正例・負例ペアの定義不要なクロスモーダル関係の定量化 –歌詞とダンスのペアデータを対象として–

渡邉 研斗[1]　後藤 真孝[1]
[1]産業技術総合研究所
{kento.watanabe,m.goto}@aist.go.jp

概要

対照学習をはじめとする深層学習手法により，クロスモーダル関係を End-to-End に定量化する研究が広く行われている．しかし，これらの手法は，学習時に正例・負例ペアが既知であることを前提としている．一方で，実データにおいては，どのペアが正例または負例に該当するかを事前に定義できない場合が多い．本研究では，正例・負例ペアを定義せずに，クロスモーダル関係をデータ駆動的に定量化する手法を提案する．歌詞とダンスのペアデータを対象とした分析により，多様な関係性を明らかにするとともに，歌詞に基づくダンス検索タスクにおいて，従来手法を大幅に上回る性能を達成した．

1　はじめに

近年，対照学習や Vision–Language Model などに代表される深層学習手法の発展により，テキスト・画像・動画・音声・身体情報といった異なるモダリティ間の関係性を End-to-End に学習するクロスモーダル学習が盛んに研究されている [1]．これらの手法は，複数モダリティを共通の表現空間に写像することで，高精度な検索や対応付けを実現してきた．

一方で，これらの手法は学習時にモデルへ与えられるペアが，何らかの意味的関係を持つ「正例のペア」であることを前提としている．しかし，実世界に存在するクロスモーダルなペアデータにおいては，全てのペアが正例であるとは限らない．たとえば，「商品の紹介文と購買ログ」や「動画概要文と動画本編」といった性質の異なるクロスモーダルデータでは，一部のペアは関係性を持つ可能性はあるものの，すべてのペアが一様に関係性を持つとは考えにくい．また，どのペアが関係性を持つかは手がかりなしに自動推定することは容易ではない．

そこで本研究では，データ中に局所的に存在するクロスモーダルペアの関係性を，正例・負例ペアを事前に定義せずに定量化する手法の構築を目的とする．その代表例として，関係性がデータ中の一部にしか存在しないと思われる「**歌詞とダンス**」のペアデータに着目する．ダンスの振り付けは歌詞から影響を受けることが指摘されており，たとえば「天国」という単語が歌われる際に，腕を天に向けて伸ばす動作が用いられる場合がある [2]．一方で，全てのダンスが歌詞に由来するわけではなく，多くのリズミカルな動作は歌詞よりも音楽の音響的特徴に誘発されていると考えられる．このように，歌詞とダンスの関係性は局所的にしか存在せず，どの歌詞とダンスが対応関係を持つかはデータ中で自明ではないため，これらの関係性を単純な End-to-End 学習で捉えることは困難である．

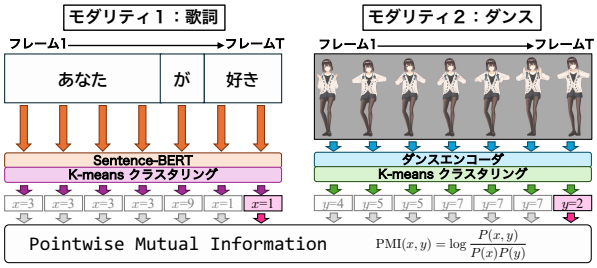

図 1　歌詞とダンスの関係性の定量化の概要．

この技術的困難に対して，我々は歌詞とダンスの関係性を数値化するデータ駆動型手法 [3] を提案する．本手法の根幹は，我々独自の「複数の楽曲にわたり同じ歌詞表現と同じダンスが繰り返し共起すれば，両者は意味的に対応している」とみなす考え方であり，本研究ではこれを「**横断共起仮定**」と名付ける．たとえば異なる楽曲で「ジャンプ」と歌われる瞬間にジャンプ動作が観測されるなら，この歌詞とダンスは対応しているとみなせる（ただし，「ジャンプ」に対して必ずジャンプ動作をするわけではなく，対応の強さは共起の度合いで評価する）．

この横断共起仮定を実装に落とし込むために，

我々は図 1 のように，まず歌詞特徴ベクトルとダンス特徴ベクトルをクラスタリングして離散的なコードブック系列へ変換し，続いて歌詞コードブックとダンスコードブックの共起頻度から Pointwise Mutual Information (PMI) [4] を算出することで，両者が強く関係する瞬間（フレーム）を定量的に抽出する．提案手法を大規模な歌詞とダンスのペアデータセットに適用したところ，高い PMI を示すフレームが多数検出された．たとえば「好き」が歌われる瞬間には手でハート形を作るジェスチャーや，「時計」と歌われる瞬間には腕を時計の針に見立てる動作などが対応していることが確認できた．

さらに，本研究ではフレーム単位の関係性だけでなく系列全体の関係性を定量化するために，PMI を置換コストとして用いる編集距離 **XMIED (Cross-Modal Mutual-Information Edit Distance)** [3] を提案する（図 2）．XMIED により，歌詞コードブック系列とダンスコードブック系列の距離を直接定量化できる．XMIED を検索スコアとして用い，入力歌詞に対応するダンスを検索する実験を行った結果，提案手法は従来の対照学習ベースの検索手法を大幅に上回る性能を示した．

2　歌詞とダンスのペアデータ

本研究では，MikuMikuDance (MMD) コミュニティで制作された 30FPS の 3D 骨格のダンスデータを合計 1,000 件（53 時間分）収集した．次に，各ダンスに対応する楽曲音源および歌詞テキストを別途収集し，強制アライメント手法 [5] を用いて，歌詞中の各単語が歌唱される時刻区間を推定した．この結果，ダンスの各フレームに対して歌詞の単語が対応付けられる．なお，歌詞が存在しないフレームにはプレースホルダとして [PAD] トークンを割り当てた．

時刻同期の後，ダウンビート推定手法 [6] を用いて，楽曲音源・歌詞・ダンスを小節単位に分割した．なお，1 秒未満の短い小節は情報量が乏しいと判断して除外した．最終的に得られたデータセットは 119,691 小節から構成されており，そのうち 92,723 小節が歌詞付きのダンス，26,968 小節が歌詞が存在しないダンスである．

3　クロスモーダル関係の定量化

本研究では，正例・負例ペアを事前に定義できないクロスモーダルデータに対して，関係性を定量化する手法 XMIED (Cross-Modal Mutual-Information Edit Distance) を提案する．XMIED は，異なるモダリティから得られた系列データを対象とした編集距離として定義され，系列長の違いや時間的順序を考慮しながら，クロスモーダルな関係性を比較できる点に特徴がある．図 2 に本手法の概要を示す．

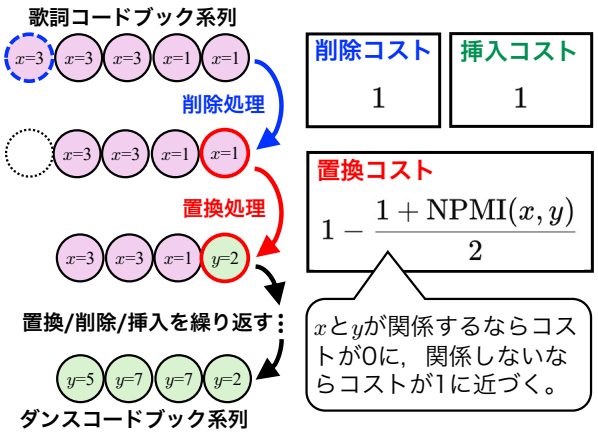

図 2　Cross-Modal Mutual-Information Edit Distance の概要．

XMIED の基本的な考え方は，異なるモダリティから得られた系列データを離散的な記号系列へ変換し，それらを文字列とみなして編集距離を計算する点にある．具体的には，歌詞およびダンスの各フレームから得られた特徴ベクトルをそれぞれ k-means クラスタリング [7] によって離散化し，コードブック系列 $x_1,\ldots,x_M$ と $y_1,\ldots,y_N$ に変換する．編集距離では，置換・挿入・削除の各操作にコストを与え，一方の系列を他方へ変換する最小コストを系列間の距離として定義する．この枠組みを異なるモダリティ間に適用するためには，歌詞コードブック x とダンスコードブック y の間の置換コストを適切に定義する必要がある．すなわち，異なるモダリティに属するコードブック同士の関係性をどのように数値化するかが，本手法の鍵となる．

本研究では，この置換コストの設計に，「異なるモダリティに属する要素が，複数のデータにわたって一貫して共起する場合，それらは意味的に対応しているとみなす」という独自の横断共起仮定を用いる．具体的には，本研究では歌詞コードブック x とダンスコードブック y の関係性を，Normalized Pointwise Mutual Information (NPMI) [8] に基づいて定量化する．NPMI は

$$\mathrm{NPMI}(x,y)=\frac{\log P(x,y)-\log P(x)-\log P(y)}{-\log P(x,y)}$$

で定義され，x と y の共起確率 $P(x,y)$ および周辺確率 $P(x)$，$P(y)$ に基づいて，両者の結び付きの強さを

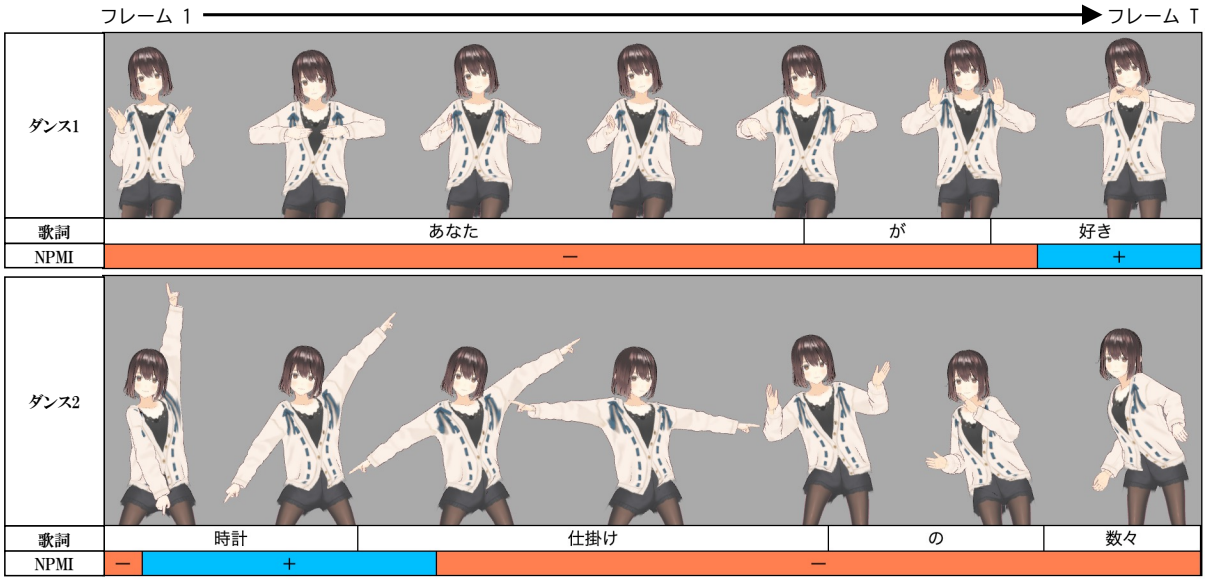

図 3　正の NPMI を持つ歌詞とダンスペアの例．負の NPMI はオレンジ色のボックスおよびマイナス記号で，正の NPMI は青色のボックスおよびプラス記号で示される．図中の歌詞は，研究目的の引用として，ダンス 1 は「エイリアンエイリアン（作詞：ナユタン星人）」，ダンス 2 は「Love Timer（作詞：emon (Tes.)）」より引用．

評価する指標である．本研究では，横断共起仮定を反映するため，これら確率の計算に「**共起が観測された楽曲数**」を用いる．全楽曲数を S とすると，

$$P(x,y) = \frac{\#(x,y)}{S}, \quad P(x) = \frac{\#(x)}{S}, \quad P(y) = \frac{\#(y)}{S}$$

と定義する．ここで $\#(x,y)$ は，同一フレーム内でコードブック x と y が共起した楽曲の数を表し，$\#(x)$ および $\#(y)$ は，それぞれ x または y が少なくとも一度出現した楽曲の数を表す．なお，共起が 1 曲でしか観測されない場合（$\#(x,y) = 1$）は，「複数楽曲にわたって繰り返し共起するペアのみを対応とみなす」という横断共起仮定を満たさないため，そのペアの NPMI は 0 として扱う．この NPMI に基づき，歌詞コードブック x とダンスコードブック y の置換コストを $1 - (1 + \mathrm{NPMI}(x,y))/2$ と定義する．挿入および削除のコストを 1 に固定した上で，この置換コストを用いて編集距離を計算したものを XMIED とすることで，歌詞系列とダンス系列のクロスモーダルな関係性を，系列レベルで定量化することが可能となる．

以上のように XMIED は，各モダリティのエンコーダの共同学習や追加学習を必要とせず，既存の学習済みエンコーダをそのまま利用できる点に特長がある．また，コードブックの頻度計算という単純な枠組みでありながら，正例・負例ペアを事前に定義できないクロスモーダルデータにも適用可能であり，高い汎用性と再利用性を備えた手法である．

4　実験

実験では，1,000 曲分のペアデータを，訓練・開発・評価用に 8:1:1 の比率で分割した．歌詞の特徴ベクトルとして，多言語対応の sentence-BERT [9] の最終層から出力される単語ベクトルを用いた．これにより，部分的に英語を含む日本語楽曲の歌詞も扱うことができる．ダンスの特徴ベクトルの算出には，本研究が独自に設計したダンスエンコーダを用いた．このエンコーダは，人体骨格の関節をノードとするグラフ構造として表現し，Graph Transformer [10] を用いたオートエンコーダとして学習される．本研究で用いたエンコーダの構造および学習パラメータは付録 A に示す．

4.1　歌詞とダンスの関係性分析

本節では，フレーム単位のクロスモーダル関係を表す NPMI が，実際のデータにおいてどのような関係を捉えているかを定性的・定量的に分析する．

図 3 は，NPMI が正となった小節から抜粋した例を示しており，歌詞とダンスの対応関係を視覚的に確認できる．たとえばダンス 1 では，感情語「好き」が歌われる瞬間に，複数の楽曲においてハート形のジェスチャーが観測されている．またダンス 2 では，「時計」など時間に関連する単語に対して，腕を用いて時計の針を模す動きが対応している．その他の対応例については付録 B に示す．これらの事例

表 1　正の NPMI を持つ曲・小節・フレームの件数と割合.

分析単位	件数	割合 (%)
曲	781	78.10
小節	2,671	2.88
フレーム	14,431	0.30

は，提案手法が歌詞とダンスの直感的に妥当な対応関係を定性的に捉えていることを示している．

さらに本研究では，このような対応関係の出現頻度をデータ全体に対して定量的に評価する．表 1 は，正の NPMI が観測された楽曲数，小節数，およびフレーム数と，それぞれの割合をまとめたものである．その結果，78.1%の楽曲において少なくとも 1 箇所は正の NPMI が観測され，多くの楽曲で歌詞とダンスの対応が存在することが確認された．一方で，正の NPMI を示す小節は全体の 2.88%，フレームに至っては 0.3%にとどまる．このことから，歌詞とダンスの意味的な相互作用は楽曲全体に一様に存在するのではなく，限られた区間に局所的に現れることが定量的に示された．

4.2　歌詞に基づくダンス検索

本研究では，評価用データに含まれる歌詞を入力として，対応するダンスを検索するタスクを設定した．本タスクは，歌詞に基づくダンス振付の検索や制作支援などへの応用が考えられる．

比較手法として，以下の 3 手法を用意した．(1) **ランダム選択**：評価対象となるダンスをデータ集合からランダムに選択する手法．(2) **対照学習**：歌詞に基づくダンス検索の既存手法は存在しないため，散文テキストに基づくモーション検索の先行研究 [11, 12, 13, 14] で用いられている対照学習に基づく検索手法を比較対象として採用した．本手法では，歌詞ベクトルとダンスベクトルを共通の埋め込み空間に写像するエンコーダを対照学習により学習し，入力された歌詞とコサイン類似度が最も高いダンスを検索する．(3) **XMIED**：提案手法では，訓練データを用いてダンスエンコーダを学習し，コードブックサイズは開発用データでのチューニングにより，最も検索性能が高い設定を採用した．

評価指標には，検索タスクで標準的に用いられる Mean Reciprocal Rank (MRR) [15] を採用した．Precision@K や Recall@K といった指標も考えられるが，K の設定によって値が大きく変動し，手法間の厳密な比較が難しくなるため，本研究では K に依存しない MRR のみを用いた．

表 2　検索タスクにおける従来手法との比較.

比較手法	MRR ↑	1/MRR ↓
ランダム選択	0.00113	884
対照学習	0.00151	663
XMIED（提案手法）	**0.01905**	**53**

表 2 より，対照学習手法はランダム選択法をわずかに上回ったものの，提案手法はさらに高い性能であることがわかる．具体的には，提案手法の MRR が 0.01905 となり，これは評価用データの 8,611 小節に対して，提案手法が平均して正解ダンスを上位 53 位以内にランク付けしたことに相当する．

対照学習手法は「訓練データ中のすべての歌詞とダンスのペアを正例とみなす」という仮定に基づいているため，実際には無関係なペアまで強制的に結び付けられ，検索性能が頭打ちになる．これに対し，XMIED は「複数の楽曲にわたって繰り返し共起するペアが意味的に対応する」という横断共起仮定に基づいて，編集距離における置換コストを設計している．その結果，関係性の高いフレーム区間が強調され，無関係な区間の影響が抑えられる．この仮定の違いが MRR の差につながったと考えられ，横断共起仮定の妥当性を裏づける結果となった．

5　まとめ

本研究では，クロスモーダルペアデータにおいて，関係性がデータ中の一部のペアにしか存在しないという課題に着目した．対照学習に代表されるクロスモーダル学習手法は，学習時に与えられるペアを正例と仮定するため，このような状況への適用には限界がある．この問題に対し，本研究では「歌詞とダンスは，複数の楽曲にわたって繰り返し共起するときに意味的な関連がある」という横断共起仮定に基づき，両者の関係性を定量化する手法を提案した．具体的には，歌詞とダンスから得られた離散シンボルが共起する楽曲の頻度に基づいて置換コストを定義し，これを編集距離に組み込むことで，クロスモーダルな関係性を評価した．その結果，感情語と象徴的ジェスチャーの対応など，直感的に妥当な関係性を可視化できることを示した．さらに，歌詞に基づくダンス検索タスクにおいて，従来手法を上回る検索性能を達成した．今後は，様々なクロスモーダルペアデータに対しても本手法を適用し，その汎用性を検証していく予定である．

謝辞

本研究の一部は JST CREST JPMJCR20D4 と JST ACT-X JPMJAX25CU の支援を受けた.

参考文献

[1] Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. **ACM Computing Surveys**, Vol. 57, No. 7, pp. 167:1–167:34, 2025.

[2] Hayley Elizabeth Powell. Modern dance choreography: Beyond the movement an analysis between lyrics and movement: Can identities be developed through modern dance choreography? **Annual Review of Education, Communication & Language Sciences**, Vol. 16, No. 2, 2019.

[3] Kento Watanabe and Masataka Goto. A data-driven method for analyzing and quantifying lyrics-dance motion relationships. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025**, pp. 7901–7916, 2025.

[4] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. **Comput. Linguistics**, Vol. 16, No. 1, pp. 22–29, 1990.

[5] Tomoyasu Nakano and Masataka Goto. LyricListPlayer: A consecutive-query-by-playback interface for retrieving similar word sequences from different song lyrics. In **Proceedings of the 13th Sound and Music Computing Conference, SMC 2016**, pp. 344–349, 2016.

[6] Sebastian Böck, Florian Krebs, and Gerhard Widmer. Joint beat and downbeat tracking with recurrent neural networks. In **Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016**, pp. 255–261, 2016.

[7] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In **Proceedings of the eighteenth annual ACM SIAM symposium on Discrete algorithms**, pp. 1027–1035, 2007.

[8] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. **Proceedings of GSCL**, Vol. 30, pp. 31–40, 2009.

[9] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019**, pp. 3980–3990, 2019.

[10] Haocong Rao and Chunyan Miao. TranSG: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In **IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023**, pp. 22118–22128, 2023.

[11] Qing Yu, Mikihiro Tanaka, and Kent Fujiwara. Exploring vision transformers for 3D human motion-language models with motion patches. In **IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024**, pp. 937–946, 2024.

[12] Junpei Horie, Wataru Noguchi, Hiroyuki Iizuka, and Masahito Yamamoto. Learning shared embedding representation of motion and text using contrastive learning. **Artificial Life and Robotics**, Vol. 28, No. 1, pp. 148–157, 2023.

[13] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In **IEEE/CVF International Conference on Computer Vision, ICCV 2023**, pp. 9454–9463, 2023.

[14] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. In **European Conference on Computer Vision, ECCV 2022**, Vol. 13682 of **Lecture Notes in Computer Science**, pp. 358–374, 2022.

[15] Nick Craswell. Mean reciprocal rank. In **Encyclopedia of Database Systems**, p. 1703. 2009.

[16] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Affective body expression perception and recognition: A survey. **IEEE Transactions on Affective Computing**, Vol. 4, No. 1, pp. 15–33, 2013.

[17] Arthur Crenn, Rizwan Ahmed Khan, Alexandre Meyer, and Saïda Bouakaz. Body expression recognition from animated 3D skeleton. In **International Conference on 3D Imaging, IC3D 2016**, pp. 1–7, 2016.

[18] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. STEP: Spatial temporal graph convolutional networks for emotion perception from gaits. In **Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020**, pp. 1342–1350, 2020.

[19] Uttaran Bhattacharya, Nicholas Rewkowski, Pooja Guhan, Niall L. Williams, Trisha Mittal, Aniket Bera, and Dinesh Manocha. Generating emotive gaits for virtual agents using affect-based autoregression. In **2020 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2020**, pp. 24–35, 2020.

[20] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2Gestures: A transformer-based network for generating emotive body gestures for virtual agents. In **IEEE Virtual Reality and 3D User Interfaces, VR 2021**, pp. 160–169, 2021.

[21] Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 1382–1390, 2022.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **Proceedings of the 7th International Conference on Learning Representations, ICLR 2019**, 2019.

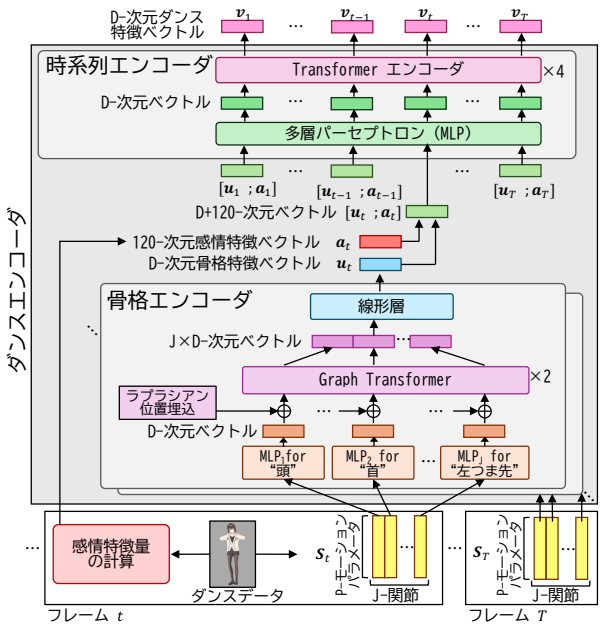

図 4　ダンスエンコーダの概要.

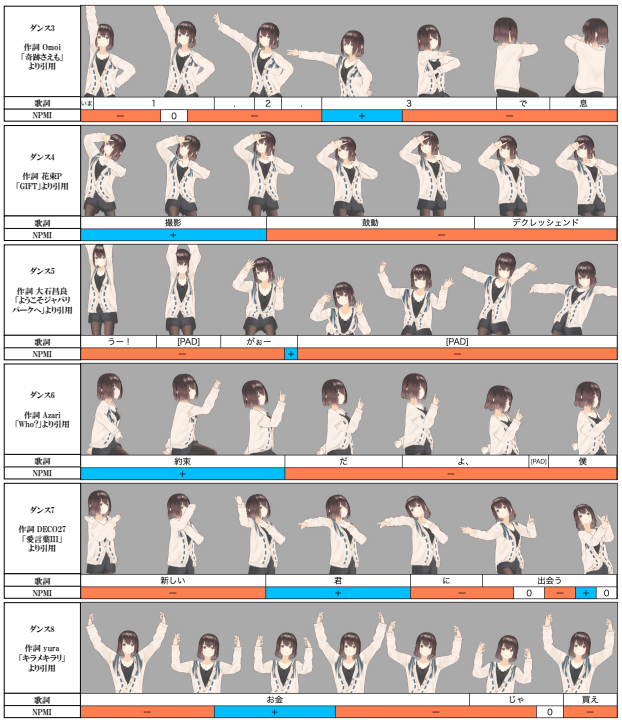

図 5　正の NPMI を持つ歌詞とダンスの例.

A　ダンスエンコーダの詳細

ダンスエンコーダは，図 4 に示す骨格エンコーダと時系列エンコーダで構成される．まず，フレーム t の骨格構造パラメータ $\boldsymbol{S}_t$ を入力とし，$\boldsymbol{S}_t$ に含まれる各関節の P 個の位置や角度情報パラメータを，各関節用の多層パーセプトロンで D 次元ベクトルへ変換する．変換後のベクトル列はラプラシアン位置埋め込みにより関節間の構造を考慮して Graph Transformer に入力される．得られた $J \times D$ 行列をひと繋ぎのベクトルへと連結した後に，線形層により骨格特徴ベクトル $\boldsymbol{u}_t \in \mathbb{R}^D$ を得る．同時に，関節間の体積・面積・長さ・曲率などの幾何学量を表した 120 次元の感情特徴ベクトル [16, 17, 18, 19, 20] を計算し，骨格特徴 $\boldsymbol{u}_t$ と連結して $D+120$ 次元の複合ベクトル $[\boldsymbol{u}_t; \boldsymbol{a}_t]$ を構成し，多層パーセプトロンで圧縮して D 次元表現へ整形する．これらの圧縮ベクトルの系列を位置埋め込みを用いない Transformer エンコーダ [21] に入力し，ダンス特徴ベクトルの系列を得る．

提案エンコーダはオートエンコーダを用いて学習する．入力データを再構成するためのデコーダはエンコーダと対称な構造をとり，再構成された骨格構造パラメータ $\hat{\boldsymbol{S}}_t$ と感情特徴 $\hat{\boldsymbol{a}}_t$ を平均二乗誤差（MSE）によって以下の損失関数で学習する．

$$\mathcal{L} = \sum_{t=1}^{T} \Big[\mathrm{MSE}(\boldsymbol{S}_t, \tanh(\hat{\boldsymbol{S}}_t)) + \mathrm{MSE}(\boldsymbol{a}_t, \mathrm{sigmoid}(\hat{\boldsymbol{a}}_t)) \Big]$$

学習には AdamW オプティマイザ [22] を用い，バッチサイズ 8 で 200 エポックにわたってパラメータ更新を行った．骨格エンコーダには，53 関節（$J = 53$）それぞれについて 29 次元の関節パラメータ（$P = 29$）を入力とし，埋め込み次元を $D = 256$ に設定した．骨格エンコーダおよびデコーダの Graph Transformer はマルチヘッド 4，層数 2 とし，時系列エンコーダの Transformer はマルチヘッド 8，層数 4 と設定した．提案モデルの学習に関するプログラムは Web 上で公開中である[1)]．

B　歌詞とダンス間の関係性の例

図 5 において，歌詞とダンスの関係性を複数観察できる．たとえば，ダンス 3 では，歌詞中の数字と，指で数を数える動きが対応している．また，ダンス 4 では，撮影に関連する単語と，指で写真のフレームを模倣する動きが対応している．ダンス 5 では，動物の鳴き声を表す単語に対して，爪を立てる動きが対応している．また，ダンス 6 では，単語「約束」に対して小指を立てる動きが対応し，ダンス 7 では「あなた」のような二人称を表す単語に対して腕を使い前方を指す動きが対応している．ダンス 8 では，単語「お金」と硬貨を示す円を指で作る動きが対応していることがわかる．

1）https://github.com/KentoW/lyrics-and-dance