャにより不誠実回答者と判別されたもののうち実際に不誠実回答者であった割合を適合率，実際に不誠実回答者であったもののうち判別プロシージャにより判別できた割合を再現率として，2 つの判別プロシージャを追加した際の性能向上率を比較した．表 4 にその結果を記す．

表4 判別プロシージャによる性能向上率の比較

	判別プロシージャA	判別プロシージャB
正解率	−0.06	0.00
適合率	−0.46	−0.08
再現率	0.18	0.08

判別プロシージャ A では再現率が向上している一方，適合率が大きく低下しており，誠実回答者の回答を過度に除外してしまう傾向があることが分かった．判別プロシージャ B では再現率は控えめに向上するが，適合率は大きく損なわずに過度に誠実回答者の回答を除外しない傾向があることが分かった．クラウドソーシングにおいては，データを大量に集めることができることがその大きなメリットの一つであるが，判別プロシージャ A の方式ではそのメリットを十分に享受できないと言え，判別プロシージャ B の方式の方がより適格であると言える．また，判別プロシージャ A の方式は同じ質問を再度ワーカーに提示する必要がある．クラウドソーシングではタスクの長さや難易度に応じて品質が低下することが予想され，その点においても判別プロシージャ B では不要な設問を追加せずに，逆転項目によって判定することができるためクラウドソーシングにおける虚偽尺度として適切であると考えられる．

3.3 クラウドソーシングによる評価構造データの大量取得

3.2 節で作成した判別プロシージャを用いて，オンラインライブを想定した映像メディアの評価グリッド法による評価構造抽出実験を行う．得られた評価構造よりオンラインライブのライブ感の評価構造を明らかにすることを目的とする．

実験刺激として，映像・残響・観客要因が異なる 24 種類のロックライブ動画像を用意した．映像要因についてはマルチカメラ，シングルカメラの 2 水準，残響要因については楽曲全体に残響を-inf dB，0dB，18dB で付与する 3 水準，観客要因については演奏開始前に観客の歓声を-inf dB，-6dB，0dB，6dB で付与する 4 水準の刺激を用意した．

実験手続きとして，クラウドソーシングサイトにより参加者を募り，簡単な聴力検査（1000Hz，4000Hz，無音の聴取の聴こえの有無の確認），動画視聴環境の確認，性格特性に関する質問（3.2 節で作成した判別プロシージャに対応），ライブ動画視聴前の気分評価を行う．その後，実際にライブ映像を視聴してその動画から得られたライブ感の程度を評価し，動画視聴後の気分評価を行い，ライブ感を感じるために必要だった構成要因を回答，オンラインライブにおけるライブ感があるとどのような良さがあるかを回答，参加者属性（年齢・性別・視聴端末・再生機器・オンラインライブへの参加経験・実際のライブへの参加経験、よく聞く音楽のジャンル、音楽経験、音楽聴取時間）を取得する．

表5 クラウドソーシングにより取得したデータの概要

	全体	CrowdWorks	Lancers	Yahoo!クラウドソーシング
データ数	3720	426	380	2914
年齢	43.0±11.9	38.0±10.4	40.6±10.7	44.0±12.0
性別（男/女/無回答）	2269/1356/95	214/207/5	230/144/6	1832/998/84
不誠実回答者の割合	39.4	12.3	20.5	43.7

Yahoo!クラウドソーシング，CrowdWorks，Lancers によりオンラインライブ経験を有する約 6,000 名に募集をかけ 3,720 名の有効データを取得した（表 5）．各動画刺激に対して 125 名程度の有効データを確保した．

また，収集したデータについて 3.1 節と同様に，手動によるラベル付との対応関係を調査した結果，正解率 0.98，適合率 0.96，再現率 1,00 と高い精度で不誠実回答者を判別できていることが分かった．

4 自動評価語構築

3.3 節で取得したライブ感の構成要因を対象に自動的に評価語を構築する手法を検討する．全体の構成図は図 2 の通りである．

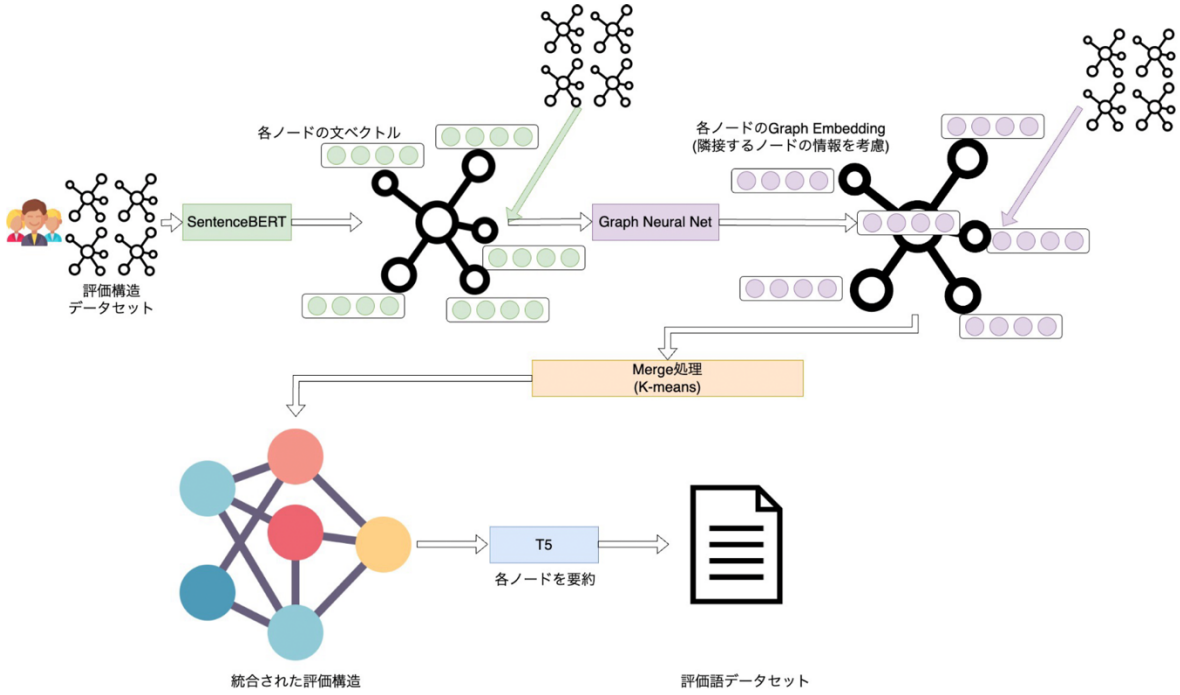

図2　評価語自動構築システム

4.1　評価構造のベクトル埋め込み

ベクトル埋め込みとは，単語や文章などのデータの関係性を表現するベクトルとして表現する手法である．

評価構造とは，人が物事の判断を行うときの構成要素の関係性のことを指し，それぞれの構成要素の意味は，構成要素の文章内容とその結合の仕方で表現されていると考えることができる．そこで本研究では，構成要素の文章の意味をベクトルとして表現するために Sentence BERT モデルを，構成要素間の関係性を VGAE モデルでベクトル埋め込みを行う．

（1）**Sentence BERT モデルによるベクトル埋め込み** Sentence BERT とは，文章単位でのベクトル埋め込みが可能で，文章間の類似度データによりファインチューニングが可能なモデルである．日本語の事前学習済みモデルを利用した（https://github.com/cl-tohoku/bert-japanese）． 3.3 節で取得した評価構造データを利用し，得られたライブ感の程度の近さによってファインチューニングを行った．ファインチューニングを行ったモデルによって，評価構造データを 1024 次元のベクトルに埋め込んだ．

（2）**VGAE(Variational Graph Auto-Encoders)モデルによるベクトル埋め込み** VGAE とは，教師なしでグラフ構造の潜在空間を学習可能なモデルである．4.1.1 項の Sentence BERT モデルにより取得したベクトルデータと評価構造の隣接行列を特徴量としてファインチューニングを行った．ファインチューニングを行ったモデルによって，評価構造データを 128 次元のベクトルに埋め込んだ．

4.2　評価構造の構成要素の結合

4.1 節で作成した 1024 次元+128 次元のベクトルを結合し，評価構造における各構成要素の特徴量を 1152 次元のベクトルとして表現した．ベクトルに変換された各評価項目について K-means でクラスタリングを行い，同一クラスターに所属する構成要素を T5 モデルにより要約して評価項目に変換した．

（1）**K-means による段階的クラスタリング** K-means によるクラスタリングで課題となるのは，クラスター数の決定である．本研究では，クラスター数を明示的に決定するのではなく，クラスター数を段階的に変化させ，要約により出現した評価項目の出現頻度をカウントして，これを評価項目の「確信度」として捉えることで評価語選定の情報として活用することを検討した．10 クラスターから 300 クラスターまで 10 クラスター刻みで評価項目を要約抽出した際の音響要因に関する評価項目の出現数は図 3 の通りである。

（2）**T5(Text-to-Text Transfer Transformer)による評価項目抽出** T5 とは，テキスト入力から新しいテキスト出力を行える抽象型モデルである．日本語の事前学習済みモデルを利用した（https://huggingface.co/sonoisa/t5-base-japanese）．

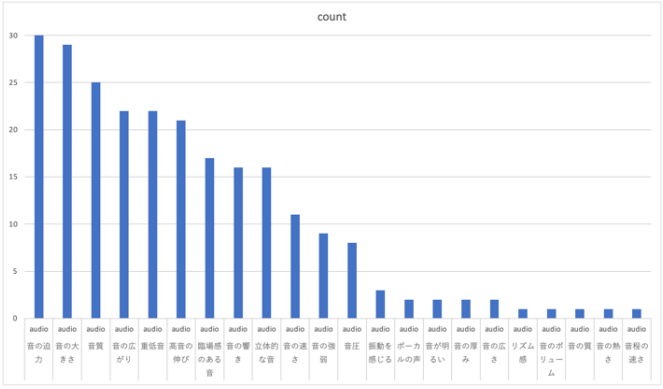

図 3 音響要因に関する評価項目の出現数

3.3 節で取得した評価構造データに対して研究者で評価項目に整理したデータと，社内でこれまで利用していた評価構造データを用いてファインチューニングを行った．ファインチューニングを行ったモデルによって，同一クラスター内の構成要素の文章を一連の文章として評価項目に要約した．

4.3 手動結果との比較

自動評価語構築システムにより構築された評価項目と，研究者によって評価項目に仕分けしたものを表 6 に記す．手動で構築された評価語と概ね類似した評価項目が自動抽出されていることが確認できた．

図6 手動と評価語自動構築システムによる評価語の比較

	手動	評価語自動構築システム
音響要因	音質が良い・悪い	音質
	音の響きが良い・悪い	音の響き
	音が大きい・小さい	音の大きさ
	音の迫力がある・ない	音の迫力
		臨場感のある音
映像要因	ステージの照明が良い・悪い	ステージの照明
	画質が良い・悪い	画質
		カメラワーク
その他内評価	観客の盛り上がりが感じられる・感じられない	観客の盛り上がり
	リアルさがある・ない	リアルさ
	迫力がある・ない	迫力
	アーティストの演奏能力が高い・低い	アーティストの演奏能力
		コールアンドレスポンス
		観客の映り込み
		アドリブ
その他外評価	緊張感がある・ない	
	臨場感がある・ない	臨場感
		インパクトがある
		距離感

※ 構築された評価語から一部抜粋

5. 結論

「ライブ感」を対象として評価項目の抽出実験において，判別プロシージャと自動評価語構築技術により，クラウドソーシングから取得するデータの品質を向上させ，手動による構築と遜色ない評価項目を構築できることが分かった．これにより従来評価項目の集約には多大な時間がかかった上，多分に分析者の主観が混入することが問題視されていたが，客観的で、かつ、省力化を図ることができる手法として活用が期待される．

参 考 文 献

［1］ Hashimoto S, Yamada A, Nagata N:A quantification method of composite impression of products by externalized evaluation words of the appraisal dictionary with review text data. Int'l J. Affective Engineering, 18(2), 59-65, 2019.

［2］ 遠藤 ルッカス良, 馬場 雪乃, 鹿島 久嗣:感性評価に基づく最適化に対するクラウドソーシングの適用, 人工知能学会全国大会論文集, JSAI2017 (0), 4O2OS17b3-4O2OS17b3, 2017.

［3］ Buhrmester, M., Kwang, T., & Gosling, S. D.: Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?, Perspectives on Psychological Science, 6(1), 3–5, 2011.

［4］ Crump MJC, McDonnell JV, Gureckis TM: Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research., PLoS ONE 8(3): e57410, 2013.

［5］ Yoshimasa Majima,The Feasibility of a Japanese Crowdsourcing Service for Experimental Research in Psychology,Sage Open,2158-2440,SAGE Publications, 7, 1, 215824401769873, 2017.

［6］ 三浦 麻子, 小林 哲郎, : オンライン調査モニタの Satisfice に関する実験的研究, 社会心理学研究, 31 巻, 1 号, p. 1-12, 2015.

［7］ Krosnick, J. A.: Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5(3), 213–236, 1991.

［8］ 中井 智己, 宮崎 啓, 山下 大貴, 高崎 祐哉, 垂水 浩幸: 評価グリッド法を用いた音楽ライブにおける評価構造の抽出, 香川大学工学研究科, 23 号, 2017.

［9］ 小塩 真司, 阿部 晋吾, Cutrone Pino: 日本語版 Ten Item Personality Inventory（TIPI-J）作成の試み, パーソナリティ研究, 21 巻, 1 号, p. 40-52, 2012.

［10］ Majima Yoshimasa, Nishiyama Kaoru, Nishihara Aki, Hata Ryosuke: Conducting Online Behavioral Research Using Crowdsourcing Services in Japan, Frontiers in Psychology, 8, 2017 .