社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

ガウス過程を用いた周波数スペクトル分析による副詞の理解

谷口　　巴† 　持橋　大地†† 　長野　匡隼††† 　中村　友昭††† 　長井　隆行††††

稲邑　哲也†††††小林　一郎†

† お茶の水女子大学　〒112–8610 東京都文京区大塚 2–1–1
†† 統計数理研究所　〒190–8562 東京都立川市緑町 10–3
††† 電気通信大学　〒182–8585 東京都調布市調布ケ丘 1–5–1
†††† 大阪大学　〒560–8531 大阪府豊中市待兼山町 1–3
††††† 国立情報学研究所　〒101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †{g1620524,koba}@is.ocha.ac.jp, ††daichi@ism.ac.jp, †††m_nagano@radish.ee.uec.ac.jp,
††††tnakamura@uec.ac.jp, †††††nagai@sys.es.osaka-u.ac.jp, ††††††inamura@nii.ac.jp

あらまし　近年，汎用言語モデルの出現などにより自然言語処理には大きな革新がもたらされ、記述されたテキストに対する意味理解の研究は飛躍的に進んだと言える．一方で，実世界における言葉の意味理解に対しては，未だ十分に研究が進んでいるとは言えない．今後，ロボットなどが家庭に導入された際に，ロボットは日常生活でより充実したサービスを提供するために，言語を通じて人と同じ実世界における感覚を共有した動作が期待される．よって，実世界環境における現象を説明するための言語の意味を理解することは，重要な課題と言える．本研究では，中でも実世界環境で用いられる副詞の意味に着目し，特に人の動作を対象として理解することを目的とする．具体的に，3 つの手法を用いて課題に取り組む．(i) Gaussian Process Latent Variable Model を用いて潜在空間における副詞表現と人の動作対応関係を捉え，(ii) 人の動作を Spectral Mixture Kernel を用いたガウス過程により解析を行い，特定の副詞を表現する動作に共通する周波数カーネルを発見する．(iii) 動作の特徴を表す周波数カーネルとその動作を表現する副詞の共通トピックを同時に学習することにより，副詞と動作の対応関係を捉えるモデルを提案する．

キーワード　ガウス過程，トピックモデル，スペクトル混合カーネル，副詞の意味理解

Understanding Adverbs Expressing Human Actions by Frequency Spectrum Analysis Using Gaussian Processes

Tomoe TANIGUCHI†, Daichi MOCHIHASHI††, Masatoshi NAGANO†††, Tomoaki NAKAMURA†††,
Takayuki NAGAI††††, Tetsunari INAMURA†††††, and Ichiro KOBAYASHI†

† Department of Information Science, Ochanomizu University, 2–1–1 Ohtsuka, Bunkyo-ku, Tokyo 112–8610 Japan
†† The Institute of Statistical Mathematics, 10–3 Midoricho, Tachikawa-shi,Tokyo 190–8562 Japan
††† Department of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, 1–5–1 Chofugaoka, chofu-shi, Tokyo 565–0456 Japan
†††† Department of Systems Science, Osaka University, 1–3 Machikaneyama, Toyonaka-shi,Osaka 560–8531 Japan
††††† National Institute of Informatics, 2–1–2 Hitotsubashi, Chiyoda-ku,Tokyo 101–8430 Japan
E-mail: †{g1620524,koba}@is.ocha.ac.jp, ††daichi@ism.ac.jp, †††m_nagano@radish.ee.uec.ac.jp,
††††tnakamura@uec.ac.jp, †††††nagai@sys.es.osaka-u.ac.jp, ††††††inamura@nii.ac.jp

Abstract　In this study, we attempt to understand the meaning of adverbs through the features of human actions. Specifically, the trajectories in the nonlinear latent space obtained by compressing human actions with a Gaussian process latent variable model (GPLVM) are represented by a Gaussian process with a Spectral Mixture kernel. We also propose a multimodal topic model in frequency space that captures the correspondence between adverbs and the multiple frequency components that make up the trajectories in each dimension.

Key words　Gaussian Process, Topic Model, Spectral Mixture Kernel, Understanding the meaning of Adverbs

1. はじめに

近年，重要性が高まってきている家庭用ロボットには，日常生活において人と同じ感覚を共有した動作が期待される．動作に対する感覚は，自然言語では副詞を通じて表現されることが多い．ゆえに，特定の副詞を表現する複数の動作に共通する特徴を見つけることができれば，ロボットはその副詞の意味を本質的に理解したといえる．副詞の意味理解についての先行研究として，Pang ら [1] が表情認識や画像に映る物体の様態の解釈に取り組んでいるが，動作との関係性を捉えた研究はほとんどない．また，実際にロボットの動作と副詞の意味を結びつける研究は例をみない．本研究ではロボットに副詞を理解させる前段階として，人の動作の特徴と副詞の意味を結び付け，理解する手法を開発する．具体的には人の動作を Gaussian Process Latent Variable Model [2] で圧縮して得られた非線形な潜在空間での軌跡を，Spectral Mixture Kernel [3] を用いたガウス過程で表現する．学習したカーネルから，各次元の軌跡を構成する複数の周波数成分を特定し，副詞との対応関係を捉えるための周波数空間でのマルチモーダルなトピックモデルを提案する．これにより，動作について副詞を用いて表現することや、副詞表現から動作を生成することを可能とする．

2. 動作と副詞の結合トピックモデル

2.1 ガウス過程潜在変数モデル (GPLVM)

本研究では，動作から得られる高次元の姿勢情報を低次元に圧縮してモデル化するため，ガウス過程に基づく教師なし学習であるガウス過程潜在変数モデル (GPLVM) [2] を用いる．GPLVM とは，ガウス過程に基づく非線形な確率的主成分分析であり，N 個の D 次元観測値をまとめた行列 $\boldsymbol{Y}$ について，式 (1) を最大化するような低次元の入力 $\boldsymbol{X}$ を計算する．

$$p(\boldsymbol{X},\boldsymbol{Y}) = p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X}) \tag{1}$$

ここで，$\boldsymbol{X}$ は未知であるため $p(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ と仮定して，$p(\boldsymbol{Y}|\boldsymbol{X})$ を考える．$\boldsymbol{Y}$ はガウス過程に従い，$\boldsymbol{X}$ がわかれば出力の各次元が独立であると仮定すると，データ全体 $\boldsymbol{Y}$ の確率は $\boldsymbol{y}^{(1)}...\boldsymbol{y}^{(D)}$ の積であるため，$p(\boldsymbol{Y}|\boldsymbol{X})$ は以下の式で表される．ただし $\boldsymbol{K}_{\boldsymbol{X}}$ は共分散行列，$k(x,x')$ はガウス過程のカーネル関数であり，$K_{i,j} = k(x_i, x_j)$ で定義される．

$$p(\boldsymbol{Y}|\boldsymbol{X}) = (2\pi)^{-\frac{ND}{2}} |\boldsymbol{K}_{\boldsymbol{X}}|^{-\frac{D}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{K}_{\boldsymbol{X}}^{-1}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})\right) \tag{2}$$

本研究では，$\boldsymbol{X} \to \boldsymbol{Y}$ の GPLVM のカーネル関数として RBF カーネルを使用し，L-BFGS 法を用いて $\boldsymbol{X}$ およびカーネルのハイパーパラメータを最適化する．図 1 に，実際の動作から計算した $\boldsymbol{X}$ の例を示した．

2.2 スペクトル混合カーネル (SM Kernel)

上で得られた潜在空間 $\boldsymbol{X}$ での動作の軌跡について，その特徴を捉えることを試みる．Wilson ら [3] はガウス過程で使用する基底を，既存の基底やその組み合わせに限定せず，フーリエ領域で混合ガウス分布を考えることでデータから自動的に学習できるスペクトル混合カーネル (Spectral Mixture Kernel, SM kernel) という手法を提案した．ここではガウス過程の基底として，値が $\tau = x - x'$ だけに依存する定常基底関数 $k(\tau)$ を考える．ボホナーの定理より，任意の $k(\tau)$ は以下の形で表される．

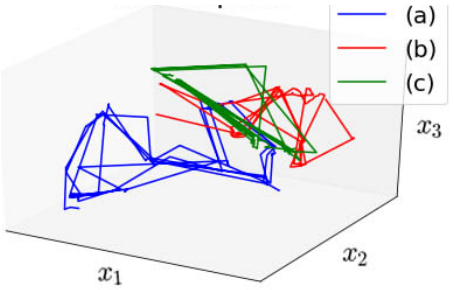


図 1　GPLVM による動作の非線形次元圧縮．ここでは (a)–(c) の異なる歩行動作が，低次元の潜在空間 $\boldsymbol{X}$ の軌跡として表現されている．

$$k(\tau) = \int_{\mathbb{R}^D} e^{2\pi i s^{\mathrm{T}}\tau} \psi ds \tag{3}$$

$k(\tau)$ は周波数領域での確率密度 $\psi(s)$ と等価なので，$\psi(s)$ に関して混合ガウス分布を考える．ガウス分布の各要素は，もとの領域では以下の基底関数を考えていることと等価となる．

$$k(\tau|\sigma,\mu) = \exp(-2\pi^2\tau^2 v^2)\cos(2\pi\tau\mu) \tag{4}$$

すなわち基底として，次の Q 個の基底関数の混合を考えていることになる．ただし，μ_q^d と v_q^d は q 個目の基底における入力 $\boldsymbol{X}$ における d 次元目の平均と分散である．

$$k(\tau) = \sum_{q=1}^{Q} w_q \cos(2\pi\tau^{\mathrm{T}}\mu_q) \prod_{d=1}^{D} \exp(-2\pi^2\tau_d^2 v_q^d) \tag{5}$$

パラメータの重み $\mathbf{w}$，平均 $\boldsymbol{\mu}$，分散 $\boldsymbol{\sigma}$ は通常のガウス過程のハイパーパラメータ最適化で学習できる．本研究ではこの手法を用いて，各動作について GPLVM で圧縮した 3 次元の潜在変数 $\boldsymbol{X}$ から，副詞と関係があると予想される Q 個の周波数成分 (平均 $\boldsymbol{\mu}$) を抽出して動作の特徴を表す観測値とする．$Q = 4$ としたときの例を図 2 に示した[1]．

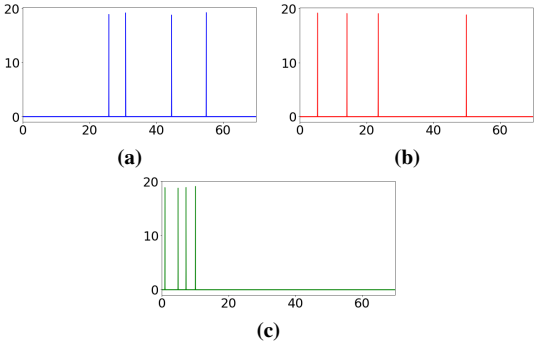


図 2　図 1 の各動作について，1 次元目の軌跡を解析したスペクトル混合カーネルによる周波数表現．縦軸，横軸はそれぞれ推定された 4 個のガウス分布の確率密度，平均を表す．

(注1)：$\boldsymbol{X}$ での軌跡を直接フーリエ変換することもできるが，その場合は関数がどこを通るか (関数の位相) と関数の特徴を分離することができない．SM kernel を用いることにより，純粋に軌跡の特徴だけを抽出することができる．

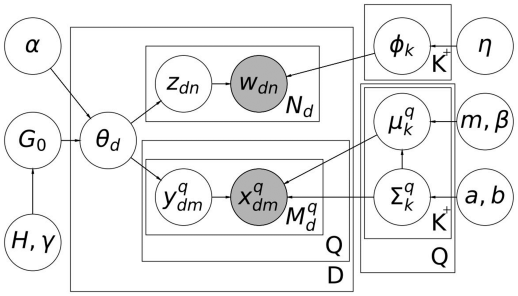


図 3　HDP-SMLDA のグラフィカルモデル.

2.3　階層的ディリクレ過程スペクトル混合潜在ディリクレ配分法 (HDP Spectral Mixture LDA)

動作から抽出された周波数成分は, その動作に付与された副詞と関係があると考えられる．ここで Gaussian-Multinomial LDA(GM-LDA) [4] を用いれば，周波数成分と副詞を同時にトピックごとにクラスタリングすることで，副詞をモデルに入力した際，その副詞と共起しやすい周波数成分を取り出すことができる．GM-LDA では，予めトピック数を決めておく必要がある．しかし，実際にはトピック数は未知であり、既知とする前提は大きな制約となっている．本研究では，GM-LDA に階層的ディリクレ過程を導入することでトピック数をデータから自動的に推定する「階層的ディリクレ過程スペクトル混合 LDA (HDP-SMLDA)」を提案する．このグラフィカルモデルを図 3 に示す. ここで，Q とは周波数成分の次元数であり，本研究では GPLVM を用いてデータを 3 次元に圧縮しているので $Q = 3$ となる．前セクションで扱った SM kernel におけるカーネルの混合数はこのモデルでは M_d として扱う．副詞はカテゴリカル分布からサンプリングされるが，周波数情報は連続データであるため，事前分布にガウス分布を仮定する．

生 成 過 程

各動作 d に潜在的なトピック分布 θ_d があると仮定する. ここでトピックの次元数は可変である．このとき, 動作に付与された副詞 $\{w_{dn}\}$ $(n = 1 \dots N_d)$ および動作の周波数成分 $\{x_{dm}\}$ $(d = 1 \dots D, m = 1 \dots M_d)$ の生成過程を以下に示す.

1. Draw $G_0 \sim \mathrm{DP}(\gamma, H)$.
2. For $d = 1 \dots D$,
   - Draw $\theta_d \sim \mathrm{DP}(\alpha, G_0)$.
3. For $n = 1 \dots N_d$,
   - Draw $z_{dn} \sim \theta_d$
   - Draw $w_{dn} \sim \mathrm{Mult}(\phi_{z_{dn}})$.
4. For $m = 1 \dots M_d$,
   - Draw $y_{dm} \sim \theta_d$
   - Draw $x_{dm} \sim N(\mu_{y_{dm}}, \sigma^2_{y_{dm}})$.

$\phi_k, \mathcal{N}(\mu_k, \sigma_k^2)$ はそれぞれ, k 番目のトピックに対応する副詞のカテゴリカル分布および周波数のガウス分布であり，互いの情報から算出されるトピック分布 θ を用いて, 各動作 d について 1 つ 1 つの副詞と周波数成分にトピックを割り当てていく.

2.4　副詞と周波数についてのサンプリング

ギブスサンプリングにより, 副詞と周波数のトピック分布を学習していく.

副詞についてのサンプリング

テーブルの割り当て集合を T, テーブルの卓番を ℓ とすると，中華料理店過程に従い，副詞 w_{dn} のトピック z_{dn} は，次式で座るテーブル t_{dn} をサンプリングすることで決定される．ここで，ℓ_{used}, ℓ_{new} はそれぞれ既存テーブルと新規テーブルを表し，L_k，L はそれぞれトピック k が振られたテーブル数，総テーブル数，V は語彙数を表す.

$$
\begin{aligned}
&p(t_{dn} = \ell | \boldsymbol{W}, \boldsymbol{T_{\backslash dn}}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta) \\
&\propto \begin{cases} (N_{dl\backslash dn} + \sum_{q=1}^{Q} M_{dl}^{q}) \dfrac{N_{kw_{dn}\backslash dn} + \eta}{N_{k\backslash dn} + \eta V} & (\ell = \ell_{used}) \\ \sum_{k=1}^{K} \dfrac{\alpha L_k}{L+\gamma} \dfrac{N_{kw_{dn}\backslash dn} + \eta}{N_{k\backslash dn} + \eta V} + \dfrac{\alpha\gamma}{L+\gamma}\dfrac{1}{V} & (\ell = \ell_{new}) \end{cases}
\end{aligned} \tag{6}
$$

新規テーブルに提供するトピックのサンプリングは以下の式を用いる．ここで，k_{used}, k_{new} はそれぞれ既存トピックと新規トピックを表す.

$$
\begin{aligned}
&p(z_{dl} = k | \mathbf{W}_{\backslash dn}, \mathbf{T}, \mathbf{Z}_{\backslash dl}, \alpha, \gamma, \beta) \\
&\propto \begin{cases} L_k \dfrac{N_{kw_{dn}} + \eta}{N_{k\backslash dn} + \eta V} & (k = k_{used}) \\ \gamma \dfrac{1}{V} & (k = k_{new}) \end{cases}
\end{aligned} \tag{7}
$$

ハイパーパラメータである η は不動点反復法により，以下の式を用いて更新する.

$$
\eta^{new} = \eta \frac{\sum_{k=1}^{K}\sum_{v=1}^{V} \Psi(N_{kv} + \eta) - KV\Psi(\eta)}{V \sum_{k=1}^{K} \Psi(N_k + \eta V) - KV\Psi(\eta V)} \tag{8}
$$

周波数についてのサンプリング

周波数成分 x_{dm} のトピック y_{dm} に関しては, 副詞の単語分布をガウス分布の確率密度関数 f に置き換え，以下の式を用いてサンプリングする.

$$
\begin{aligned}
&p(t_{dm} = \ell | \boldsymbol{W}, \boldsymbol{T}_{\backslash dm}, \boldsymbol{Z}, \boldsymbol{Y}, \alpha, \gamma, \eta) \\
&\propto \begin{cases} (N_{dl} + \sum_{q=1}^{Q} M_{dl\backslash dm}^{q}) f(x | \mu_k, \sigma_k^2) & (\ell = \ell_{used}) \\ \sum_{k=1}^{K} \dfrac{\alpha L_k}{L+\gamma} f(x | \mu_k, \sigma_k^2) & \\ \quad + \dfrac{\alpha\gamma}{L+\gamma} f(x | \mu_{k_{new}}, \sigma^2_{k_{new}}) & (\ell = \ell_{new}) \end{cases}
\end{aligned} \tag{9}
$$

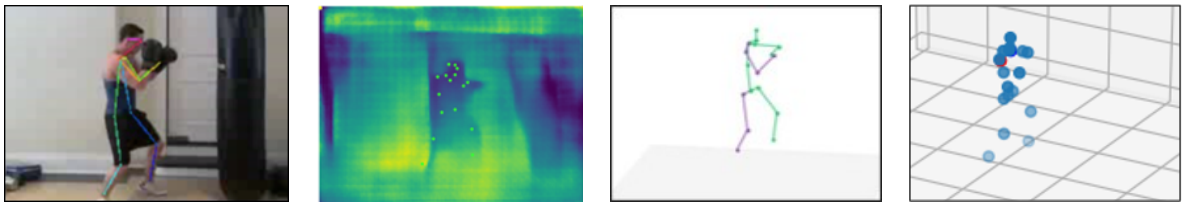
$$
\begin{aligned}
&p(z_{dl} = k | \mathbf{X}_{\backslash dm}, \mathbf{T}, \mathbf{Y}_{\backslash dl}, \alpha, \gamma, \beta) \\
&\propto \begin{cases} L_k f(x | \mu_k, \sigma_k^2) & (k = k_{used}) \\ \gamma f(x | \mu_{k_{new}}, \sigma^2_{k_{new}}) & (k = k_{new}) \end{cases}
\end{aligned} \tag{10}
$$

パラメータであるガウス分布の分散 σ^2 は固定値として学習する．ガウス分布は平均を 0 としたとき，およそ -3σ から 3σ の範囲にデータが収まることから，データの範囲にガウス分布が均等に配置されるよう以下の式を用いて算出した．ここで K^+ はイテレーション時のトピック数である.

$$
\sigma^q = \frac{max(\mathbf{X}^q) - min(\mathbf{X}^q)}{6K^+} \tag{11}
$$

ガウス分布の平均 μ は以下の事後分布からサンプリングする．ここで精度 λ はガウス分布の分散を σ^2 としたとき，$\lambda = 1/\sigma^2$ である.

$$
p(\mu | \mathbf{Y}) = \mathrm{N}(\mu | m, (\beta\lambda)^{-1}) \tag{12}
$$



(a) Openpose による画面座標推定　**(b)** FCRN-depth による深度推定　**(c)** 3 次元の骨格座標の推定　**(d)** 回転行列による方向正規化

図 4　動画データの前処理による動作の骨格座標の抽出.

ただし β_0, m_0 を事前分布のパラメータとして

$$\beta = M + \beta_0,\ m = \frac{1}{\beta}\left(\sum_{m=1}^{M} x_m + \beta_0 m_0\right). \qquad (13)$$

ここで新規トピックに対応するガウス分布の平均 $\mu_{k_{new}}$ はクラスタに所属するデータがないため、直接的な推定ができない. 平均は適当なパラメーターを使ってガウス分布からサンプリングしてある程度学習させた後，従来通り推定を行う.

2.5　集中度 α の推定

よりデータにフィットしたトピック数を推定するため，集中度 α の事前分布にガンマ分布を仮定し推定を行う.

$$p(\alpha|\pi, s, Z, c_1, c_2) = Ga(\alpha|c_1 + K^+ - s, c_2 - \log\pi) \qquad (14)$$

ただし，π と s は次のようにサンプリングする.

$$p(\pi|\alpha, s, Z, c_1, c_2) = Beta(\pi|\alpha + 1, N + M) \qquad (15)$$

$$p(s|\alpha, \pi, Z, c_1, c_2) = Bernoulli\left(s\left|\frac{N+M}{N+M+\alpha}\right.\right). \qquad (16)$$

3.　実　　験

3.1　使用するデータ

本研究では 100 種類の異なる歩行動作を集めた動画[2]とダンスに関するデータセット AIST++[3]を用いて実験を行った.

100 Walks

YouTube に掲載されている 2 次元の動画である．入力データとして 3 次元の姿勢情報が必要であるため，動画を動作の切れ目で 100 個に分割し，以下の 4 つの手法を用いて推定した.

（1）　Openpose [5] を用いて動画データから 2 次元の骨格座標を推定 (図 4(a))

（2）　FCRN-depth prediction [6] を用いて動画データの深度を推定 (図 4(b))

（3）　1,2 の推定結果と 3d-pose baseline [7] を用いて動画データから 3 次元の骨格座標を推定 (図 4(c))

（4）　回転行列を用いて人の体の向きを正規化 (図 4(d))

表 1　学習に使用したデータの詳細.

	動画数	副詞の種類	平均副詞数
100 Walks	100	264	12.93
AIST++	1199	1767	16.18

（注2）：https://www.youtube.com/watch?v=HEoUhlesN9E

（注3）：https://google.github.io/aistplusplus_dataset/

AIST++

産業技術総合研究所が公開しているダンスに関するデータセットである．音楽を入力として，それに合わせたダンスを出力するモデルに関する論文 [8] が公開されており，学習の際に用いた 16 個の関節点の 3 次元の姿勢情報が収録されている．中でも Basic Dance と分類される単調なダンスに関するデータセット 1199 個を使用した．10 種類のジャンルのダンスに 10 種類の振付がされており，計 20 名の踊り手がそれぞれの動画を担当している．踊り手はジャンルに合わせた音楽に合わせて指定された振付を踊っており，音楽のテンポは 6 段階でそれぞれ用意されている.

3.2　副詞アノテーション

クラウドソーシングシステム Lancers[4]を用いて，アノテーターに動画の各動作について思いつく限り自由に副詞をアノテーションしてもらうよう依頼した．全動画で 3 個以上出現した副詞に限定し，満たない副詞はノイズとして除去した． 100 Walks データセットには 100 動画につき 20 名，AIST++ データセットには 50 動画につき 5 名のアノテーターにアノテーションを依頼した．作成した副詞データセットについての詳細を表 1 に示す．ただし平均副詞数とは，1 つの動画に対して付与された副詞数の平均を表す．先行研究と比較して，どちらのデータも副詞の種類が多く収集できた.

3.3　方向ベクトルの算出

元の姿勢情報を復元するため，各関節間の方向ベクトルを入力データとした．また腕の長さなど，個人差を無くすため，単位ベクトルを算出した．100 walks データセットは 16 本，AIST++ データセットは 14 本の方向ベクトルを算出し，それぞれ 1 フレーム毎に 3 次元座標を列方向に結合した．したがって，それぞれ 48 次元と 42 次元のデータになる.

3.4　周波数成分の抽出

以下の 2 つの手法を用いて, 前処理した動画データから周波数成分を抽出した. カーネルの混合数 M_d は 4，または 10 として推定を行った.

（1）　GPLVM を用いて高次元の姿勢データを 3 次元の潜在変数に圧縮する (図 1)

（2）　SM kernel を用いて 3 次元の潜在変数から, 各次元について周波数成分を抽出する (図 2)

100 Walks の学習データのうち, 3 個の動作を GPLVM で圧縮した 3 次元の潜在空間にプロットしたものを図 1 に示す. 歩行

（注4）：https://www.lancers.jp/

表 2　AIST++ データ ($M_d = 4$): HDP-SMLDA で得られたトピック別副詞上位 5 語.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
激しく	楽しそうに	規則正しく	しなやかに	力強く	踊るように	慣れたように	テンポ良く
力強く	リズミカルに	テンポよく	優雅に	激しい	ステップを踏み	安定的に	スタイリッシュに
はっきりと	軽やかに	躍動的に	なめらかに	激しく	嬉しそうに	くねくねと	気持ち良さそうに
熱心に	弾むように	生き生きと	軽やかに	素早く	躍動するように	キビキビと	流れるように
上品に	元気に	大胆に	くるくると	大胆に	つまらなそうに	ダイナミックに	格好良く
Topic 9	**Topic 10**	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**
ゆったりと	ダイナミックに	はずむように	格好よく	キビキビと	小刻みに	確かめるように	軽く
滑らかに	激しく	ひろがるように	カクカクと	機械のように	回るように	ひょうきんに	揺れているような
ゆっくりと	くねくねと	たどたどしく	おおらかに	コミカルに	細かく	丁寧に	波のような
機械的に	おおきく	ぐんぐんと	楽しそうな	しっかりと	クルクルと	慎重そうに	細かい動作で
ゆるやかに	キレよく	落ち着いた	機械のように	ロボットのように	リズム感よく	探すように	ロボットのような

動作は繰り返しの動作であるため, 潜在変数は図のように円を描くような動きになる. $M_d = 4$ の時，SM kernel によって各動画の 1 次元目について最適化された平均 $\boldsymbol{\mu}$ と分散 $\boldsymbol{\sigma}$ をパラメータとしてガウス分布を描画したものを図 2 に示した[5]. 式 (5) より平均 μ の値が大きいほど周期が小さくなることから，値の変動が低速な動画データほど基底を表すスペクトルは左側に多く見られると推測できる. よって, (a) は遅い動きの成分が多く, (c) は速い動きの成分が多く，(b) はその中間的な動きということがわかる. SM kernel はパラメータとして重みが最適化されている．この重みは各周波数成分の重要度を表すものであり，各動画の動きの特徴として用いる周波数成分は重みを用いてイテレーションごとにサンプリングすることとした.

3.5 実 験 結 果

AIST++ データ ($M_d = 4$) について，学習されたトピック–単語分布から各副詞について NPMI [9] を計算した上位 5 語を表 2 に示す. また学習された平均 $\boldsymbol{\mu}$ を用いて，各トピックに対応するガウス分布から 100 個ずつサンプリングし，3 次元空間に描画したものを図 5 に示す. 各サンプルはトピックを象徴する周波数表現であり，距離が近いことは周波数表現が似ていることを意味する．分散は推定していないため，図の点の広がりは一定である．モデルはパープレキシティを用いて評価する．パープレキシティは評価用文書を構成する各単語確率の幾何平均の逆数であるが，値が小さいほうがより優れたモデルといえる. 各トピックモデルにおける訓練時のパープレキシティを表 3 に示す．ここで Unigram モデルの値は学習前の単語分布を用いて評価用動画におけるパープレキシティを算出したものである.

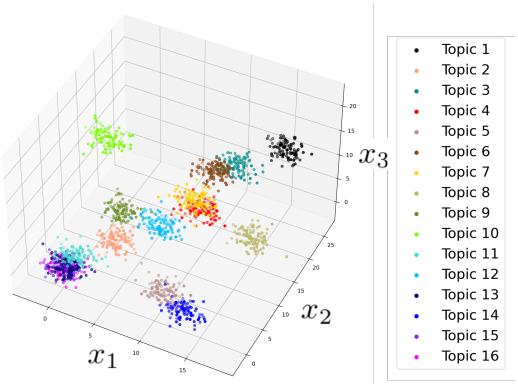


図 5　トピックと動作特徴の関係．HDP-SMLDA で学習した各トピックに対応するガウス分布から抽出された 100 個のサンプルを 3 次元空間に描画.

(注5)：推定された分散がきわめて小さいため, 図ではガウス分布がデルタ関数状に描画されている.

表 3　各トピックモデルにおける訓練時のパープレキシティ.

	Unigram	LDA	HDP-SMLDA ($M_d = 4/10$)
100 Walks	156	99	**52** / 57
AIST++	558	331	**218** / 249

3.6 周波数情報から副詞の生成

周波数が正しく副詞と紐づいているかを確認するため，学習した単語分布を用いて，評価用動画 (図 6) の周波数情報から副詞を生成する実験を行った．実際にアノテーションされた正解の副詞と HDP-SMLDA が算出した確率の高い副詞上位 7 語を表 4 に示す. 評価用動画の多くは，上位単語をみると $M_d = 10$ の方が適切な副詞を推定していることを確認した.

3.7 考　　察

図 5 では，狙い通りデータの幅を均等に分けるようにガウス分布が配置されている．中でも，トピック 5 とトピック 14，トピック 11 とトピック 13 とトピック 16 は μ の値が近くなっているが，表 4 より，トピックごとの上位副詞から動きの内容が近いということが推察される．また，トピック 1，トピック 8，トピック 9，トピック 10 は他のトピックと比べてそれぞれ μ が離れた位置にプロットされている．この 3 つのトピックはほ

図 6　評価用動画. 動画内でダンサーはジャズバレエを踊っている.

表 4　正解の副詞と推定された副詞の上位 7 語.

正解の副詞	$M_d = 4$ の場合	$M_d = 10$ の場合
情熱的に	力強く	テンポ良く
陽気に	激しい	スムーズに
テンポ良く	激しく	スタイリッシュに
スムーズに	大胆に	流れるように
流れるように	堂々と	陽気に
力強く	キビキビと	悲しそうに
大胆に	ダイナミックに	気持ちよさそうに

かのトピックに比べて，副詞の特徴が周波数に顕著に表れていることがわかる．トピック 1 と 10 は内容が重複しているように見えるが，上位 20 単語を観察するとトピック 1 は「勇ましく」,「重々しく」など感情的なダンス、トピック 10 は「キレ良く」,「シャキッと」などラフなダンスが想像できるような副詞がランクインしている．以上の結果からモデルは周波数情報を用いることで副詞を意味的にも動作的にもうまくクラスタリングしていることが示された．モデルの評価に関して，表 3 より，各データともに LDA の訓練時でのパープレキシティよりかなり小さくなっており，これは周波数情報が副詞のトピック分類に貢献していると言える．またカーネルの混合数を増やすことでパープレキシティが小さくなるはずが，実際は大きい値を記録している．周波数情報から副詞の生成の実験において，$M_d = 10$ とした時，妥当な副詞を推定できていることから，モデルが推定した副詞を，アノテーション時にアノテータが該当の語彙を思いつかなかった可能性があると考えられる．

4. NN モデルとの比較

4.1 マルチラベル学習

比較のため NN モデルを用いて実験を行った．本研究で扱うデータは，1 つの入力に対して複数のラベルが付与されているため，マルチラベル学習をする必要がある．一般的なクラス分類の学習では出力された確率と，入力に対する 1 つのラベルとの差異を誤差として逆伝播していくが，今回は動画につけられたすべての副詞ラベルに対して誤差をとり，その平均を逆伝播することで学習を行なう．具体的に LSTM，多層パーセプトロン (MLP) にそれぞれ以下の 4 つのデータを入力し実験を行った．

（1） LSTM に GPLVM で圧縮したデータを入力
（2） LSTM に圧縮前のオリジナルデータを入力
（3） MLP に周波数成分 ($M_d = 4$) を入力
（4） MLP に周波数成分 ($M_d = 10$) を入力

4.2 実験結果と考察

各モデルにおける評価時のパープレキシティを表 5 に示す．

LSTM は，GPLVM で圧縮したデータと，圧縮前の元データとを比較すると，圧縮したデータの方がパープレキシティは低くなっていることから，データの次元圧縮がクラス分類に有効であることが示された．また NN モデルはどれも値がかなり高く，うまく副詞を学習できているとは言えない．どちらのデータセットも提案手法が最高スコアとなっており，少ないデータで正しく副詞の推定ができていることが実験により示された．

表 5　各モデルにおける評価時のパープレキシティ.

	LSTM (GPLVM/Original)	MLP ($M_d = 4/10$)	HDP-SMLDA ($M_d = 4/10$)
100 Walks	210 / 402	253 / 284	**89** / 117
AIST++	1068 / 1794	994 / 1027	**320** / 382

5. まとめと今後の課題

本研究は，動作に係わるあいまいな副詞の意味理解を目的とし，動作特徴と副詞の関係を学習する結合トピックモデル HDP-SMLDA を提案した．このモデルでは副詞を，動作の潜在空間における軌跡を表現するガウス過程において，そのカーネルの周波数空間での混合分布の形で表現することにより，あいまいな副詞の動作の成分を推定することができる．また逆に，動作を複数の周波数成分の混合として考えることで，実験にて動作特徴から適切な副詞の生成に成功している．最後に比較として単純な NN モデルより優れたクラス分類性能を示した．今後の課題として，より正しくモデルを評価するため，人手評価であったり，BERT [10] を用いて類似度を算出して評価する方法を検討している．またどちらの実験もイテレーション数は大きな値に設定しているが，学習中のパープレキシティの様子から，さらにパープレキシティが下がる可能性があるため，モデルに収束判定を組み込む必要があると考える．

文　献

[1] Bo Pang, Kaiwen Zha, and Cewu Lu. Human Action Adverb Recognition: ADHA Dataset and A Three-Stream Hybrid Model. *CoRR*, Vol. abs/1802.01144, pp. 2438–2447, 2018.

[2] Michalis K. Titsias and Neil D. Lawrence. Bayesian Gaussian Process Latent Variable Model. In *AISTATS 2010*, pp. 844–851.

[3] Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian Process Kernels for Pattern Discovery and Extrapolation. In *ICML 2013*, pp. 1067–1075.

[4] David M. Blei and Michael I. Jordan. Modeling annotated data, 2003.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 43, No. 1, pp. 172–186, 2021.

[6] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *3DV*, pp. 239–248, 2016.

[7] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A Simple yet Effective Baseline for 3D Human Pose Estimation. In *ICCV 2017*, pp. 2640–2649, 2017.

[8] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021.

[9] Gerlof Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, pp. 31–40, 2009.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, Vol. abs/1810.04805, , 2018.