聴覚刺激に対する生理信号・音響情報・言語報告を統合した感情概念形成モデルの検討

Study of Emotion Concept Formation by Integrating Audio, Physiology, and Word Information

植村 奏[1*]　谷内 洋介[1]　日永田 智絵[1]　池田 和司[1]
Minato Uemura[1]　Yosuke Taniuchi[1]　Chie Hieida[1]　Kazushi Ikeda[1]

[1] 奈良先端科学技術大学院大学
[1] Nara Institute of Science and Technology

Abstract: Emotion understanding in human agent interaction requires models that adapt to emotions constructed through individual experience. This study proposes a computational model grounded in constructionist emotion theory that forms emotion concepts by integrating physiological signals, auditory features, and verbal reports during auditory stimulation. Auditory stimuli were presented to 36 participants, and cardiac activity, electrodermal activity, and subjective emotion reports were collected. Emotion concepts were modeled as latent topics using multilayered Multimodal Latent Dirichlet Allocation (mMLDA). Evaluation using the Rand index showed that models integrating multiple physiological signals outperformed single-modality and random baselines, and that adding linguistic information further improved alignment with subjective emotional responses.

1　はじめに

2000 年頃から機械知能に感情知能を組み込むべきだと指摘されてきたが [1]，現代においても十分に実現されているとは言い難く，感情知能実現のための感情の形成メカニズムを考慮した計算モデルの開発に注目が集まっている．近年の研究では，感情の形成メカニズムとして，構成主義的情動理論 [2] が提案され，内受容感覚（内臓などの身体内部の感覚）と外受容感覚（五感などの身体外部の感覚）の統合により感情が形成され，各感覚の対応関係を媒介する枠組みとして感情概念が存在すると考えられている．こうした理論を利用し，計算モデルを構築することで，感情知能実現に近づくことが期待できる．

先行研究として，弦牧ら [3] は，人に視覚刺激を提示する実験を通して，内受容感覚として生理信号，外受容感覚として視覚刺激特徴および言語情報を統合し，感情概念を形成する計算モデルを提案した．計算モデルは確率的生成モデルの一種である多層マルチモーダル LDA（mMLDA）[4] を用いており，モデルが形成した感情概念は人の主観報告と 7 割程度一致した．しかしながら，弦牧らの研究で扱われている外受容感覚は視覚刺激に限定されており，感情概念形成において，外受容感覚の種類が果たす役割については十分に検討されていない．

そこで本研究では，外受容感覚の中でも聴覚刺激に焦点をあて，感情概念形成モデルを構築する．視覚と聴覚は，前者が空間的に比較的安定した構造として処理されやすいのに対し，後者は時間的・逐次的に展開する情報として処理されるという点で，処理様式が大きく異なることが知られている．加えて，注意・予測の観点からも，視覚では空間的注意が，聴覚では時間的期待がより顕著に機能することが報告されている [5]．以上のように，視覚と聴覚では情報の構造や処理様式が大きく異なるため，感情概念がどのような手がかりに基づいて形成されるかについてもモダリティによって異なる可能性がある．実際に，聴覚刺激は，必ずしも明確な対象や意味内容を伴わない場合であっても，主観的な感情体験や生理反応を喚起し得ることが報告されている [6]．さらに，同一の音響刺激に対しても，個人の過去経験や置かれた文脈によって，喚起される感情の意味づけが大きく異なることが報告されている [7], [8]．

具体的なアプローチとして，本研究では，人に聴覚刺激を提示し，内受容感覚として生理信号，外受容感覚と

*連絡先：植村　奏，奈良先端科学技術大学院大学，
〒 630-0192 奈良県生駒市高山町 8916-5
uemura.minato.uk7@naist.ac.jp

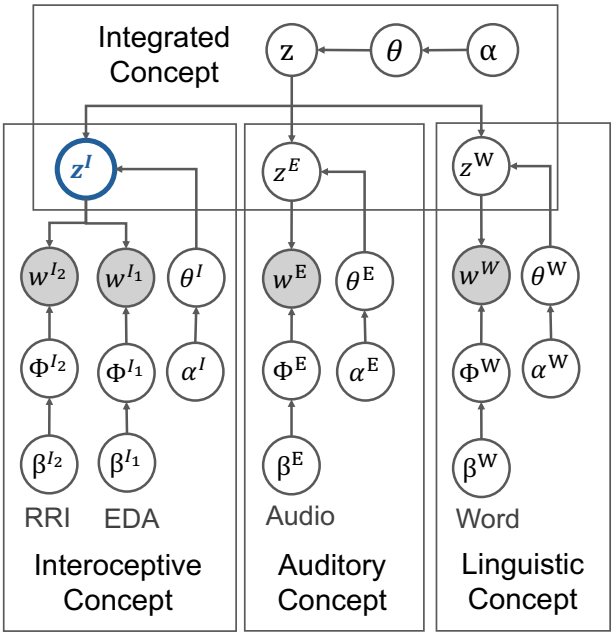


図 1: 感情概念形成のグラフィカルモデル

して聴覚刺激特徴および言語情報を取得し，mMLDA を用いて統合・学習し，主観的感情報告と整合的なカテゴリが形成されるかを評価した．

2　感情概念形成モデル

本提案モデルは mMLDA[4] により構成される．mMLDA とは，LDA[9] や LDA をマルチモーダル情報に拡張したマルチモーダル LDA（MLDA）[10] を，階層状に重ねることで複数の概念間の関係を確率的に表現するモデルである．ここでの概念とは異なる種類の入力を範疇化することで得られるカテゴリであり，LDA や MLDA はそうしたカテゴリを表現する生成モデルである．mMLDA は，各情報に対応する下位層として複数の LDA と MLDA を用意し，それらを統合する上位層として MLDA を配置することで，下位層で表現されたカテゴリ，すなわち概念間の関係性を表現する．

mMLDA を採用する理由は，本研究が目的とする感情概念形成を，高い予測性能や再構成精度の獲得ではなく，複数モダリティにまたがる概念間の対応関係を解釈可能な確率構造として記述する問題として位置づけているためである．

深層生成モデルや VAE 系手法は，高次元データから表現を学習し，再構成や予測に有効な潜在表現を獲得できる点で有力である．一方で，本研究の目的は性能最適化それ自体ではなく，生理信号や音響，言語といった異質なモダリティがどの概念単位で対応づけられているかを被験者ごとに解釈可能な形で記述することである．

また，連続潜在変数を用いる確率的生成モデルでは，観測分布の近さに基づくカテゴライズが主となり，概念の混合や意味的曖昧性を明示的に扱うことが難しい．

これに対し，mMLDA は，各モダリティに対応する下位層で概念を形成し，それらの関係性を上位層で統合するという階層構造を持つため，外受容感覚・内受容感覚・言語的意味づけといった異なる種類の概念がどのように結びついて感情概念が構成されるかを，トピックという解釈可能な単位で表現できる．

感情概念形成モデルのグラフィカルモデルを図 1 に示す．図中の z は統合概念を表し，z^I，z^E，z^W はそれぞれ下位概念に相当する内受容感覚，聴覚，言語カテゴリである．また，w^{I_1}，w^{I_2}，w^E，w^W はそれぞれ，生理信号，聴覚，言語に関する観測情報である．θ^*，ϕ^* は多項分布のパラメータであり，α^*，β^* はこれらの事前分布であるディリクレ分布のハイパーパラメータを表す．本研究では，カテゴリ数は全て 4 とし，ハイパーパラメータ α^* ，β^* はすべて 1.0 とした．

本研究では，感情概念が内受容感覚カテゴリ z^I に表現されると仮定する．この内受容感覚カテゴリはモデル内において，統合カテゴリを介して，聴覚および言語カテゴリからの情報がトップダウンに付与されることで形成される．この枠組みは，構成主義的情動理論において，感情概念が内受容感覚を基盤としつつ，外受容感覚に由来する情報を段階的に統合することで構成されるとする見方と整合的である．また，内受容感覚が感情体験の中核的役割を果たすとする近年の感情研究の知見にも一致する [11]．

3　実験

3.1　データ収集

本研究では，聴覚刺激として感情を喚起する音響刺激を人に提示する実験が実施され，提示中の生理信号および音響に対する印象が収集された．被験者は 36 名で（20 代から 40 代の男性 18 名，女性 18 名），各被験者に対し，60 個の音響刺激が提示された．音響刺激は，International Affective Digitized Sounds（IADS）[12] より，60 個の刺激（各刺激に付与された Valence と Arousal の評価値が広く分布するよう選定）が使用された．生理信号の計測には，E4 wristband[13] および myBeat WHS-1 を用いられ，皮膚電気活動（EDA）および心拍波形が取得された（EDA：4 Hz，心拍波形：128 Hz）．音響に対する印象収集は自由な単語発話と Self-AssessmentManikin（SAM）[14] が用いられた．単語発話は，Google Speech Recognition[15] を用いることで，テキスト形式で収集された．SAM は感情の主観的な測定尺度であり， Valence（快度），Arousal（覚醒

度），Dominance（支配度）の 3 つの項目に関して，イラストを用いた 9 件法で構成される．Dominance は不安定な尺度であることが指摘されているため，本研究では Valence と Arousal の 2 項目が解析対象となった．

1 サイクルは 75 秒で構成され（音響提示 6 秒，単語発話 50 秒，SAM 回答 15 秒，待機時間 4 秒），10 サイクルを 1 セッションとして，計 6 セッション実施された．各セッション間には 3 分間の休憩が設けられた．実験前には，安静時測定および 1 セッション分の練習が行われ，実験後には，個人特性に関するアンケートが実施された．実験全体の所要時間は，1 人あたり約 2 時間半から 3 時間であった．本調査は，奈良先端科学技術大学院大学倫理審査委員会の承認を得て実施した．

3.2 モデル学習

本研究では，全被験者のうち，生理信号に欠損のない 27 名（男性 14 名，女性 13 名）のデータを学習対象とした．mMLDA の観測情報 w^*（図 1）として，生理信号および音響刺激，単語発話を用いた．以下より，各前処理について説明する．

内受容感覚モジュールの観測情報として，音響提示時 6 秒間の生理信号を使用した．EDA は 0.05 Hz 以下をカットするハイパスフィルターを適用したのちに，区間 1 秒の移動平均で平滑化し，標準化（z スコア化）を行った．心拍波形は R-R 間隔（RRI）を算出し，akima 補間（サンプリング間隔：128 Hz）[16] によって等間隔に変換後，標準化を行なった．その後，標準化した生理信号を Vector Quantized-VAE（VQ-VAE）[17] を用いて特徴抽出し，128 次元の頻度情報に変換した．ここで，VQ-VAE は安静時区間のデータで学習したモデルを用いた．

聴覚モジュールの観測情報には，被験者に提示した 60 個の音響刺激を用いた．各音響刺激に対してメルフィルタバンク [18]（80 次元）を適用し，標準化を行った．その後，標準化した音響刺激を VQ-VAE[17] を用いて特徴抽出し，128 次元の頻度情報に変換した．ここで，VQ-VAE は IADS と類似の音響データである ESC-50[19] の 2000 サンプルを対象として事前学習したモデルを用いた．

言語モジュールの観測情報には，被験者の音響刺激に対する印象の単語発話を用いた．全被験者の単語発話に対して，テキスト誤変換の修正を行い，学習済みの Word2Vec[20] を用いて，各単語を 200 次元の埋め込み表現に変換し，中心化を行なった．その後，Deep Compositional Code Learning（DCCL）[21] を用いて圧縮し，24 次元の頻度情報に変換した．DCCL は，全被験者の全単語発話を対象として学習したモデルを用いた．

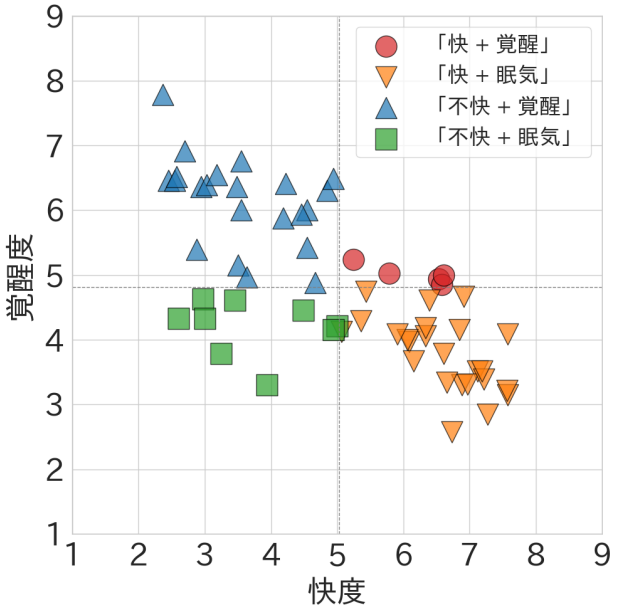


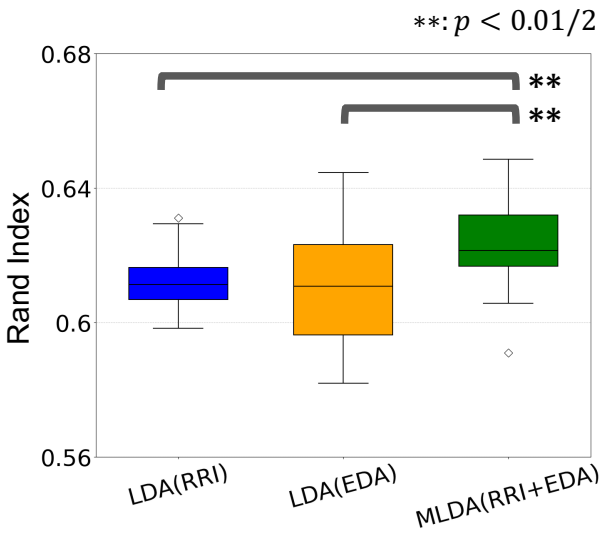
図 2: 音響ごとの SAM 評価重心およびカテゴリ

上記のデータを用い，図 1 の mMLDA（FULL）のモデルに加えて，情報を欠損させた 5 つのモデルを学習した．具体的には，生理信号単体の LDA（EDA）および LDA（RRI），EDA と RRI を用いた MLDA（RRI＋EDA），その MLDA に言語情報のみを追加した mMLDA（WORD），音響情報のみを追加した mMLDA（AUDIO）の 5 つとした．これらのモデルを比較することにより，各情報についての議論が可能となる．なお，モデル学習は被験者ごとに行った．

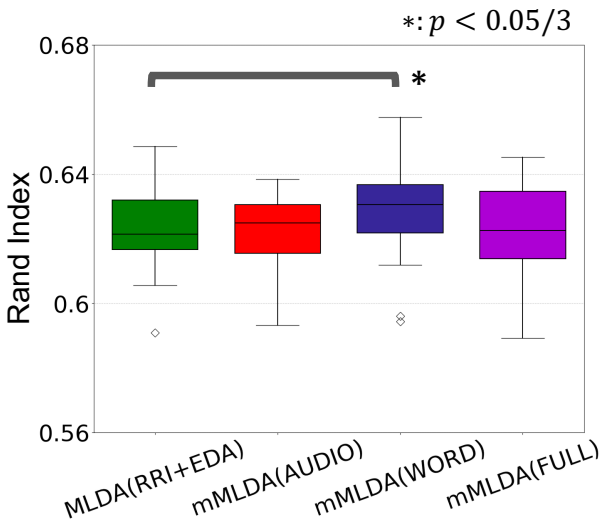
3.3 検証方法

モデルが学習した内受容感覚のカテゴリ z^I と人の主観的な感情報告の類似度をランド指数（RI）を用いて検証した．RI は 2 つのカテゴリの当てはまりの良さを定量化する指標であり，0 から 1 の範囲の値をとる．0 は 2 つのカテゴリが全く一致しておらず，1 は 2 つのカテゴリが完全に一致していることを示す．主観的な感情報告に関しては，SAM の回答を元にしたカテゴリを使用した．具体的には，縦軸を Valence の全被験者平均，横軸を Arousal の全被験者平均とし，データ重心を原点として各象限を 1 カテゴリとすることで音響を「快+覚醒」「快+眠気」「不快+覚醒」「不快+眠気」の 4 カテゴリにラベリングした（「快+覚醒」: 5 個，「快+眠気」:24 個，「不快+覚醒」:22 個，「不快+眠気」:9 個，図 2）．また，本研究においては被験者間で比較可能な基準を与えるため，SAM のカテゴリは全被験者の重心により決定した．

RI をチャンスレベルと比較することで，偶然カテゴ



(a) LDA および MLDA



(b) MLDA および mMLDA

図 3: 全データにおける各モデルのランド指数

リが一致する確率をこえてモデルが構成したカテゴリが主観的な感情報告を表現しているかを検証した．チャンスレベルは，SAM カテゴリの出現回数を固定したまま，カテゴリをランダムに並べ替えた割当を多数回生成し，その平均をモンテカルロ推定した（5000 回）．結果，チャンスレベルは全データにおいて 0.56 となった．また，モデル間の差はウィルコクソンの符号付き順位検定を用いて検定した．有意水準は 0.05 とし，ボンフェローニ法で補正した．

4 結果

4.1 チャンスレベルとの比較

本研究では聴覚刺激に対する感情概念形成において，生理信号，言語情報，音響情報が果たす役割を検証した．評価には，モデルが推定したカテゴリと主観的情動評価（SAM）との一致度を測る指標として RI を用いた．すべてのモデルにおいて RI はチャンスレベル（0.56）を明確に上回った（図 3）．

4.2 LDA と MLDA の比較

LDA（RRI）と LDA（EDA）を MLDA（RRI+EDA）と比較した（図 3(a)）．その結果, MLDA は LDA（RRI）および LDA（EDA）に対して有意に高い RI を示した（vs. LDA（RRI）: $p = 9.78 \times 10^{-4} < 0.01/2$，vs. LDA（EDA）: $p = 4.31 \times 10^{-4} < 0.01/2$）．この結果は，RRI と EDA を統合することで，主観的感情報告との一致度が高まることを示す．

4.3 MLDA と mMLDA の比較

MLDA と mMLDA の比較を図 3(b)に示す．mMLDA（AUDIO）と MLDA の比較では有意差は認められなかった（$p = 5.97 \times 10^{-1} > 0.05/3$）．mMLDA（WORD）と MLDA の比較では，有意差が確認された（$p = 4.24 \times 10^{-3} < 0.05/3$）．生理信号と言語情報，音響情報をすべて統合した mMLDA（FULL）は，MLDA との比較において有意差が認められなかった（$p = 6.07 \times 10^{-1} > 0.05/3$）．したがって，言語情報を階層的に統合することで，本モデルでの聴覚刺激に対する感情概念の主観的感情報告との一致度が高まることを示された．

5 考察

5.1 聴覚刺激について

本研究では，聴覚モダリティのみを階層的に統合した mMLDA（AUDIO）は，MLDA に対して有意差がみられなかった．考えられる要因として，前処理により刺激の時間構造に関する情報が縮約された結果，各被験者が 6 秒間の音響刺激のどの時点に注意を向けたかという情報が失われた可能性が挙げられる．本研究が対象とした聴覚刺激は，時間的に展開する一過性の情報であり，視覚刺激のように空間的に保持される構造とは異なる．この性質により，同一刺激であっても，被験者が注意を向けた時間区間や刺激中に形成された期待の違いによって，抽出される手がかりが変化しやすいと考えられる．

5.2 言語情報の寄与

言語モダリティのみを階層的に統合した mMLDA（WORD）が MLDA より有意に高い RI を示した．考えられる要因として，単語の埋め込み表現を利用したことで，各単語が意味空間上に配置され，聴覚刺激に対する被験者固有の解釈や文化的背景が言語モダリティの追加によってモデルに明示的に取り込まれた可能性が挙げられる．

聴覚刺激は必ずしも明確な対象や意味内容を伴わず，同一刺激であっても，被験者の経験や置かれた文脈によって解釈が大きく変化し得ることが知られている．このような特性を持つ刺激に対して，言語情報は主観的解釈を明示的な記号表現として与える手段となり得る．本研究で観測された一致度の改善は，こうした言語情報の特性がカテゴリ形成に寄与した結果である可能性を示唆している．

5.3 マルチモダールモデルの特性

聴覚モダリティ，言語モダリティの両者を追加した mMLDA（FULL）は，RI 向上という点において MLDA に対して有意差がみられなかった．この結果は，聴覚モダリティの追加が無意味であることを示すものではなく，本研究の設定においては，一致度指標（RI）としてその効果が顕在化しなかったことを意味する．具体的には，トピック数を固定した設定や前処理による情報縮約により，聴覚モダリティの寄与が小さく評価された可能性が考えられる．また，言語モダリティがすでに SAM カテゴリと整合する手がかりを十分に含んでいたため，聴覚モダリティの追加が冗長になった可能性も否定できない．

複数モダリティを統合するモデルにおいては，各モダリティが提供する情報の相補性や冗長性が，性能向上の程度に大きく影響することが指摘されている [22]．本研究の結果は，言語モダリティが主観感情に対して内受容感覚カテゴリが持たない情報を提供した一方で，聴覚モダリティは相補性を有していなかったことを示唆する．これらの点については，特徴表現や統合構成を変更した条件での再検証が必要である．

5.4 限界と今後の課題

本研究では，カテゴリ数を 4 に固定したため，被験者ごとのデータ（生理信号や単語発話）に応じたカテゴリ数の探索の検討が必要である．また，聴覚カテゴリの観測情報について聴覚刺激の時間構造をより表現した特徴抽出を導入する余地がある．さらに，本稿では評価が RI に限定されており，未観測モダリティ予測や mMLDA のカテゴリ可視化など，異なる観点からの検証を追加する必要がある．

6 おわりに

本研究では構成主義的情動理論にもとづき，人への聴覚刺激提示時の生理信号，音響特徴，言語情報を統合し，感情概念を形成する計算モデルの構築を試みた．多層マルチモーダル LDA により形成したカテゴリと主観的感情評定との一致度をランド指数で評価した結果，複数の生理信号を統合したモデルは単一の生理信号およびランダム水準であるチャンスレベルを有意に上回り，さらに言語情報を加えることで一致度が向上した．これにより，本研究にて構築した感情概念形成モデルは，聴覚刺激に対する主観感情を表現できることが示唆された．本研究により，感情概念が複数の感覚情報の統合として計算的に構成され得る可能性が示された．本モデルが，人とエージェントの相互作用において，聴覚刺激に基づく主観的感情の理解と推定を行うための計算的枠組みとして，今後の HAI 研究に寄与することを期待する．

謝辞

本研究は JST ACT-X JPMJAX21AL および JSPS 科研費 JP24K17239，JP25H00579 の助成を受けたものである．

参考文献

[1] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 23, No. 10, pp. 1175–1191, 2001.

[2] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, Vol. 12, No. 1, pp. 1–23, 2017.

[3] Kazuki Tsurumaki, Chie Hieida, and Kazuki Miyazawa. Study of emotion concept formation by integrating vision, physiology, and word information using multilayered multimodal latent dirichlet allocation. *IEEE Transactions on Affective Computing*, 2025.

[4] Muhammad Attamimi, Muhammad Fadlil, Kasumi Abe, Tomoaki Nakamura, Kotaro Funakoshi, and Takayuki Nagai. Integration of various concepts and grounding of word meanings using multi-layered multimodal lda for sentence generation. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2194–2201. IEEE, 2014.

[5] A. Wilsch, et al. Spatial attention and temporal expectation exert differential effects on visual and auditory discrimination. *Journal of Cognitive Neuroscience*, 2020.

[6] Julie Kirwan, Deniz Başkent, and Anita Wagner. The time course of the pupillary response to auditory emotions in pseudospeech, music, and vocalizations. *Trends in hearing*, Vol. 29, p. 23312165251365824, 2025.

[7] Patrik N Juslin and Daniel Västfjäll. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, Vol. 31, No. 5, pp. 559–575, 2008.

[8] Laurie M Heller and Jessica M Smith. Identification of everyday sounds affects their pleasantness. *Frontiers in psychology*, Vol. 13, p. 894034, 2022.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[10] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi. Grounding of word meanings in multimodal concepts using lda. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3943–3948. IEEE, 2009.

[11] 寺澤悠理, 梅田聡. 内受容感覚と感情をつなぐ心理・神経メカニズム. 心理学評論, Vol. 57, No. 1, pp. 49–66, 2014.

[12] Margaret M Bradley and Peter J Lang. The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual. Technical report, Technical report B-3. University of Florida, Gainesville, Fl, 2007.

[13] Hans Stuyck, Leonardo Dalla Costa, Axel Cleeremans, and Eva Van den Bussche. Validity of the empatica e4 wristband to estimate resting-state heart rate variability in a lab-based context. *International Journal of Psychophysiology*, Vol. 182, pp. 105–118, 2022.

[14] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, Vol. 25, No. 1, pp. 49–59, 1994.

[15] Aqeel Rajput, Dilbag Singh, and Nishant Kumar. Comparing speech recognition systems (microsoft api, google api and cmu sphinx). *IOSR Journal of Computer Engineering (IOSR-JCE)*, Vol. 19, No. 3, pp. 20–24, 2017.

[16] Hiroshi Akima. A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the ACM (JACM)*, Vol. 17, No. 4, pp. 589–602, 1970.

[17] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, Vol. 30, , 2017.

[18] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, Vol. 28, No. 4, pp. 357–366, 1980.

[19] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM Press.

[20] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会 (NLP2016), 東京, 日本, March 2016. 言語処理学会.

[21] Raphael Shu and Hideki Nakayama. Compressing word embeddings via deep compositional code learning. *arXiv preprint arXiv:1711.01068*, 2017.

[22] Daheng Wang, Tong Zhao, Wenhao Yu, Nitesh V Chawla, and Meng Jiang. Deep multimodal complementarity learning. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 34, No. 12, pp. 10213–10224, 2022.