擬態語による歩容の記述と生成

加藤 大貴

要　旨

歩行動作は人間にとって最もなじみが深い動作の一つである．人間の歩行動作の様子は歩容と呼ばれ，見慣れた動作であるがゆえに，人間は人間の歩容の細かな違いを弁別することができ，「強そう」，「軽やか」など，様々な印象を感じ取ることができる．このような能力を計算機に獲得させることは，直感的インタフェースの実現や，より人間味がある人工知能の開発などに有益であると考えられるが，歩容認証や行動認識などのタスクが盛んに研究される一方で，従来は歩容を記述するための適切なラベルが提案されていないことを大きな要因として，歩容そのものの直感的で精緻な記述・生成を試みた研究は少なく，限定的なのが現状である．そこで本研究では，擬態語を利用することで，このような歩容の記述・生成を試みる．擬態語には音象徴性という性質があり，擬態語を構成する音素の音響的印象が事象の様態と対応するため，人間は擬態語に対して共通のイメージを想起するとされている．この擬態語を構成する音素によって印象が決まるという性質をふまえると，例えば「とことこ」歩く，「どこどこ」歩く，というように，擬態語を構成する音素の一部を入れ替えることによって，微妙な印象の違いを表現することが可能である．言い換えれば，この性質は，動作に関する細かい印象の基底として擬態語の音素を利用可能であることを示唆している．

本論文では，以上の仮説に基づき，擬態語を「音韻空間」上で表現したベクトルである「音韻ベクトル」と歩容との対応関係を獲得することにより，歩容の印象の細かな違いを表現し分けることが可能な，人間の直感に近いモデルを構築し，これをもって（1）擬態語により歩容を記述する手法，および（2）擬態語から歩容を生成する手法の二つを提案する．

本論文は全 5 章からなり，第 1 章では，上で述べたような本研究の背景，目的，位置付けなどを詳述する．

次に第 2 章で，両手法の学習に用いるために新たに構築した，擬態語がアノテーションされた歩容データセット「HOYO」について紹介する．本研究では歩容の細かな違いをモデルに学習させる必要があることから，歩容自体もできるだけ多様なものが収録されているデータセットを用いるのが望ましい．そこで，既存の歩容認証データセットにアノテーションを追加するのではなく，歩容の動画像も新規に撮影した．この際，歩行者に対して擬態語による動作の教示を行なうことにより，歩容そのものの多様性を豊かにした．しかし，歩行者が自身のイメージ通りに体を動かせるとは限らないため，得られた歩容は客観的に見て，教示された擬態語を表現できているとは限らない．そこで，第三者が歩容を見た際に想起する擬態語を，歩容を表現する真の擬態語と定義し，第三者による評価に基づいて，改めて歩容に対する擬態語のアノテーションを行なった．ここで，本データセットでは各歩容に 2 種類の擬態語アノテーションを付与した．一つは選択式アノテーションであり，主観ラベルと同様のカテゴリカルな擬態語ラベルを付与するものである．もう一つ

は自由記述アノテーションであり，アノテータに複数の擬態語を自由に回答させるものである．構築したデータセットは Web 上で公開しており，擬態語がアノテーションされた唯一無二の公開歩容データセットとして，感性工学，行動認識，異常検出などのさまざまなタスクへの応用が期待される．

続く第 3 章では，擬態語により歩容を記述する，すわわち，擬態語を入力として歩容を出力する手法を提案する．音韻ベクトルは，音象徴性に基づく表現であるため，人間の直感をよく反映した形式となっていることが期待できる．そこで，提案手法では歩容から End-to-End に擬態語を求めるのではなく，（1）歩容を入力として推定音韻ベクトルを求めるモジュールと，（2）推定音韻ベクトルを擬態語に変換するモジュールの 2 段階に分けて歩容を記述する．中間にこのような人間の直感をよく反映した形式の表現を明示的に導入することで，提案手法も人間の直感に近い歩容記述が可能になると期待できるほか，学習に用いていない新奇的な擬態語を出力することも可能となる．そして，評価実験により，提案手法は歩容の記述の正確さを損なわずに，より自然な擬態語を出力できることを確認する．また，提案手法は無作為に音素を選んで生成した擬態語よりも正確な擬態語を出力できることを確認する．更に，提案手法に新奇的な語を含む多様な擬態語を出力する能力があること，自然さの考慮度合いを制御するハイパパラメータ α を調節することで，用途に応じて擬態語の新奇性を調節できることを確認する．

そして第 4 章では，第 3 章とは逆に，擬態語から歩容を生成する手法を提案する．提案手法では，歩容の生成をスタイル変換問題として捉え，音韻ベクトルをスタイルとして入力し，歩容を生成するモデルを構築する．そして，評価実験により，入力に一般的な擬態語を用いた場合，クエリとした擬態語と生成された歩容モーションが正しく対応することを確認する．

以上のように，本研究では，歩容の印象の細かな違いを表現し分けることが可能な，人間の直感に近いモデルを構築し，また，これを用いた擬態語による記述と生成のための方法論を確立する．そして，評価実験を通して，擬態語による歩容の記述と生成の双方を一定の水準で実現できることを確認する．

目次

表目次

図目次

第 1 章

序論

本論文は，擬態語による歩容の記述と生成に関する筆者の研究成果をまとめたものである．

本章ではまず 1.1 節で，擬態語がもつ性質やその歩容との関係，歩容のモデル化のための擬態語の有用性など，本研究の背景について論じる．次に 1.2 節で，擬態語や歩容に関する従来の研究について概観し，1.3 節で，本研究の目的と位置づけについて述べる．最後に 1.4 節で，本論文の構成について述べる．

1.1　研究の背景

歩行動作は人間にとって最もなじみが深い動作の一つである．人間の歩行動作の様子は歩容と呼ばれ，見慣れた動作であるがゆえに，人間は人間の歩容の細かな違いを弁別することができ，「強そう」，「軽やか」など，様々な印象を感じ取ることができる．

コンピュータビジョン分野の研究においても歩容から得られる情報は歩行者の行動認識 [2]，感情認識 [3]，異常検知 [4]，属性認識 [5, 6] 等の様々な用途において有用とされている．近年は OpenPose [7, 8] に代表されるような，可視光画像から高精度に人体部位の座標を検出する手法が手軽に利用できるようになったことで，人間の骨格情報やその動きの情報を利用することが容易になった．

しかし，歩行者が「歩いている」，「走っている」などの動作を認識する手法の研究はあれど，歩行者が「どのように」歩いているかに注目した研究，いわば，歩容そのものの精緻な認識を試みた研究は少ない．上で挙げたような，従来の歩容に関する研究は，歩行者の感情や体調，急いでいるか否かといった意図など，そのような歩容になるに至った内部

的な原因に着目した研究が中心である．一方，より直接的に歩容自体に着目した研究が少ないのは，従来，歩行者の動き方を精緻かつ直感的に記述するための適切なラベルが提案されていないことが大きな要因であると考えられる．

一方，近年の Zoom* などの遠隔ビデオ会議システムなどを利用したオンラインでの催事の増加や，いわゆる Virtual Youtuber のようなアバタを利用した動画像コンテンツの普及に伴い，個人が仮想空間上のアバターに動き（モーション）を設定したいという需要が今後増えていくことが予想される．この際，専門的なコンピュータグラフィックス（Computer Graphics; CG）に関する知識がなくとも，直感的にモーションを生成できるのが望ましい．これを実現するためには，記述とは逆処理となる，直感的なクエリから歩容モーションを生成する技術が必要であるが，これも精緻な認識と同様に，直感的なクエリとして利用できる適切なラベルが提案されていないという問題から，従来は探求されてこなかった．

これら両方の問題を同時に解決するために，本研究では擬態語を利用することを考える．擬態語とは，日本語，Bengal 語，韓国語，Tamil 語など，世界各地のいくつかの言語でよく用いられる表現方法であり，多種多様な質感的・動的な印象を直感的に表現することができる [9–13]．日本語の場合，例えば「つるつる」，「のろのろ」等の語が挙げられる．擬態語には音象徴性という性質がある．音象徴性とは，擬態語を構成する音素の音響的印象が事象の様態と対応するという性質であり，この性質のため，人間は擬態語に対して共通のイメージを想起するとされている [14, 15]．このことから，擬態語は表現が容易ではない直感的な印象を他者に対して伝える手段として有効だと考えられており，例えば藤野らは，人間が運動動作の感覚を他者に教示する際に擬態語の利用が効果的であることを指摘している [16]．また，擬態語は直感的な印象を計算機に伝える手段としても有効であると考えられており，神原らは，「ぎざぎざ」と発話しながら線を描くことでぎざぎざな線を描くことができる直感的な描画インタフェースを提案している [17]．

Köhler が示したブーバ／キキ効果 [18, 19] に代表されるように，音象徴性は言語によらない普遍的な性質であるが，例えば英語においては，動作の細かな違いを表現する際，「Swagger（威張って歩く）」，「Trot（急いで歩く）」などのように動詞の語彙を増やすことで多様な表現を実現しているのに対し，日本語のように擬態語が豊富な言語では，「どしどし」歩く，「すたすた」歩くなどのように擬態語による動詞の副詞的な修飾によって多様

* https://explore.zoom.us/ja/products/meetings/

な表現を実現している [20, 21]．印象の情報が副詞として独立していることや，擬態語は構造が単純であることから，擬態語表現は音象徴性の影響が色濃く表れていると考えられる．また，擬態語を構成する音素によって印象が決まるという性質をふまえると，例えば「とことこ」歩く，「どこどこ」歩く，というように，擬態語を構成する音素の一部を入れ替えることによって，微妙な印象の違いを表現することが可能である．言い換えれば，この性質は，動作の細かい印象の基底として擬態語の音素を利用可能であることを示唆している．このような擬態語が存在する言語のうち，日本語（話者数約 1.3 億人）は Bengal 語（話者数約 2.6 億人）に次いで 2 番目に話者数が多い言語であり，擬態語の代表として扱うのに申し分ない言語だと考えられることから，本研究では日本語の擬態語を取り扱う．

以上に述べたように，擬態語には多種多様な事象の様態を直感的に表現する能力があるが，類似した概念である擬音語（「ばちばち」，「ぽきっ」等といった，音を模倣した語）が比較的早くから工学的な研究を試みられていた [22–26] のに対し，擬態語の工学的な研究例は少ない．さらに，既存の擬態語の工学的な研究では，「つるつる」，「ざらざら」等の質感を表す擬態語が注目されることが多く [27, 28]，動的な印象を表す擬態語はほとんど着目されてこなかった．しかし，日本語において歩容を表す擬態語（56 語）は，食感を表す擬態語・擬音語（73 語）に次いで 2 番目に語彙が多く，全身を使う動作としては最も語彙が多い [10] ことからもわかるように，擬態語を用いれば，歩容の直感的な印象という観点から，歩容をより直感的に精緻に描写することが可能である．

1.2　従来研究

1.1 節で述べたような背景に基づき，本節では，1.2.1 項で擬態語に関する従来研究を，1.2.2 項で歩容の認識に関する研究を，1.2.3 項で歩容の生成に関する研究についてそれぞれ概観する．

1.2.1　擬態語に関する研究

まず，音象徴性を利用した擬態語の定量化に関する従来研究について述べる．Sakamoto らは，擬態語は患者が医者に病状を説明するのに有効だが，痛みに関する擬態語の語彙は日本特有であり海外の医者には通じないという背景から，擬態語を 35 種類の医学的な尺度に変換することで，医者による擬態語の意味理解を助けるシステムを提案してい

る [29–31]．戸本らは，音象徴性に基づいて擬態語を 32 次元のベクトルに定量化し，それを恒等写像学習によって 2 次元に写像することで，食感に関する擬態語のシソーラスマップを作成する手法を提案している [32]．秋山らは，日本語の音から感じるイメージを複数の形容詞対で評価する主観評価実験を実施し，その結果を因子分析にかけることによって，各音素を「キレ・俊敏さ」,「柔らかさ・丸み」,「躍動感」,「大きさ・安定感」と名づけられた四つの属性値で数値化する手法を提案している [33, 34]．このように，擬態語を定量化する研究はいくつか例があるが，これらは擬態語自体の可視化や意味理解を目的としている．

次に，擬態語と他メディアとの関連性を調べた研究について述べる．前節でも述べた通り，最も多いのは質感画像を取り扱った研究である．Shimoda らは「ふわふわ」,「ぴかぴか」等の質感を表す擬態語をクエリとして検索した Web 上の画像からデータセットを作成し，特定の擬態語と対応する画像を無関係な質感画像から分離させる実験を行ない，画像特徴として学習済み深層学習モデルの中間層出力を利用することで高精度に分離可能であることを示している [28]．權らは，質感画像に擬態語のアノテーションを付与し，その擬態語の構造をベクトルで表現したものと画像の対応関係を学習させる試みを通じて，質感画像と擬態語の音韻に関連性が存在することを示唆している [27]．これらの画像に関する研究例は近年の畳み込みニューラルネットワーク（Convolutional Neural Network; CNN） [35, 36] の発展に伴って行なわれたものが多く，2015 年ごろに研究例が集中している．本研究は，画像から受ける静的な印象ではなく，動的な印象に着目しているという点で対象が異なる．

動画像と擬態語の関係性に関する工学的な研究例は筆者が知る限り従来存在せず，心理的観点から分析した研究のみが行なわれている．鍵谷らは CG 映像作成ソフトウェアを用いて作成した粘性をもつ液体の映像を実験参加者に提示し，想起される擬態語を回答させる実験を行なうことで，動画像と，映像から想起される擬態語を構成する音韻の種類に関連性があることを明らかにしている [37]．動画像と擬態語の関係が従来あまり追究されてこなかったのは，関係性の学習に必要なデータセット，特に擬態語のアノテーションが不足していたことが大きな要因であると思われる．中でも映像をそのまま入力とする 3 次元畳み込みニューラルネットワーク（3 Dimensional Convolution Neural Network; 3D-CNN） [38] などの学習手法は，画像を入力するネットワークと比較して計算コストが高いことから，特に学習に必要な量のデータを確保するのが難しいという問題がある．

Shimoda らの質感画像の研究では，Web 画像検索結果を利用することでデータセット不足の問題の解決を試みているが，Shimoda らはこの方法の問題点として，例えば「ごつごつ」で検索して得られる画像は山肌の画像がほとんどであることから，このようなデータを用いて学習すると「ごつごつ」と「山肌という概念」との関係が学習されてしまい，質感の学習という目的が十分に果たせない可能性を指摘しており，Web 検索を利用したデータセット構築には課題が多い．更に，この方法はキャプション等，文字列のラベルが付与されていることが多い静止画像であればこその手法であり，動画像に応用するのは現実的ではない．權らは既存の質感画像データセット [39] に，鍵谷らは自作の CG 映像に対して人手で擬態語のアノテーションを実施している．このように，擬態語がアノテーションされた公開データセットというものは類がなく，研究者が自前でデータセットを構築するのが常となっている現状があり，これが研究の裾野が広がらない一因であると考えられる．本研究は，提案手法の学習・評価のために筆者が構築したデータセットを公開することにより，このデータセット不足問題にも貢献する．

1.2.2　歩容の認識に関する研究

まず，歩容からの個人認証や，年齢・性別といった Soft biometrics 関連の属性の認識について述べる．これらは広く研究されており，例えば Li らは，Joint intensity [40] という画像間の相違度を推定するニューラルネットワークを用いて，衣服などの条件変化に頑健な歩容からの個人認証手法を提案している [41]．Sakata らは，歩容から歩行者の年齢を推定するために，複数の CNN を用いて段階的に年齢層を絞り込んでいく手法を提案している [42]．深層学習登場以前の研究としては，Cao らは，歩行者の全身画像を入力とし，HOG（Histograms of Oriented Gradients）特徴 [43] のアンサンブル学習により歩行者の性別を推定する手法を提案している [44]．Ge らは，同様に SIFT（Scale Invariant Feature Transform）特徴 [45] のスパースコーディングにより歩行者の年齢を推定する手法を提案している [46]．年齢や性別推定のために歩容の時系列情報を利用する研究は複数存在し，例えば Davis らはモーションキャプチャにより取得した歩容の関節点データから，身長，歩幅，周期などの特徴に基づいて大人と子供を分類する手法を提案している [47]．また，Yu らは，Gait Energy Image （GEI） [48] という，歩容シルエットの時間平均画像を利用して性別を推定する手法を提案している [49]．しかし，これらの研究は動きを特徴として利用しているだけであり，動き自体を記述しているとは言い難い．

次に，動作認識などの，歩行に限らない人間の動きの認識について述べる．Carreira らは，3D-CNN を用いた動作認識手法を提案している [50]．Liu らは，歩行者の骨格情報とグラフ畳み込みネットワーク（Graph Convolutional Network; GCN） [51] を利用した動作認識手法を提案している [52]．十分に大規模なデータセットが普及する以前は，入力に Optical flow [53] を利用した 2 次元畳み込みネットワーク（2 Dimensional Convolution Neural Network; 2D-CNN）を用いる手法 [54] や，Fisher vector [55] を利用する手法 [56] が有効であった．これらの研究は，本研究と比較して薄く広い範囲の認識を目的としていると言える．また，近年の高精度な動作認識モデルを学習させるためには大規模なデータセットが必要であり，歩容に限ってこれらの手法の学習に耐える規模のデータセットを確保するのは容易でないという問題がある．

そのほか本研究に比較的近いタスクとして，歩容からの感情認識や，異常歩容の検知の研究が挙げられる．Bhattacharya らは，歩容の 3 次元骨格情報から，三つの関節点で張られる平面の面積比などの特徴量を用いて，「happy」，「sad」，「angry」，「neutral」の 4 種類の感情を認識する手法を提案している [3]．本研究はこのような感情認識の研究と比較して，動作によって表現される内的な状態ではなく，動作自体を認識するという点で観点が異なる．Temuroglu らは，正常な歩容の骨格情報を用いて学習した Auto encoder [57] を利用し，酔っ払いなどの異常な歩行動作をしている歩行者を検知する手法を提案している [4]．ここでは，異常な歩容が Auto encoder に入力された際には再構成誤差が大きくなりやすいことを利用して異常歩容を検出している．このような異常検知の研究は，普通でない歩容に関心があるという点で本研究と類似しているが，あくまで検知することが目的であり，どのように普通でないかを認識することに関心がない点で本研究とは異なる．

1.2.3　歩容の生成に関する研究

従来，歩容の生成といえば，既存の歩容データを入力として，後続の歩容を予測する研究 [58, 59] や，複数の歩容を合成する [60] 研究が一般的である．Bhattacharya らは，「happy」，「sad」，「angry」，「neutral」の 4 種類の感情ラベルを指定して，対応した歩容を生成する手法を提案している [59]．Kovar らは，動作生成（Motion synthesis）の文脈で，合成元として「普通の歩容（normal walking)」や「こそこそした歩容（sneaking movement)」を使い分けることによって，合成した歩容の印象を制御可能な手法を提案している [60] が，原理上，この手法で生成できるのは合成元の歩容の中間の歩容のみである

という問題がある．

このほか，Zhang らはドメイン変換の枠組みを用いて，「child」や「senior」などの 5 種類の年代ラベルを入力とし，その年代らしい歩容を生成する手法を提案している [61]．また，Holden らは，歩容の種類を条件付け可能なモーション生成手法を提案している [62] が，この条件付けには「老人風」，「ゾンビ風」といったクラスラベルが用いられており，このような特定のクラスラベルを用いて学習する手法では，離散的で，ごく限られた特殊なクラスの歩容しか生成できないという制限がある．

より連続的な条件付けによって動作を生成する研究としては，Chen らの音楽を入力としてダンスの動作を生成する研究 [63] などが存在するもの，歩容の生成に関しては，筆者が知る限りでは過去に研究例は存在しない．

1.3　本研究の目的と位置づけ

1.3.1　本研究の目的

ここまでで述べたように，歩容の人間の直感に近い形でのモデル化，すなわち，何らかの直感的なラベル・クエリと歩容とを相互変換する枠組みの提案は従来行なわれてこなかった．本研究の目的は，この直感的なラベル・クエリとして擬態語を利用することにより，歩容の印象の細かな違いを表現し分けることが可能な，人間の直感に近いモデルを構築し，これをもって（1）擬態語による歩容を記述する手法，および（2）擬態語から歩容を生成する手法の二つを提案することである．また，これらの手法で必要となる学習データを確保するために，（3）多様な擬態語アノテーションを付与した歩容データセットの構築も併せて行なう．

1.3.2　本研究のアプローチ

本研究では，図 1.1 に示すように，擬態語を「音韻空間」上で表現したベクトルと歩容との対応関係を獲得することにより，擬態語による歩容の記述・生成を行なう．

本研究で提案する音韻空間は，音象徴性に基づき，擬態語を構成する音素の種類を空間の基底として利用する．例えば，「すたすた」という擬態語は，第 1 子音/s/，第 1 母音/u/，第 2 子音/t/，第 2 母音/a/の四つの音素から構成されていると考えることができる．擬態語で表現される事象の印象は，擬態語を構成する音素によって決まるという音象

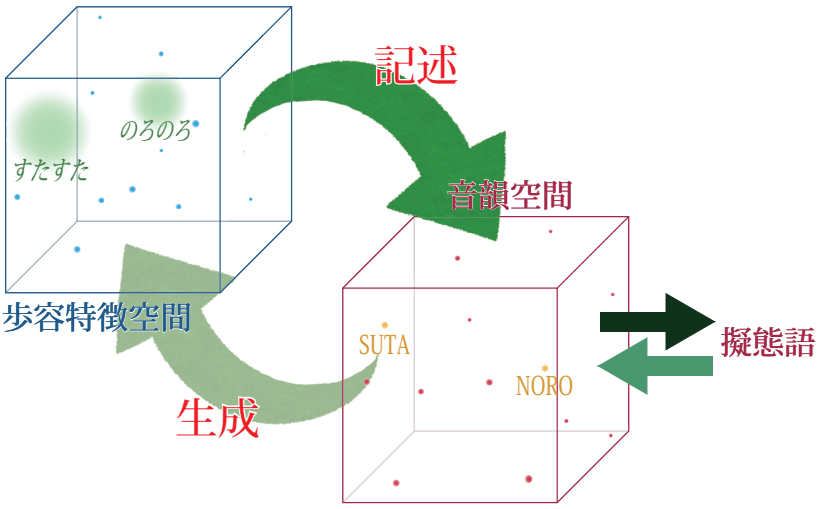


図 1.1　本研究の大まかな枠組み

徴性をふまえると，「すたすた」で表現される歩容の印象は，/s/，/u/，/t/，/a/の四つの変数に 1 が立った 4-hot ベクトルの形で表現することができる．

本研究では一つの歩容に対して複数の擬態語がアノテーションされたデータセットを利用し，アノテーションされた複数の擬態語を音素単位に分解，各音素の出現頻度を集計したものを音韻ベクトル，これによって張られる空間を音韻空間と定義する．ここで，複数の擬態語を集計して利用するのは，擬態語の使い方や，擬態語から受ける印象にはある程度の個人差があると考えられるためである．出現頻度を集計したものは，印象の個人差が，平均的な印象を中心とした音韻空間上でのぶれであるとみなし、これを平滑化したものと考えることができる．このようにして算出した音韻ベクトルは，連続的であり，かつ音象徴性により人間の直感に近い歩容の特徴表現となっていることが期待できるため，これを明示的に歩容の記述モデルや生成モデルに組み込むことにより，歩容の印象の細かな違いを表現し分けることができる記述・生成手法を設計する．

また，モデルを学習するにあたり必要となる，擬態語アノテーションが付与された歩容データセットが従来存在しなかったことから，これを新たに構築する．本研究では歩容の印象の細かな違いを学習させたいことから，普通の歩容だけではない，動きそのものの多様性に富んだ歩容を学習に用いることが望ましい．しかし，一般的な歩容データセットや動作認識データセットの歩行クラスは，歩容の多様性が周囲環境変化や服装の違い程度に留まっており，動きそのものの多様性には乏しいことから，本研究に転用するのは難しい．この問題を解決するため，歩容の映像も新規に撮影したものを用意し，これに擬態語

のアノテーションを付与する．

歩容の記述に際しては，一般的な認識モデルのような，擬態語ラベルを直接学習・出力する多クラス分類タスクとは異なり，歩容から音韻ベクトルを推定するモジュールと，推定した音韻ベクトルを擬態語に変換するモジュールを用意し，2 段階にわけて歩容を記述する．これにより，学習時に用いていないような新奇的な擬態語も含めた多様な出力が可能となる．逆に，歩容の生成に際しては，擬態語クエリを音韻ベクトルに変換するモジュールと，音韻ベクトルに応じて適当な歩容を生成するモジュールを用意し，2 段階にわけて歩容を生成する．

1.3.3　本研究の位置づけ

本研究は，歩容を人間の直感に近い形でモデル化することを試みる先駆的な研究である．これは，計算機に人間の感覚を理解させるということでもあり，より人間味がある人工知能の開発へとつながることが期待できる．

本研究では歩容の記述と生成の二つのタスクを取り扱う．記述タスクに関しては，歩容に特化した精緻な動作認識と解釈することができる．1.1 節で述べた通り，従来の動作認識は「歩いている」,「走っている」，といった粒度の認識をしているものと捉えられるが，本研究はそれよりも細かな違いを表現しわけることが可能である．更に，学習に用いていない擬態語も出力可能な Zero-shot 学習である．感情認識と比較すると，動作によって表現しているものではなく，動作そのものを認識している点で観点が異なる．また，例えば「ふらふら」歩いている人を検出することができれば，そのまま異常検知タスクへの転用が可能である．記述タスクの別の解釈としては，歩容という動きの情報を，擬態語という文字列の形にエンコードしていると捉えることもできる．例えば動画像のクリップをサムネイル画像と擬態語で表現するなどして，超高圧縮率な映像要約に応用することができる可能性もある．一方，生成タスクに関しては，クエリの自由度，任意性の高さという点で，指定できるクラスの種類が明確に決まっていた従来研究とは一線を画している．また，クエリが直感的であることから，直感的インタフェースとしての有用性もあると考えられる．

このほか，記述タスクと生成タスクを組み合わせることにより，例えば歩容を擬態語に変換し，それを歩容に再変換することで，歩容の印象を保持したまま歩容を匿名化する等の応用が考えられる．

1.4　本論文の構成

本論文は全 5 章からなる．第 1 章は序論であり，本研究の背景，目的，位置付けなどを述べた．第 2 章では，本研究を実施するうえで新規に構築した公開データセット「HOYO」について紹介する．次に第 3 章では，擬態語により歩容を記述する手法を提案する．更に第 4 章では，擬態語から歩容を生成する手法を提案する．最後に第 5 章で，本論文を総括し，今後の展望について述べる．

第 2 章

データセットの構築

本章では，本研究で用いるために新たに構築した擬態語がアノテーションされた歩容データセット「HOYO」について紹介する．

第 1 章で述べた通り，擬態語がアノテーションされた歩容データセットは従来存在しない．更に，本研究では歩容の細かな違いをモデルに学習させる必要があることから，歩容自体もできるだけ多様なものが収録されていることが望ましい．1.2.2 項で述べた歩容の認識に関する従来研究において用いられている歩容データセットを概観すると，歩容認証の学習に用いられる大規模な公開データセットはいくつか存在する [64–68] ものの，これらはいずれも個人認証や性別・年齢等の学習目的で収集されたデータセットであり，多様性と言っても周囲環境変化や服装の違い程度に留まっている．本研究では，歩行者の属性ではなく，歩行者の動きそのものの差異を学習したい都合上，歩行者の動きそのものにより多様性があるデータセットを学習に用いるのが望ましく，これらのデータセットをそのまま本研究に転用するのは適切でない．

動作認識に用いられる大規模な公開データセットとしては Kinetics データセット [69] が挙げられる．これは動画共有サイト Youtube* から収集された大量の動画像クリップに 400 種類の動作ラベルが付与されたデータセットであるが，「jogging」，「running on treadmill」などのクラスは存在するものの，大半の歩行・走行動作は「playing tennis」などのより意味的に高次なクラスに内包されてしまっていると考えられ，歩行動作に限ってデータを利用するのは困難である．

このような問題点から，本研究では既存の歩容データセットにアノテーションを付与す

* https://www.youtube.com/

表 2.1　HOYO データセットと既存の歩容データセットとの比較

データセット名	擬態語	歩容の多様性	動画像数	歩行者数
HOYO	○	11 種類（擬態語を表現した動作）	292 本	10 人
OULP [64]	×	2 種類（所持物）	178,018 本	62,528 人
TUM-GAID [65]	×	16 種類（服装・所持物・時間経過）	3,370 本	305 人
CASIA-B [66]	×	4 種類（服装・所持物）	1,240 本	124 人
USF HumanID [67]	×	16 種類（服装・所持物・時間経過）	1,870 本	122 人
SOTON Large [68]	×	3 種類（撮影環境）	2,128 本	115 人

るのではなく，歩容の動画像も新規に撮影した．そして，撮影した多様な歩容に対し，第三者により擬態語のアノテーションを施した．

本データセットと主要な既存の歩容データセットとの比較を表 2.1 に示す．前述の通り，既存の歩容データセットは，基本的に歩容認証に類するタスクに利用することを目的として構築されており，歩行者数が多い一方で，歩容の多様性に乏しい．更に，その多様性もほとんどの場合，服装・所持物などであり，歩容そのものの多様性については考慮せずデータが収集されているという問題点がある．そこで，本データセットの撮影時には，歩容そのものの多様性を豊かにするために，歩行者に対して擬態語による動作の教示を行なう．

以下では，本データセットの構築手順と，アノテーション結果について述べる．まず 2.1 節で，歩容の撮影方法について述べる．次に 2.2 節で，歩容への擬態語のアノテーション方法および，アノテーション結果について述べる．そして 2.3 節で，歩容の動画像から歩容モーションを取得する方法について述べる．更に 2.4 節で，公開した HOYO データセットの仕様を紹介する．最後に，2.5 節で本章をまとめる．

2.1　歩容の撮影

歩行者の前面及び背面から，歩容の動画像を撮影した．側面からの撮影に関しては，十分な長さかつ解像度の動画像を撮影するためには歩行者に合わせてカメラを移動させる，または複数台のカメラを設置する等の大規模な撮影環境の構築が必要であるが，本研究ではそれが困難であったため断念した．奥行き方向の移動による歩行者の大きさの変化を最小限に抑えるために，歩行者から十分離れた位置にカメラを設置した．撮影には Point Gray Research 社製のカメラ Flea3 [70] を用いた．カメラレンズの焦点距離は 35 mm，

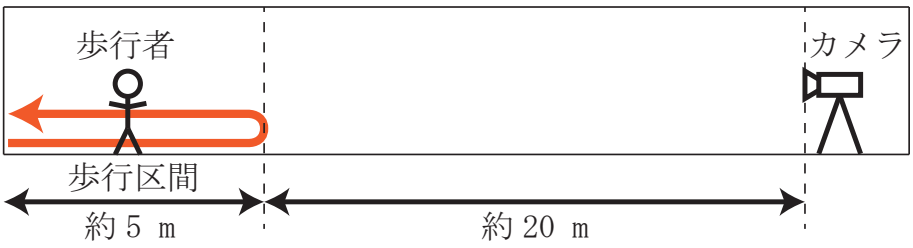


図 2.1　歩容の撮影状況

センサの大きさは 2/3 inch であり，35 mm 判換算の焦点距離は約 138 mm であった．歩容の撮影状況を図 2.1 に示す．歩行区間は約 5 m，歩行区間とカメラとの距離は約 20 m とした．

図 2.1 に示すように，撮影実験協力者は 1 回の試行でまずカメラに近づく向きに歩行し，歩行区間の端に達したところで一旦静止し，180 度向きを変えてカメラから離れる向きに歩行した．各試行において，通常の歩行，「すたすた」，「のろのろ」，「よろよろ」，「どっしどっし」，「せかせか」，「てくてく」，「とぼとぼ」，「のしのし」，「よたよた」，「ぶらぶら」の 11 種類のうち，実験者が指定した 1 種類を表現するよう指示した．協力者及び試行によって異なる擬態語を指示し，各協力者が 6〜16 回試行するようにした．その際に，表 2.2 に示した各擬態語に関する辞書上の意味も参考として提示した．これらの擬態語は，歩行に関する擬態語としてオノマトペ辞典 [10] に掲載されているもののうち，構成する音韻の多様性を考慮して筆者が選択した．この教示の種類を主観ラベルと呼ぶ．主観ラベルは，行動認識データセットにおける動作クラスに相当し，歩行動作のクラスを細分化したものとみなすことができる．このような教示に基づいて被験者を歩行させることにより，動きそのものに多様性がある歩容の動画像を収集した．歩行者は日本語を母語とする 20 代の男性 10 名であった．動画像はすべて 527 × 708 画素，60 fps で撮影し，最終的に 292 本の動画像を得た．撮影した歩容の動画像の例を図 2.2 に示す．

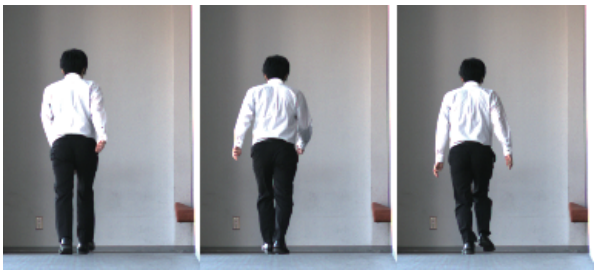
表 2.2　教示として用いた擬態語の辞書上の意味（オノマトペ辞典 [10] より引用）

擬態語	意味
すたすた	足どり軽く見向きもしないで
のろのろ	にぶい動きでなかなか進まず
よろよろ	今にも倒れそうな足取りで
どっしどっし	体重をかけ力強く踏みつけて
せかせか	せかされるように小走りで
てくてく	遠い距離を踏みしめながら
とぼとぼ	遠い距離を肩を落として
のしのし	力強く重々しい足取りで
よたよた	老人や病人が足どり弱く
ぶらぶら	目的もなくさまよい歩いて

(a) 「通常の」歩行（正面）

(b) 「通常の」歩行（背面）

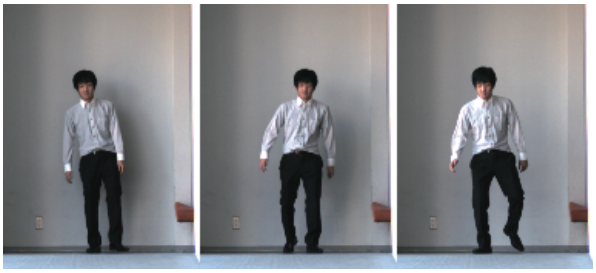
(c) 「すたすた」した歩行（正面）



(d) 「すたすた」した歩行（背面）

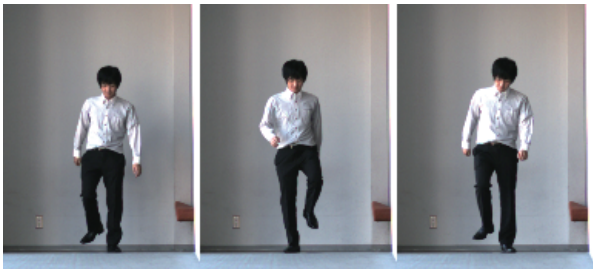
(e) 「のろのろ」した歩行（正面）

(f) 「のろのろ」した歩行（背面）



(g) 「よろよろ」した歩行（正面）

(h) 「よろよろ」した歩行（背面）



(i) 「どっしどっし」した歩行（正面）

(j) 「どっしどっし」した歩行（背面）

(k) 「せかせか」した歩行（正面）

(l) 「せかせか」した歩行（背面）

(m) 「てくてく」した歩行（正面）

(n) 「てくてく」した歩行（背面）

(o) 「とぼとぼ」した歩行（正面）

(p) 「とぼとぼ」した歩行（背面）

(q) 「のしのし」した歩行（正面）

(r) 「のしのし」した歩行（背面）

(s) 「よたよた」した歩行（正面）

(t) 「よたよた」した歩行（背面）

(u) 「ぶらぶら」した歩行（正面）

(v) 「ぶらぶら」した歩行（背面）

図 2.2　歩容の動画像の例

2.2　擬態語アノテーションの付与

データセットの撮影時に，歩行者には特定の擬態語を表現するように教示したが，歩行者が自身のイメージ通りに体を動かせるとは限らないため，得られた歩容は客観的に見て教示された擬態語を表現できているとは限らない．そこで，第三者が歩容を見た際に想起する擬態語を，歩容を表現する真の擬態語と定義し，第三者による評価に基づいて，改めて歩容に対する擬態語のアノテーションを行なった．

本データセットでは，各歩容に 2 種類の擬態語アノテーションを付与した．一つは選択式アノテーションであり，主観ラベルと同様のカテゴリカルな擬態語ラベルを付与するものである．もう一つは自由記述式アノテーションであり，アノテータに複数の擬態語を自由に回答させるものである．以降，2.2.1 項で，選択式アノテーションの方法とその結果を，2.2.2 項で，自由記述式アノテーションの方法とその結果について述べ，2.2.3 項で，アノテーション結果について考察する．

2.2.1　選択式アノテーション

選択式アノテーション実験には，2.1 節で得られた動画像のうち，歩行者の前面を撮影した 146 本を用いた．日本語を母語とする 20 代の男女 14 名のアノテータに対して歩容の動画像を提示し，その歩容に対応すると思う擬態語を，主観ラベルとしても利用した表 2.2 の 10 種の擬態語の中から，複数回答を許して選択させた．また，選択式アノテーション実験のアノテータには撮影実験の協力者が含まれており，自身の歩容に対してアノテーションを行なう場合もあるが，撮影実験から 1 ヶ月以上期間をあけて選択式アノテーション実験を実施することで，影響を低減した．選択式アノテーション実験で用いたインタフェースを図 2.3 に示す．歩容一つあたり 7 名から回答を得て，その過半数である 4 名以上が対応付いていると回答した擬態語を歩容にアノテーションした．各擬態語に対応付いた歩容の数を表 2.3 に示す．表 2.3 の各行が主観ラベル，各列が第三者によって選択された擬態語である．なお，実験の結果，複数の擬態語が過半数票を得る場合や，いずれの擬態語も過半数票を得ない場合があるため，撮影した歩容の本数（146 本）と表 2.3 の合計値は一致しない．具体的には，146 本の歩容のうち，いずれの擬態語もアノテーションされなかったものが 18 本，一つの擬態語がアノテーションされたものが 103 本，二つの

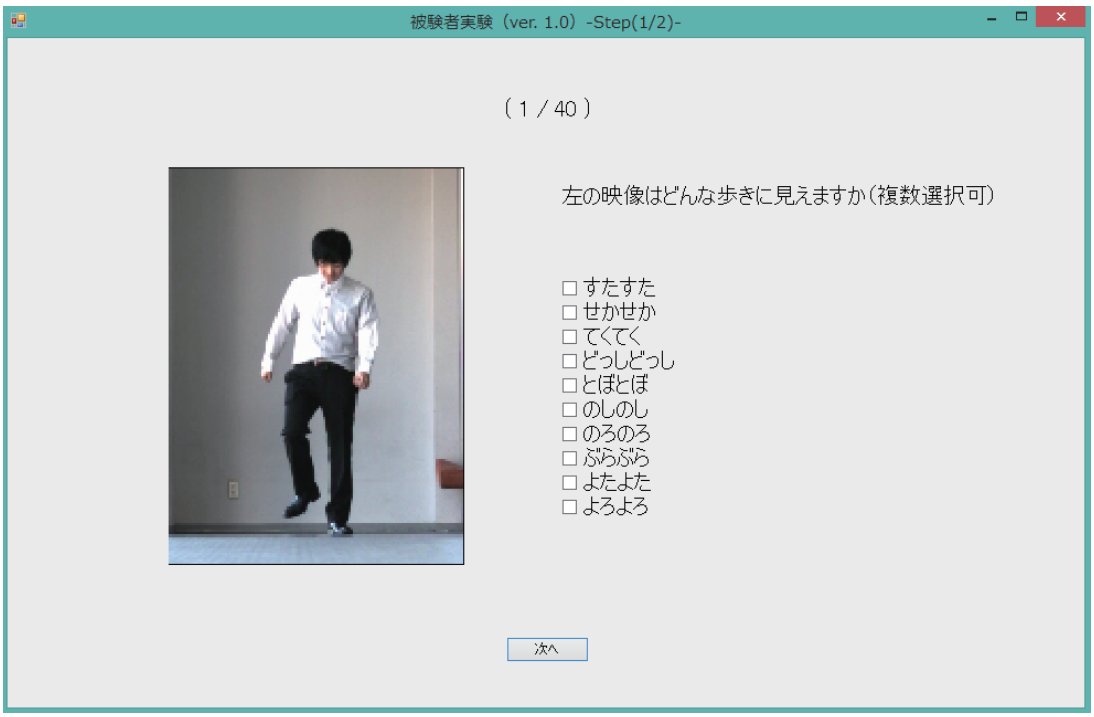


図 2.3　選択式アノテーション実験で用いたインタフェース

擬態語がアノテーションされたものが 25 本存在し，三つ以上の擬態語がアノテーションされた歩容は存在しなかった．ここで，歩行者の背面から撮影した歩容については，対になる前面から撮影した歩容と同じ擬態語をアノテーションした．

表 2.3 のアノテーション結果を見ると，主観ラベルと同様の擬態語がアノテーションされている場合は全体の 55% 程度であり，主観ラベルと客観的な擬態語アノテーションは類似する傾向があるが，一致しない場合も多々あることがわかる．また，「のしのし」と「どっしどっし」，「よたよた」と「よろよろ」など，音素列的に類似した擬態語は混同されやすいこともわかる．さらに，通常の歩容は，擬態語を用いて表現するなら「すたすた」ないし「てくてく」した歩容と表現されることがわかる．

2.2.2　自由記述式アノテーション

2.2.1 項で述べた選択式アノテーションは 10 カテゴリのラベルであり，取り扱いやすい反面，ラベル数の少なさゆえに，付与されたラベルが，歩容から真に想起される擬態語と厳密には一致していない可能性がある．また，回帰モデルを学習する際にも，目標変数（音韻空間上の点）が粗になってしまい，内挿外挿の学習に支障をきたす可能性がある．

表 2.3　第三者による選択式アノテーション結果（擬態語名は略記）

主観	アノテーションされた歩容数										
ラベル	すた	のろ	よろ	どっし	せか	てく	とぼ	のし	よた	ぶら	合計
通常	6	0	0	0	0	11	0	0	0	0	17
すた	15	0	0	0	5	3	0	0	0	0	23
のろ	0	10	0	0	0	0	4	0	0	2	16
よろ	0	0	11	0	0	0	0	0	5	2	18
どっし	0	0	0	12	0	1	0	2	0	0	15
せか	3	0	0	0	8	0	0	0	0	0	11
てく	1	0	0	0	1	5	0	1	0	0	8
とぼ	0	1	0	0	0	0	10	0	0	0	11
のし	0	1	0	4	0	0	2	4	1	0	12
よた	0	1	6	0	0	0	4	0	2	0	13
ぶら	0	0	2	0	0	0	0	0	0	7	9
合計	25	13	19	16	14	20	20	7	8	11	153

そこで，より大規模なアノテーション実験を実施し，アノテータに擬態語を自由記述させることにより，より正確でより多様な擬態語アノテーションを付与する．1.3 節でも述べたように，擬態語の使い方にはある程度個人差があると考えられるので，一つの歩容に対して複数人のアノテータを割り当てるとともに，各アノテータには複数の擬態語を回答させた．

自由記述式アノテーション実験で用いたインタフェースを図 2.4 に示す．アノテータには計算機上で歩容の動画像を見て，その歩容から想起される擬態語を三つ記入するよう指示した．この際，記入する擬態語は ABAB 型（「すたすた」のように 2 音を 2 回繰り返す形式）のものに限定した．また，促音，拗音，長音が付与されたもの（「どっしどっし」等）と，第 2 モーラに撥音を用いたもの（「どんどん」等）も許容した．ここで，モーラとは音韻論上の音の単位であり，本論文では ABAB 語の擬態語と言った場合，A が第 1 モーラおよび第 3 モーラ，B が第 2 モーラおよび第 4 モーラを表している．ABAB 型は，歩容を表す日本語の擬態語の中で最も種類が多い型である [10]．その他，主要な型として「さっさっ」などの AA 型，「しゃなりしゃなり」などの ABCABC 型がある他，「どたばた」，「たたた」などの特殊な形状の擬態語もあるが，第 1 章で述べた通り，本研究におい

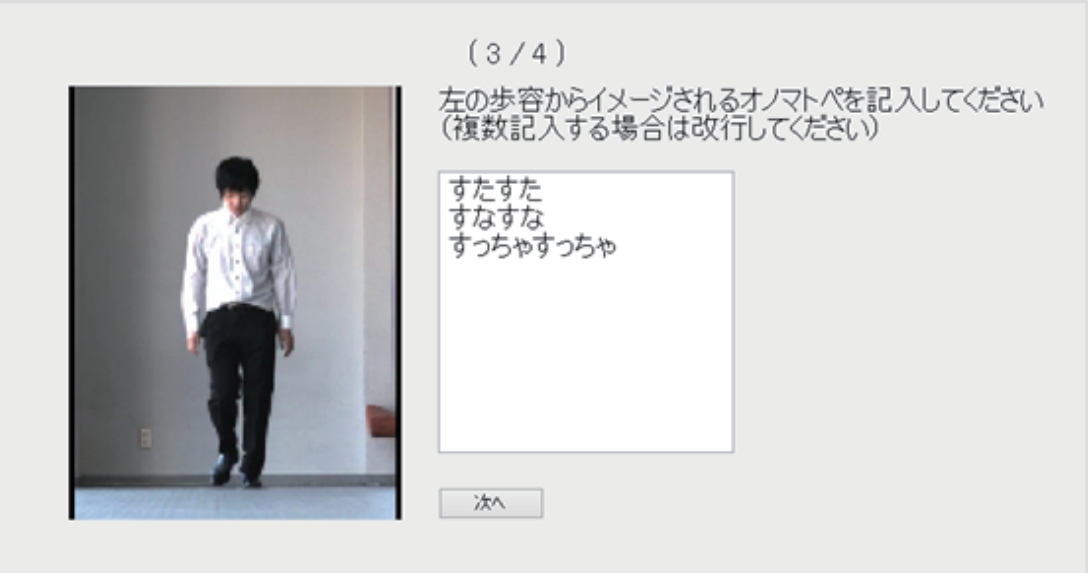


図 2.4　自由記述式アノテーション実験で用いたインタフェース

ては擬態語を音素単位に分解して取り扱いたいことから，擬態語を分解した際に得られる構成要素の数を一定にするために，その形状を ABAB 型に限定する．また，従来研究の中には撥音を拗音や長音などと同じ補助的な要素とみなして，「どんどん」などの語を AA 型，「がしゃんがしゃん」などの語を ABAB 型の擬態語と解釈しているものもある [71] が，前述のように本研究では撥音は母音の一種である [33] と解釈する．

実験には 2.1 節で得られた歩容の動画像のうち，歩行者の前面から撮影した 146 本を用いた．アノテータは，日本語を母語とする男女 30 名であり，歩容一つあたり 15 名から回答を得た．なお，この 30 名には撮影実験参加者は含まれていない．歩行者の背面から撮影した歩容については，対になる前面から撮影した歩容と同じアノテーションを施した．

自由記述アノテーション実験の結果，6,570 件の回答を得て，そのうち 6,322 件が有効な回答であった．例えば，図 2.2(g) に示した歩容に対しては，おぞおぞ，がったがった，ぐらぐら（2），そろそろ，たらたら，ちょんちょん，とことこ，とたとた，とてとて，とろとろ（3），とんっとんっ，どんどん，のしのし（2），のそのそ，のろのろ，はくはく，ひょこひょこ，びくびく，ふらふら（8），ふわふわ，ぶらぶら，へいへい，へなへな，ぺすぺす，ほいほい，ゆっさゆっさ，ゆらゆら（3），よいよい，よたよた（2），よちよち，の計 45 件 31 種類の回答が得られた（括弧内の数字は重複した回答件数）．有効ではない回答としては，入力誤りと思われるものが 9 件，指示を無視して AA 型，ABCABC 型の

擬態語を回答したものが 239 件存在した．

2.2.3　アノテーション結果の考察

データセット全体での擬態語のアノテーション状況について，表 2.4 にまとめる．表 2.4 は出現回数が多い順に，自由記述アノテーション実験で回答された有効な擬態語全 604 種類のうち上位 34 種類の擬態語について掲載している．ここで，自由記述アノテーション実験時には一つの歩容に対して 15 名がアノテーションを施しているため，一つの歩容に対して同じ擬態語が最大で 15 回アノテーションされる可能性がある．ある歩容に対してその擬態語がアノテーションされた回数を調べ，その歩容全体（146 本）での平均と標準偏差を算出したものが表 2.4 の右の 2 列である．出現回数の平均値が大きいほど，その擬態語は歩容を表現するために使われやすい語だと言える．出現回数の標準偏差は，その擬態語が汎用的に使われるのか，何らかの特殊な歩容に対して使われるのかを表していると言える．例えば，「きょろきょろ」と「ぷらぷら」は出現回数の平均は同じだが，標準偏差が大きく異なっている．これは，標準偏差が小さい「ぷらぷら」は多くの歩容に対して薄く広くアノテーションされている一方で，標準偏差が大きい「きょろきょろ」は，一部の歩容に対してアノテーションが集中していることがわかる．

また，得られた擬態語を第 1 モーラ，第 2 モーラに分けて出現頻度を調べたものを表 2.5 に示す．ここで，N は撥音である．出現頻度の平均は，各モーラについて合計が 100 となるように正規化している．第 1 モーラは，出現頻度が高い順に to，su，te，no，hu となっており，特に to と su の音が多い．第 2 モーラは，ra，ta，N，ku，ro の順に頻度が高く，特に ra，sa，N の音が多い．これを表 2.4 の結果と照らし合わせてみると，to は「とんとん」，「とことこ」，「とぼとぼ」など多くの擬態語で用いられることで出現頻度を伸ばしているのに対し，su は「すたすた」の 1 語に支えられている部分が大きいことが推察できる．実際 su は to と比較して，平均ではほぼ同水準であるにもかかわらず，標準偏差が 2 倍近い値となっており，to の音は汎用的に使われるのに対し，su は特定の特徴を持った歩容に集中的に使われていることが伺える．また，一度も出現しなかった音の数は第 1 モーラで 8 種，第 2 モーラで 26 種あり，歩容を表す擬態語は，第 2 モーラで現れる音は第 1 モーラよりも多様性が小さいことがわかる．

表 2.4　自由記述式アノテーション実験で得られた擬態語の出現回数の平均と標準偏差

擬態語	出現回数	出現回数の平均	出現回数の標準偏差
すたすた	547	3.75	4.24
てくてく	445	3.05	2.77
ふらふら	418	2.86	3.64
とんとん	224	1.53	1.45
とことこ	179	1.23	1.29
のろのろ	170	1.16	1.62
ゆらゆら	169	1.16	1.52
とぼとぼ	168	1.15	1.80
のそのそ	157	1.08	1.38
たんたん	121	0.83	0.97
のしのし	119	0.82	1.06
ぶらぶら	104	0.71	0.89
よろよろ	101	0.69	1.15
とろとろ	100	0.68	1.00
どんどん	95	0.65	1.35
どしどし	80	0.55	1.12
くらくら	69	0.47	0.85
ぐらぐら	69	0.47	0.98
てっくてっく	62	0.42	0.70
だらだら	59	0.40	0.75
とてとて	56	0.38	0.62
ずんずん	51	0.35	0.67
たらたら	51	0.35	0.66
すらすら	49	0.34	0.70
すいすい	49	0.34	0.61
のっしのっし	46	0.32	0.67
どすどす	46	0.32	0.79
そろそろ	46	0.32	0.62
すたっすたっ	44	0.30	0.60
きょろきょろ	43	0.29	1.19
ぷらぷら	43	0.29	0.56
のっそのっそ	42	0.29	0.56
かくかく	40	0.27	0.58
ずしずし	37	0.25	0.56

表 2.5: データセット全体での各音の出現頻度の平均と標準偏差

第 1 モーラ			第 2 モーラ		
音	平均	標準偏差	音	平均	標準偏差
a	0.11	0.67	a	0.02	0.18
i	0.63	1.26	i	2.28	2.39
u	0.53	1.03	u	0.15	0.58
e	-	-	e	-	-
o	0.49	1.18	o	-	-
ka	1.29	2.27	ka	1.41	2.12
ki	0.64	1.55	ki	0.46	1.14
ku	1.68	2.50	ku	9.39	7.14
ke	0.05	0.32	ke	0.35	0.95
ko	1.02	2.85	ko	4.56	3.94
sa	1.08	2.05	sa	1.17	1.54
si	0.14	0.55	si	5.79	6.63
su	14.17	14.06	su	1.76	2.68
se	0.41	1.20	se	0.03	0.27
so	1.41	2.15	so	3.76	3.67
ta	4.07	3.24	ta	16.26	10.89
ti	0.20	0.69	ti	0.40	1.06
tu	0.28	0.76	tu	0.79	1.60
te	9.81	7.60	te	2.10	2.75
to	14.45	7.89	to	1.36	1.73
na	0.05	0.32	na	0.31	0.86
ni	-	-	ni	0.14	0.55
nu	0.20	0.70	nu	-	-
ne	0.02	0.19	ne	0.14	0.61
no	9.31	8.30	no	-	-
ha	0.61	1.17	ha	0.02	0.19
hi	0.17	0.60	hi	-	-
hu	8.18	9.26	hu	-	-
he	0.92	1.82	he	-	-

第 1 モーラ			第 2 モーラ		
音	平均	標準偏差	音	平均	標準偏差
ho	0.77	1.26	ho	0.02	0.18
ma	0.02	0.18	ma	0.22	0.67
mi	-	-	mi	0.21	0.80
mu	0.03	0.28	mu	0.08	0.41
me	0.02	0.19	me	0.02	0.18
mo	0.13	0.52	mo	0.08	0.41
ya	0.11	0.48	ya	0.11	0.50
yi	-	-	yi	-	-
yu	3.80	4.27	yu	0.03	0.26
ye	0.02	0.20	ye	-	-
yo	2.71	3.95	yo	0.02	0.19
ra	0.29	1.02	ra	18.57	17.25
ri	0.03	0.28	ri	0.33	0.97
ru	0.61	1.80	ru	1.09	1.57
re	0.08	0.41	re	0.08	0.41
ro	0.02	0.18	ro	8.51	8.84
wa	0.03	0.27	wa	0.58	1.41
wi	0.05	0.33	wi	-	-
wu	-	-	wu	-	-
we	0.02	0.19	we	-	-
wo	-	-	wo	-	-
ga	0.86	2.30	ga	-	-
gi	0.09	0.45	gi	-	-
gu	1.92	3.17	gu	-	-
ge	0.05	0.32	ge	-	-
go	0.07	0.40	go	0.02	0.19
za	0.30	0.81	za	0.06	0.38
zi	-	-	zi	-	-
zu	2.28	3.26	zu	-	-
ze	0.02	0.19	ze	-	-
zo	0.08	0.49	zo	0.02	0.18
da	1.64	2.26	da	0.22	0.77

第 1 モーラ			第 2 モーラ		
音	平均	標準偏差	音	平均	標準偏差
di	-	-	di	-	-
du	0.03	0.26	du	-	-
de	0.37	1.06	de	0.16	0.70
do	5.03	9.16	do	0.05	0.32
ba	0.60	1.20	ba	-	-
bi	0.03	0.28	bi	0.55	1.52
bu	2.12	2.20	bu	-	-
be	0.11	0.49	be	-	-
bo	0.26	0.77	bo	3.15	4.97
pa	0.60	1.15	pa	0.08	0.43
pi	0.20	0.73	pi	0.05	0.33
pu	0.79	1.36	pu	0.02	0.19
pe	0.98	1.48	pe	-	-
po	0.96	1.50	po	0.02	0.20
			N	13.05	11.36

次に，アノテーションされた擬態語をさらに細かく分解し，音素単位で集計したものを表 2.6 に示す．表 2.5 と同様に，出現頻度の平均は，第 1 子音，第 1 母音，第 2 子音，第 2 母音のそれぞれについて総和が 100 となるように正規化されている．ここで，第 1 子音とは，ABAB 型の擬態語の第 1 モーラである A の音の子音，第 1 母音は A の音の母音，第 2 子音は第 2 モーラである B の音の子音，第 2 母音は B の音の母音である．音素単位の分解の場合は，表 2.5 の場合と異なり，全ての変数が非零となっている．第 1 子音は/t/，/s/，/h/，/n/の順に出現頻度が高く，/s/は/t/よりも平均値が小さいにもかかわらず，標準偏差の値が大きくなっている．表 2.5 と照らし合わせてみると，/t/は to の他に te や ta の構成要素としても用いられているのに対し，/s/が使われる場合の大部分が su の音，さらに言えば「すたすた」という語であることがわかる．一般的に/s/の音は滑らかさや素早さを想起させることが多い [72–74] という知見があることから，第 1 子音/s/は素早い動きの歩容に対して集中的に利用されていると考えることができる． 実

表 2.6　データセット全体での各音素の出現頻度（上段）と標準偏差（下段）

第 1 子音	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	1.75	4.68	17.21	28.82	9.57	10.66	0.19	6.63	1.02	0.10	2.98	2.67	7.07	3.12	3.54
	2.07	4.33	15.80	13.23	8.46	9.89	0.63	7.42	2.71	0.46	4.08	3.59	10.39	2.63	2.85
第 1 母音	/a/	/i/	/u/	/e/	/o/										
	11.65	2.18	36.61	12.86	36.70										
	6.57	2.45	12.83	7.67	14.33										
第 2 子音	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	15.49	16.17	12.52	20.92	0.59	0.03	0.59	0.16	28.57	0.58	0.02	0.08	0.42	3.71	0.16
	11.72	10.57	9.35	12.64	1.22	0.26	1.19	0.58	24.40	1.41	0.19	0.42	1.23	4.89	0.59
第 2 母音	/a/	/i/	/u/	/e/	/o/	ん									
	39.02	10.20	13.31	2.87	21.54	13.05									
	13.62	7.46	7.90	3.08	13.15	11.36									

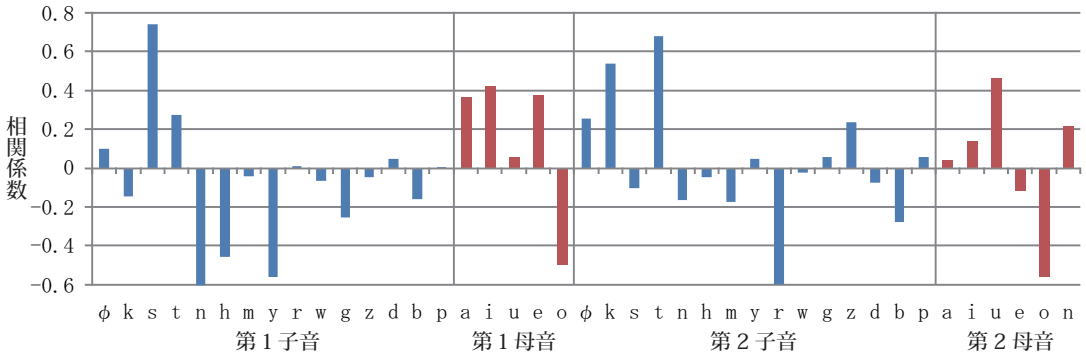


図 2.5　各音素の出現頻度と歩容の速さの相関

際，データセット内の各歩容の速さ（系列長の逆数）と第 1 子音/s/の出現頻度の相関係数を計算すると約 0.74 となり，強い正の相関がみられた．このような各音素の出現頻度と歩容の速さの相関の大きさを列挙したものを図 2.5 に示す．横軸が音素の種類であり，左から第 1 子音，第 1 母音，第 2 子音，第 2 母音の順に並んでいる．縦軸は相関係数である．ここから，前述の第 1 子音/s/の他に，第 2 子音/k/や/t/などが速さと強めの正の相関があることがわかる．母音では，第 1 母音/a/，/i/，/e/，第 2 母音/u/などが比較的正の相関が大きい方であるが，前述の子音ほどの強い相関ではない．ただし，母音は子音と比較して種類が少ないことから，子音以上に複数の用途に用いられ，線形な相関関係でなくなりがちである可能性には留意する必要がある．一方，速さと負の相関がある，すなわち遅い歩容と関わりが深い音素としては，第 1 子音/n/，/h/，/m/，第 1 母音/o/，第 2 子音/r/，第 2 母音/o/などが挙げられる．また，第 2 子音/s/と歩容の速さの相関を調べると，約 −0.10 であり，ほぼ相関がない．このことから，同じ/s/という音素であっても，

図 2.6　CPM [1] で検出される人体の 14 部位

出現する位置によって意味合いが異なっていることがわかる．なお，表 2.5 の第 1 モーラ su の出現頻度と速さの相関は 0.75，表 2.4 の語「すたすた」の出現回数と速さの相関は 0.72 であった．

2.3　歩容モーションの取得

本研究では歩容の骨格情報を利用する．そのため，撮影された歩容の動画像から人体の部位座標系列を取得した．本論文では，このような部位座標系列を歩容モーションと呼ぶ．

まず Convolutional Pose Machines （CPM）[1] を利用し，図 2.6 に示す 14 箇所の部位座標の系列を取得した．CPM は，Pose Machine [75] の畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）版であり，部位座標の尤度マップ推定器を複数段にすることにより，高精度に部位座標を推定する手法である．そして，検出結果が誤っている部分を人手で修正することで最終的な歩容モーションを得た．なお，撮影実験参加者のプライバシーに配慮し，公開されているデータセットには元の動画像を含めずに歩容モーションのみを収録している．

更に，取得した歩容モーションに対して人手で歩容モーションに位相情報を付与した．これは，各歩容モーションにおいて右足が接地する瞬間，および左足が接地する瞬間のフレーム番号を列挙したものである．

2.4　公開データセット

本節では公開した HOYO データセット † の仕様をまとめる．本データセットは歩容モーションと，そこから想起される擬態語によるアノテーションからなる．

歩容モーションの仕様は以下のとおりである．歩容モーションは 60 fps で，各 100〜500 フレームの長さをもつ時系列である．各フレームの表現形式は Euclidean 座標系の 2 次元部位座標であり，関節点の種類は CPM [1] の形式に準ずる，頭，首，右肩，右肘，右手，左肩，左肘，左手，右腰，右膝，右足，左腰，左膝，左足の 14 点である．このような形式の歩容モーションが，前向きの歩容 146 本，後ろ向きの歩容 146 本の計 292 本収録されている．

各歩容モーションに付与されている擬態語アノテーションは以下のとおりである．

- 主観ラベル（11 種類；2.1 節を参照）
- 選択式アノテーション（10 種類；2.2 節を参照）
- 自由記述式アノテーション（擬態語群；2.2 節を参照）

このほか，歩行者 ID（0〜9），歩行者の向き（前向きまたは後ろ向き），歩容の位相情報（2.3 節を参照）が含まれている．

2.5　まとめ

本章では，本研究で用いるために新たに構築した公開データセット，「HOYO」について紹介した．第 1 章で述べた通り，擬態語がアノテーションされた歩容データセットは従来存在しない．更に，本研究では歩容の細かな違いをモデルに学習させる必要があることから，歩容自体もできるだけ多様なものが収録されていることが望ましい．そこで，本データセットでは，歩行者に対して擬態語による教示を行なうことで多様な歩容を撮影した．そして，第三者による選択式アノテーション，自由記述式アノテーションを実施することで，擬態語のアノテーションを付与した．

本データセットは 2.4 節で述べたような内容で Web 上で公開しており，擬態語がアノテーションされた唯一無二の公開歩容データセットとして，本論文で取り組む歩容の記

† https://www.cs.is.i.nagoya-u.ac.jp/ja/opensource/hoyo/

述・生成タスクはもちろんのこと，感性工学，行動認識，異常検出などのさまざまなタスクへの応用が期待される．

第 3 章

擬態語による歩容の記述

第 1 章で述べたように，本論文では擬態語による歩容の記述および，擬態語からの歩容の生成の二つの手法を提案するが，本章ではこのうち，擬態語により歩容を記述する手法について述べる．以降，まず 3.1 節で，本研究の問題設定と，提案手法のアプローチについて述べる．続いて 3.2 節で，提案手法について詳述する．更に 3.3 節で，提案手法の有効性を確認するための評価実験について報告し，3.4 節で考察する．最後に 3.5 節で，本章をまとめる．

3.1　はじめに

本研究では，2.3 節でアノテーションに用いたものと同じ形式の擬態語である，ABAB 型，すなわち「すたすた」のように 2 音を 2 回繰り返すような形式の擬態語を取り扱う．

第 1 章で述べた通り，音韻ベクトルは，音象徴性に基づく表現であるため，人間の直感をよく反映した形式となっていることが期待できる．そこで，提案手法では歩容から End-to-End に擬態語を求めるのではなく，（1）歩容モーションを入力として推定音韻ベクトルを求めるモジュールと，（2）推定音韻ベクトルを擬態語に変換するモジュールの 2 段階に分けて歩容を記述する．中間にこのような人間の直感をよく反映した形式の表現を明示的に導入することで，提案手法も人間の直感に近い歩容記述が可能になると期待できる．モジュール（1）の学習段階では，歩容モーションから特徴量を抽出し，対応する音韻ベクトルとの関係性を回帰モデルに学習させる．記述段階には，学習した回帰モデルに未知歩容モーションを入力することで，推定音韻ベクトルを得る．モジュール（2）では，モジュール（1）が出力した推定音韻ベクトルを入力とし，単一の擬態語を記述結果として

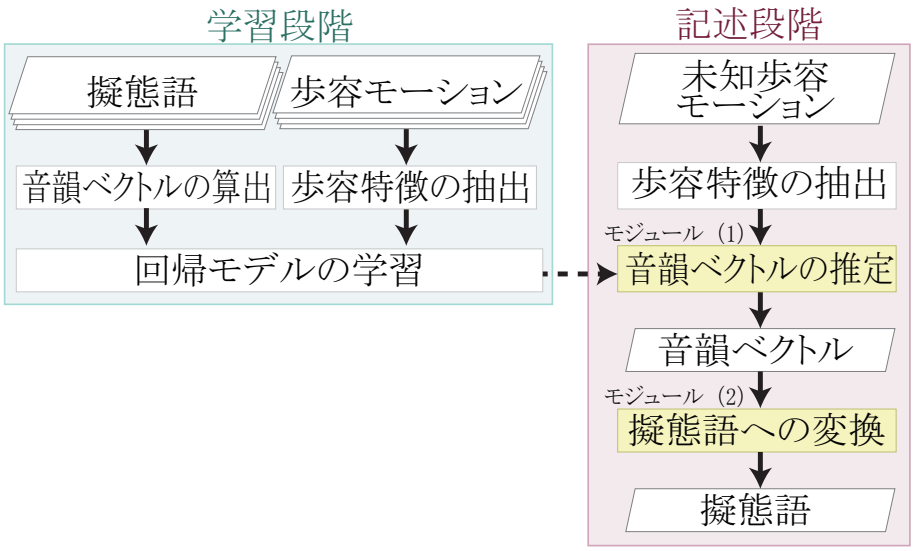


図 3.1　擬態語による歩容記述手法の処理手順

出力する．また，中間表現として音韻ベクトルを導入することで，記述の際に，学習に用いていない新奇的な擬態語を出力することも可能となる．第 1 章で述べた通り，人間は，「とことこ」歩く，「どこどこ」歩く，というように，擬態語を構成する音素の一部を入れ替えることによって微妙な印象の違いを表現しているが，新奇的な擬態語を出力できるようにすることで，提案手法はこの能力をもつ記述モデルとなることが期待できる．擬態語の新奇性の許容範囲は，応用によって異なると考えられるが，提案手法では，モジュール（2）で出力擬態語の自然さを制御するようにすることで，回帰モデルであるモジュール（1）を再学習させることなく，様々な自然さの度合いを持つ擬態語を出力することが可能である．

3.2　擬態語による歩容の記述手法

本節では，音象徴性に基づく歩容の記述手法を提案する．提案手法の処理手順を図 3.1 に示す．学習段階では，まず 3.2.1 項で述べるようにして擬態語から音韻ベクトルを算出する．同時に，3.2.2 項で述べるようにして入力歩容モーションから歩容特徴を抽出する．次に，3.2.3 項で述べるようにして歩容特徴から音韻ベクトルを推定する回帰モデル（モジュール（1））を学習する．記述段階では，未知の歩容モーションを入力として，特徴抽出，学習した回帰モデルによる音韻ベクトルの推定（モジュール（1））を経て，3.2.4 項

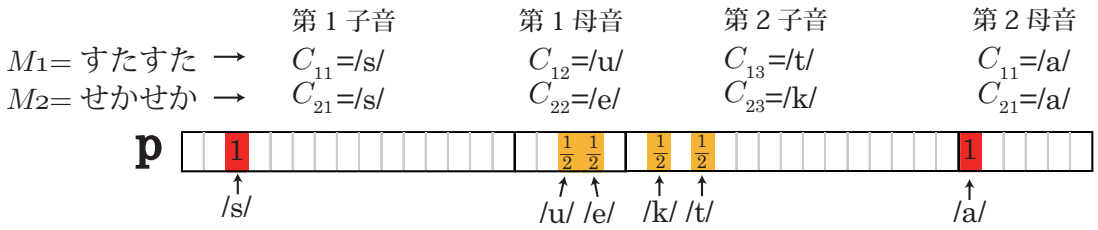


図 3.2　音韻ベクトルの計算例

で述べるようにして推定音韻ベクトルを擬態語に変換する（モジュール（2））ことで，最終的な記述結果を得る．なお，各処理の詳細な実装に関しては 3.3.1 項で述べる．

3.2.1　擬態語の音韻ベクトル化

擬態語を構成する音素から直感的にイメージが想起される，という音象徴の性質を考慮すると，擬態語を構成する音素を特徴の基底とすることで，人間の直感に近い特徴空間を設計できると考えられる．この際，擬態語や歩容から受ける印象には，ある程度個人差があると考えられるため，本研究では複数人によってアノテーションされた複数個の擬態語を利用する．具体的には，学習サンプルとしてある単一の歩容 G に対し複数の擬態語 $M_i \in \{M_1, \ldots, M_n\}$ がアノテーションされているデータについて，これらの擬態語 M_i を音素単位に分解し，各音素の出現割合をベクトルで表現したものを音韻ベクトル $\mathbf{p}$ として定義し，これを特徴量として用いる．このような，複数の擬態語を平均するような処理をすることにより，擬態語や歩容から受ける印象の個人差をある程度軽減できると考えられる．音韻ベクトルの計算例を図 3.2 に示す．ここで，M_i は四つの構成要素 C_{i1}，C_{i2}，C_{i3}，C_{i4} に分解できるとする．このとき，$\mathbf{p}$ の部分ベクトル $\mathbf{p}_j$ を

$$\mathbf{p}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{q}(C_{ij}) \qquad (j = 1, 2, 3, 4) \tag{3.1}$$

と定義する．ここで $\mathbf{q}(\cdot)$ は C_{ij} の音素に該当する変数のみ 1，それ以外が 0 となるような 1-hot ベクトルである．歩容 G に対する最終的な音韻ベクトル $\mathbf{p}$ は，四つの部分音韻ベクトル $\mathbf{p}_1$，$\mathbf{p}_2$，$\mathbf{p}_3$，$\mathbf{p}_4$ を結合させることで得られる．前述の通り，本研究で用いる HOYO データセットにアノテーションされている擬態語は ABAB 型，すなわち「すたすた」のように 2 音を 2 回繰り返すような形式のものである．例えば，歩容に「すたすた」という擬態語がアノテーションされている場合，これは第 1 子音/s/，第 1 母音/u/，第 2 子音/t/，

第 2 母音/a/の四つの要素から構成されていると考える．アノテーションされた擬態語を全てこのように分解し，構成要素ごとに音素の出現頻度を集計する．日本語においては，第 1 母音は/a/，/i/，/u/，/e/，/o/の 5 種類の音素のいずれかであるので，音韻ベクトルのうち第 1 母音に相当する部分音韻ベクトル $\mathbf{p}_2$ は 5 次元であり，その各要素にはそれぞれ，その歩容に付与された擬態語の第 1 母音が/a/，/i/，/u/，/e/，/o/である割合が入ることになる．このように，それぞれの構成要素がとりうる音素の種類が，部分音韻ベクトルの次元数となる．具体的には，第 1 子音および第 2 子音は 15 次元，第 1 母音は 5 次元，第 2 母音は 6 次元であり，音韻ベクトル全体は 41 次元の特徴表現となる．ここで，第 2 母音が 6 次元となっているのは撥音を母音の一種として扱っているためである．第 1 母音には撥音が出現しえないので，第 1 母音は 5 次元としている．また，子音に関しては濁音，半濁音，清音は別の音素として扱っており，第 1 子音，第 2 子音いずれも，子音なし（以降 ϕ で表記），/k/，/s/，/t/，/n/，/h/，/m/，/y/，/r/，/w/，/g/，/z/，/d/，/b/，/p/の 15 種類である．なお，本研究では母音が撥音である場合，子音は ϕ として取り扱う．本手法では，音素の並び順の違いを区別するために，四つの部分音韻ベクトルを結合している．そのため，同じ音素であっても出現する場所が異なる場合は別の変数として区別されることに注意されたい．例えば，「すたすた」と「たすたす」は，どちらも/s/，/u/，/t/，/a/の 4 種類の音素からなるが，並び順が異なるので，異なる音韻ベクトルで表現されることになる．

3.2.2　歩容特徴の抽出

Li らは異常歩容を検出するために人体部位の動きを利用する手法を提案しており，特徴的な歩容を捉える特徴量として，人体部位の動きが有用であることを示唆している [76]．一方，杉山らは犬型ロボットの歩行シミュレータを用いて，被験者に擬態語を表現したロボットの歩行パターンを設計させる実験を行ない，動きに対応した擬態語の種類を人間が判別するためには，肩と足，右足と左足など，体の部位の相対的な運動に着目することが重要であると示唆している [77]．これらをふまえ，提案手法では人体部位の相対的位置関係に基づく特徴を用いることとした．

まず，2.3 節で紹介した HOYO データセット中の歩容モーションにおける，各部位の位置座標系列を Coordinate(p, t) とする．ここで，$p \in \{0, \ldots, P-1\}$ は部位数を P とした時の各部位の識別子である．また，$t \in \{1, \ldots, T\}$ はフレーム番号である．ここで T は

歩容モーションの系列長である．

得られた位置座標系列 Coordinate(p,t) から，任意の 2 部位 $p_1, p_2 \in \{0,\ldots,P-1\}$ のすべての組み合わせにおける部位間相対距離系列 $D_{p_1,p_2}(t)$ を計算する．ここで，相対距離の計算には Euclidean 距離を用いる．

また，各フレームにおける頭の y 座標と，より低い位置にある方の足の y 座標の差 $H(t)$ を計算し，歩容モーション全体での $H(t)$ の平均 $\bar{H}$ を求める．そして，すべての $D_{p_1,p_2}(t)$ を $\bar{H}$ で除することにより，正規化された部位間相対距離系列 $L_{p_1,p_2}(t)$ を得る．

$$L_{p_1,p_2}(t) = \frac{D_{p_1,p_2}(t)}{\bar{H}} \tag{3.2}$$

$$\bar{H} = \frac{1}{T}\sum_{t=1}^{T} H(t) \tag{3.3}$$

ここで，p_1 と p_2 の組み合わせは ${}_P\mathrm{C}_2$ 通りである．

3.2.3　回帰モデルの学習

3.2.2 項で得た正規化された部位間相対距離系列 $L_{p_1,p_2}(t)$ を説明変数，3.2.1 項で求めた対応する音韻ベクトルを目的変数として回帰し，相対距離系列を音韻空間上に射影する回帰モデル f_1 を構築する．学習ののち，未知の歩容モーションから計算された相対距離系列をこのモデルに入力することで，音韻ベクトル $\hat{v}$ を推定できる．

$$\hat{v} = f_1(L) \tag{3.4}$$

ここで，入力 L は一つの歩容映像から得られた相対距離系列をまとめたもので，$T \times {}_P\mathrm{C}_2$ の行列である．

3.2.4　推定音韻ベクトルの擬態語への変換

3.2.1 項で述べた音韻ベクトルの算出方法では，各音韻の共起を一切考慮していない．そのため，例えば 2.3 節で例示した図 2.2（g）に対する擬態語アノテーションから音韻ベクトルを計算すると，図 3.3 のようなベクトルとなるが，この音韻ベクトルを単純な最近傍法で擬態語に変換すると，「ほらほら」なる擬態語が出力されることになる．音韻ベクトルは複数の擬態語アノテーションの統計であるので，音韻ベクトルを単純に単一の擬態語に変換しようとすると，このように不自然な擬態語が出力されてしまうことがある問題

第1子音	φ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	0.0222	0	0.0222	0.2000	0.0889	0.3333	0	0.1778	0	0	0.0667	0	0.0222	0.0444	0.0222
第1母音	/a/	/i/	/u/	/e/	/o/										
	0.0667	0.0222	0.3556	0.0667	0.4889										
第2子音	φ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	0.1333	0.0889	0.1111	0.1333	0.0222	0	0	0	0.4667	0.0222	0	0.0222	0	0	0
第2母音	/a/	/i/	/u/	/e/	/o/	ん									
	0.4889	0.1333	0.0667	0.0222	0.2222	0.0667									

図 3.3　図 2.2（g）の歩容に対応する音韻ベクトル

がある．そこで，新奇的な擬態語を出力する能力は残しつつ，ある程度自然な擬態語を出力できるようにするために，出力候補となる擬態語の不自然さに対して罰則を科すような変換手法を提案する．擬態語の新奇性に関する許容範囲は，応用によって異なると考えられるが，提案手法では，本節で述べる出力擬態語の自然さを制御するモジュール f_2 が 3.2.3 項で述べた回帰モデルから独立しているため，回帰モデルを再学習させることなく，様々な自然さの擬態語を出力できる利点がある．以下では，このような擬態語の自然さを Naturalness と呼ぶ．

このモジュールは，推定音韻ベクトル $\hat{\mathbf{p}}$ を入力とし，Naturalness を考慮して単一の擬態語を出力する．本論文においては，擬態語を構成する音素の共起頻度を Naturalness として利用する．

まず，推定音韻ベクトル $\hat{\mathbf{p}}$ を，第 1 子音，第 1 母音，第 2 子音，第 2 母音にそれぞれ対応する部分ベクトル $\hat{\mathbf{p}}_1$，$\hat{\mathbf{p}}_2$，$\hat{\mathbf{p}}_3$，$\hat{\mathbf{p}}_4$ に分解する．提案手法では，以下の損失関数 $\mathcal{L}$ を最小化するような 4 音素 C_j（$j = 1, 2, 3, 4$）の組み合わせを求め，それを結合することで出力擬態語を得る．

$$\mathcal{L} = \mathcal{L}_d + \alpha \mathcal{L}_c \tag{3.5}$$

$$\mathcal{L}_d = ||\hat{\mathbf{p}}_1 - \mathbf{q}(C_1)|| + ||\hat{\mathbf{p}}_2 - \mathbf{q}(C_2)|| + ||\hat{\mathbf{p}}_3 - \mathbf{q}(C_3)|| + ||\hat{\mathbf{p}}_4 - \mathbf{q}(C_4)|| \tag{3.6}$$

$$\begin{aligned}\mathcal{L}_c = {} & w_{12}N_{12}(C_1, C_2) + w_{23}N_{23}(C_2, C_3) + w_{34}N_{34}(C_3, C_4) \\ & + w_{13}N_{13}(C_1, C_3) + w_{24}N_{24}(C_2, C_4) + w_{14}N_{14}(C_1, C_4)\end{aligned} \tag{3.7}$$

ここで，$\mathbf{q}(\cdot)$ は C_j の音素に該当する変数のみ 1，それ以外が 0 となるような 1-hot ベクトルである．$\mathcal{L}_d$ は，推定音韻ベクトル $\hat{\mathbf{p}}$ と，任意の音素の組み合わせから構成される 4-hot 音韻ベクトル（出力候補となる擬態語）との音韻空間上の距離を計算している．$\mathcal{L}_c$ は，不自然な音素の組み合わせに対して罰則を与える項である．なお，α は調整用のパラメータであり，$\alpha = 0$ の場合は単純な最近傍法と等価となる．$N_{jj'}(\cdot)$ は，2 音素の組み合わせに対する罰則項である．例えば，$N_{12}(C_1, C_2)$ という項は，第 1 子音 C_1 と第 1 母

音 C_2 の組み合わせが不自然な場合に値が大きくなる．最終的に，$\mathcal{L}$ を最小化するような C_1，C_2，C_3，C_4 の組み合わせを求め，この 4 つの音素を並べてできる擬態語を出力する．

$N_{jj'}(\cdot)$ の具体的な値は，第 2 章で述べた HOYO データセットに付与されている擬態語群から算出する．まず，自由記述アノテーションで得られた擬態語群を音素に分解し，2 音素の組み合わせで共起ヒストグラムを計算する．ここで，2 音素の組み合わせは ${}_4\mathrm{C}_2 = 6$ 通りであるため，6 種類の共起ヒストグラムが得られる．このようにして得られた 6 種類の共起ヒストグラム $N'_{jj'}(\cdot)$ は，よく共起する音素の組に対して高い値を返すので，罰則項として用いるために $N_{jj'}(\cdot) = 1 - (N'_{jj'}(\cdot)/N_{\mathrm{words}})$ とする．ここで，N_{words} は計算に用いた擬態語の数であり，実際には $N_{\mathrm{words}} = 6,322$ である．

$w_{jj'}$ は，六つある罰則項 $N_{jj'}(\cdot)$ のそれぞれの重みの大きさを示す．具体的にこの重み $w_{jj'}$ の値を決定するために，予備実験を実施した．本予備実験では，人間による主観評価によって得られる擬態語の自然さと，提案する罰則項から算出される擬態語の自然さができるだけ一致するような重み w の値を求めることを考える．まず，任意の音素の組み合わせで得られる全ての ABAB 型の擬態語 6,750 語（$= 15 \times 5 \times 15 \times 6$）について，6 つの w 全てが同じ値の条件下で $\mathcal{L}_c$ を計算し，$\mathcal{L}_c$ が小さい順に並び変えた．6,750 語すべてについて主観評価を実施するのは困難であるため，代表として 10 種類の擬態語を等間隔に抽出した．具体的には，$\mathcal{L}_c$ が小さい順に，「ゆらゆら」，「ぐりぐり」，「ずぜずぜ」，「さこさこ」，「どねどね」，「まばまば」，「ろやろや」，「ぷぷぷぷ」，「はぞはぞ」，「はぺはぺ」の 10 種類であった．続いて，この 10 種類の擬態語を用いて Thurstone の一対比較法 [78] による自然さの評価実験を行ない，その結果に基づいて 10 種類の語を自然な順に並び変えた．一対比較実験では，10 種類の語のうちの 2 語を評価者に提示し，擬態語としてより自然だと思う方を回答するよう指示した．設問数は ${}_{10}\mathrm{C}_2 = 45$ 問であり，評価者は 20 代の男性 4 人であった．この一対比較実験の結果に基づき，10 語を自然な順（選択された回数が多い順）に並べ替えた．具体的には，自然な順に，「さこさこ」（34 回），「ゆらゆら」（33 回），「ずぜずぜ」（21 回），「はぺはぺ」（20 回），「ぐりぐり」（18 回），同率で「どねどね」と「はぞはぞ」（16 回），「まばまば」（13 回），「ろやろや」（8 回），「ぷぷぷぷ」（1 回）の順番であった．括弧内の数字は一対比較実験で選択された回数である．最後に，一対比較実験で得られた 10 語の自然さの順位と，様々な $w_{jj'}$ の条件下で算出された $\mathcal{L}_c$ による 10 語の自然さの順位が最も近くなるような $w_{jj'}$ の値をグリッドサーチで求めた．各 $w_{jj'}$ の値は 1 刻みで 0 から 9 までの 10 段階を試行した．順位の類似尺度に

表 3.1　第 1 子音と第 2 子音の共起回数

第 1 子音	第 2 子音														
	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
ϕ	2	5	19	7	2	0	1	5	66	1	0	1	2	0	0
/k/	15	47	12	50	7	0	3	0	130	0	0	0	0	31	1
/s/	83	50	54	706	2	0	1	0	131	5	1	0	0	26	6
/t/	416	797	25	199	0	0	10	1	189	0	0	0	0	175	0
/n/	2	8	384	28	0	0	0	0	190	0	0	0	0	0	0
/h/	64	26	14	27	21	1	21	3	490	15	0	0	0	0	1
/m/	1	2	2	6	0	0	0	0	1	0	0	0	0	0	0
/y/	9	4	17	92	1	0	0	0	300	1	0	0	0	5	0
/r/	53	0	2	5	0	1	0	0	3	0	0	0	0	0	0
/w/	5	0	0	0	0	0	0	0	1	0	0	0	0	0	0
/g/	23	5	39	11	4	0	0	0	77	10	0	0	21	0	0
/z/	87	14	45	7	0	0	0	0	12	1	0	4	1	0	0
/d/	148	37	155	21	0	0	1	1	84	0	0	0	3	1	0
/b/	41	5	10	28	0	0	0	0	113	1	0	0	0	0	0
/p/	28	5	15	120	1	0	1	0	47	3	0	0	0	0	2

は Spearman の順位相関係数 [79] を用いた．

グリッドサーチの結果，$w_{12} = 0$, $w_{23} = 0$, $w_{34} = 1$, $w_{13} = 9$, $w_{24} = 0$, $w_{14} = 1$ とした場合に，順位相関が最も高い値 0.8389 となった．ここで，最も重みが大きい w_{13} は第 1 子音と第 2 子音の共起，次点である w_{34}，w_{14} はそれぞれ第 2 子音と第 2 母音の共起，第 1 子音と第 2 母音の共起に対応する重みであり，擬態語の自然さにはこれらの音素の組み合わせが特に重要であることがわかる．

表 3.1 に第 1 子音と第 2 子音の共起回数の表を示す．最も共起回数が多いのは第 1 子音/t/と第 2 子音/k/の組であり，HOYO データセット中の全擬態語 6,322 語のうち 797 語がこのパターンである．これは「とことこ」,「てくてく」などといった，よく使われる擬態語でよく見られるパターンであり，擬態語がこのようななじみ深い共起パターンを持っていると，人はその擬態語を自然に感じやすいと考えられる．

この値を式 3.7 にあてはめて書き下すと，

$$\mathcal{L}_c = C_{34}(o_3, o_4) + 9C_{13}(o_1, o_3) + C_{14}(o_1, o_4) \tag{3.8}$$

であり，以下ではこの重みの値を用いて実験を行なう．

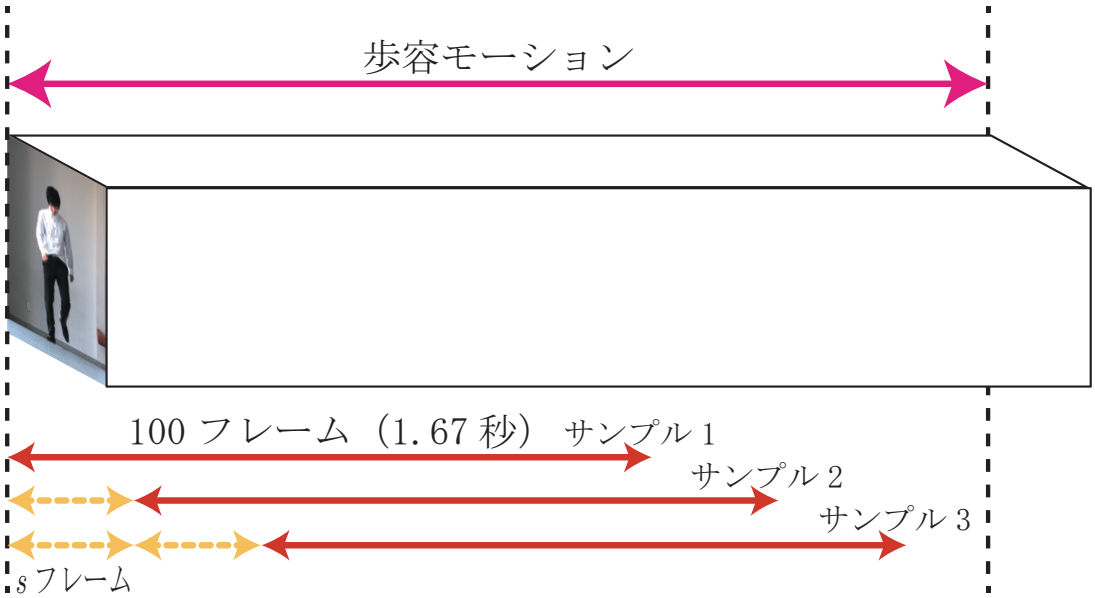


図 3.4　サンプルの切り出し方法

3.3　評価実験

本節では，3.2 節で提案した手法の評価実験について報告する．まず 3.3.1 項で，評価実験を行なった際の提案手法の実装について述べる．次に 3.3.2 項で，提案手法における音韻ベクトルを擬態語に変換するモジュールを評価するための主観評価実験について報告する．続いて 3.3.3 項で，提案手法における回帰モデルおよび音韻空間の定義方法について評価するための実験について報告する．

3.3.1　実装

3.3.1.1　サンプルの作成

第 2 章で紹介したデータセットは，歩容モーションの長さが一定ではないため，これを実験で扱いやすくするために，元の歩容モーションから固定長の部分歩容モーションをサンプルとして切り出した．具体的には，図 3.4 に示すように，開始フレームを s フレームずつずらしながら 100 フレーム分（約 1.67 秒）の歩容モーションを順次切り出した．すなわち回帰モデルに入力するサンプルの長さ $T = 100$ とした．長さを 100 フレームとしたのは，歩容の 1 周期（2 歩）が十分収まる長さであるためである．ずらし幅は経験的に

表 3.2　実験に用いた CNN の構造

Input	Units: 100	Channels: 91
Convolution 1	Kernel size: 10	Channels: 128
	Max-pooling size: 10	
Convolution 2	Kernel size: 10	Channels: 128
	Max-pooling size: 10	
Output	Units: $4N_p$	

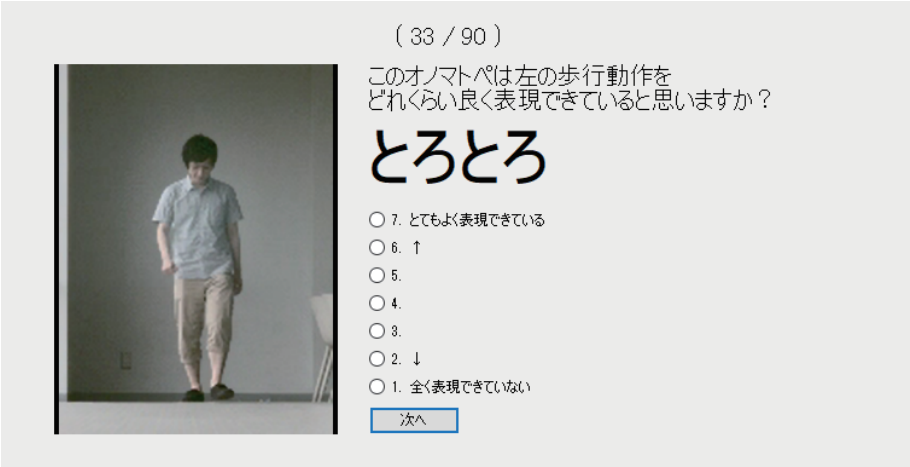


図 3.5　Correctness の主観評価実験に用いたインタフェース

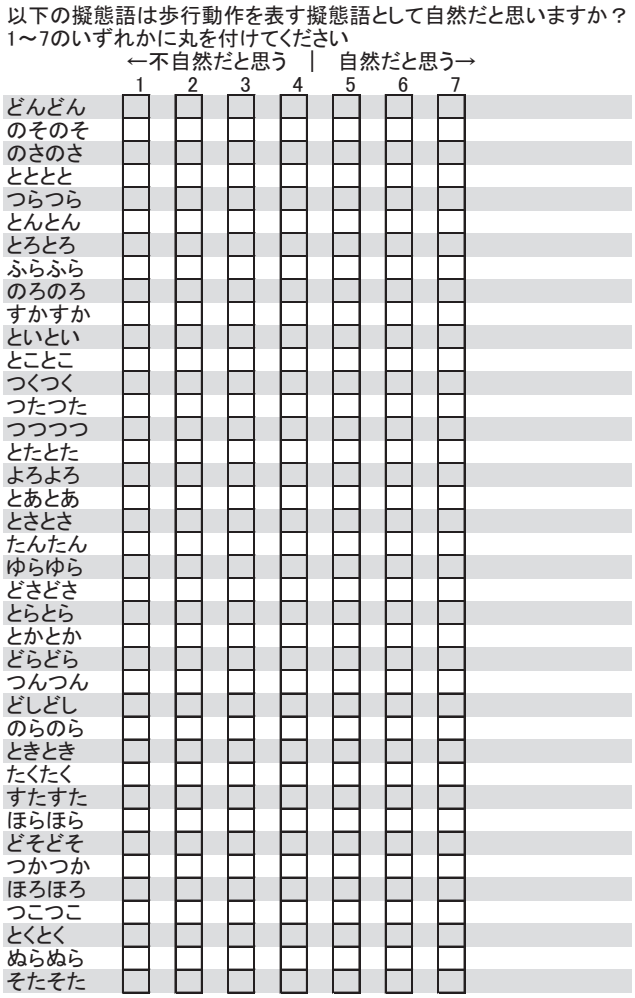
$s = 5$ とした．

3.3.1.2　モデルアーキテクチャ

本論文では回帰モデルとして，深層学習モデルの一種である 1 次元の畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）を用いた．CNN は入力層で正規化された部位間相対距離系列 $L_{p_1,p_2}(t)$ それぞれをチャネルとみなしてチャネル数 91，ユニット数 100 の入力を受け付け，出力層は 3.2.1 項で述べた 41 次元の音韻ベクトルを出力する．実験で用いた CNN の構造を表 3.2 に示す．これらの深層学習モデルの実装には Keras [80] を用い，パラメータは実験的に設定した．

3.3.2　音韻ベクトルを擬態語に変換するモジュールの評価

音韻ベクトルを擬態語に変換するモジュールの有効性を評価するために，このモジュールが出力した擬態語に関して評価実験を行なった．本実験では，出力された擬態語につい


以下の擬態語は歩行動作を表す擬態語として自然だと思いますか？
1～7のいずれかに丸を付けてください

	←不自然だと思う						自然だと思う→
	1	2	3	4	5	6	7
どんどん							
のそのそ							
のさのさ							
とととと							
つらつら							
とんとん							
とろとろ							
ふらふら							
のろのろ							
すかすか							
といとい							
とことこ							
つくつく							
つたつた							
つつつつ							
とたとた							
よろよろ							
とあとあ							
とさとさ							
たんたん							
ゆらゆら							
どさどさ							
とらとら							
とかとか							
どらどら							
つんつん							
どしどし							
のらのら							
ときとき							
たくたく							
すたすた							
ほらほら							
どそどそ							
つかつか							
ほろほろ							
つこつこ							
とくとく							
ぬらぬら							
そたそた							

図 3.6　Naturalness の主観評価実験に用いた質問用紙

て，擬態語が元の歩容をどれだけ正確に表現できているか（Correctness），擬態語が自然かどうか（Naturalness）の 2 つの観点で評価した．評価はいずれも主観評価実験により，7 段階の Likert 尺度を用いて評価を行なった．変換元となる音韻ベクトルは，回帰モデルによる推定値ではなく，HOYO データセットから算出した真値を用いた．

Correctness の評価実験では，評価者に歩容映像と擬態語のペアを提示し，このオノマトペがどの程度歩容を表現できていると思うかを質問した．ここで，オノマトペとは擬態語を包含する概念を指す用語 [10] であり，ここでは擬態語とほぼ同義である．Correctness の主観評価実験に用いたインタフェースを図 3.5 に示す．本実験では，$\alpha = 0, 1, 3, 6$ の四つの条件で歩容から擬態語を出力し，それぞれの歩容-擬態語ペアについて主観評価を実施した．3.2.4 項で述べた通り，α は出力擬態語の自然さをどの程度重視するか決定する

表 3.3　Correctness および Naturalness の主観評価値

条件	Correctness	Naturalness
$\alpha = 0$	4.434 ± 0.109	4.962 ± 0.109
$\alpha = 1$	4.452 ± 0.088	5.217 ± 0.077
$\alpha = 3$	4.275 ± 0.053	5.356 ± 0.043
$\alpha = 6$	4.192 ± 0.067	5.553 ± 0.052

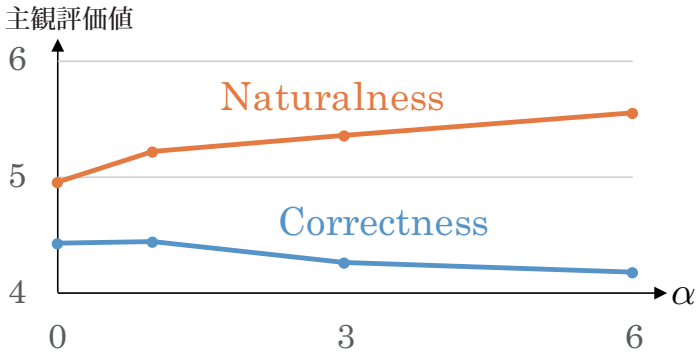


図 3.7　Correctness および Naturalness の評価結果のグラフ

ハイパパラメータであり，α が大きいほど自然さを重視した擬態語が出力される．また，$\alpha = 0$ のときは単純な最近傍法と等価である．評価者は 5 人であり，全員が日本語を母語とする大学生であった．

Naturalness の評価実験では，評価者に擬態語を提示し，その擬態語が歩行動作を表す擬態語としてどのくらい自然だと思うかを質問した．質問は図 3.6 に示したような紙面上で行ない，評価者は日本語を母語とする大学生 4 名であった．こちらの評価実験でも $\alpha = 0, 1, 3, 6$ の四つの条件で擬態語を出力し，得られた全ての擬態語を提示に用いた．

表 3.3 および図 3.7 にこれらの評価実験の結果を示す．表の値は Likert 尺度の平均 ± 標準誤差である．結果から，α の値が大きくなるにつれて，Naturalness が上昇していくことがわかる．また，Correctness と Naturalness はトレードオフの傾向にあるが，$\alpha = 1$ の条件においては，$\alpha = 0$ の条件と比較して Correctness を維持したまま Naturalness が上昇しており，提案手法は歩容の記述の正確さを損なわずに，より自然な擬態語を出力できることを確認できたと言える．

3.3.3　回帰モデルおよび音韻空間の評価

本研究では 3.2.1 項で定義した音韻ベクトルによって張られる音韻空間を用いて歩容をモデル化している．この音韻ベクトルの定義が妥当であること，および音韻ベクトルを基底として学習した回帰モデルの有効性を確認するために，異なる定義による音韻ベクトルを用いて評価実験を行なった．なお，回帰モデルの学習および擬態語の出力は，HOYO データセットの歩行者 ID に基づく Leave-one-person-out 交差検証により行なった．

本実験では，比較手法として，提案手法とは異なる定義の音韻ベクトルを使用した手法 2 種類を実装して比較した．以下ではこの異なる実装を比較手法 A，比較手法 B と呼ぶ．それ以外の実装や，学習に用いたデータなどは提案手法と同一とした．

比較手法 A

提案手法における音韻ベクトルは，ひとつの ABAB 型の擬態語を四つの音素に分解して扱ったが，比較手法 A ではこの分解方法を変更し，文字単位で分解する．例えば，「すたすた」という語は，第 1 モーラ「す」と第 2 モーラ「た」が 2 回繰り返されているものであると考え，「す」と「た」の二つの要素に分解する．それ以外は提案手法と同じように，ある歩容にアノテーションされた擬態語を全て分解し，構成要素ごとに音の出現頻度を集計することで音韻ベクトルを得る．このとき，提案手法と同様に，濁音，半濁音は清音とは別の種類の音として数える．このようにして算出された音韻ベクトルは第 1 モーラとして出現しうる音が 75 種類，第 2 モーラとして出現しうる音が 76 種類であるため，合計 151 次元のベクトルとなる．第 1 モーラの次元が一つ少ないのは，第 1 モーラには撥音が出現しないためである．

比較手法 B

比較手法 B では，まず提案手法と同じように擬態語を四つの音素に分解したのち，秋山らによって提案されている音素の数値化表 [33] に基づいて各音素を 4 次元のベクトルに変換する．秋山らが提案するこのベクトルは，一つの音素に対して「キレ・俊敏さ」，「柔らかさ・丸み」，「躍動感」，「大きさ・安定感」の四つの属性値が割り当てられたもので，これらは主観評価実験と因子分析によって決定された実数値である．具体的には，まず音素のそれぞれについて，大学生 115 名を対象とした，音の印象を問う主観評価実験が実施

表 3.4　秋山らによる 4 次元属性ベクトル [33]

		キレ・俊敏さ	柔らかさ・丸み	躍動感	大きさ・安定感
母音	/a/	0.05	0.29	0.83	1.38
	/i/	0.71	−0.88	0.15	−1.23
	/u/	−0.73	0.73	−0.95	−0.02
	/e/	0.29	−0.45	−0.08	−0.61
	/o/	−0.69	0.55	−0.15	1.83
	/N/	−0.77	0.08	−1.73	0.28
子音	/k/	2.05	−2.43	1.54	−0.46
	/s/	1.67	−0.92	1.15	−1.55
	/t/	1.20	−1.51	1.13	0.08
	/n/	−1.26	0.94	−1.55	0.04
	/h/	−0.01	0.45	0.17	−0.26
	/m/	−1.42	1.31	−1.36	0.82
	/y/	−0.75	0.74	−0.47	−1.43
	/r/	0.10	0.31	0.67	−0.37
	/w/	−0.43	0.78	0.65	1.51
濁音		−0.07	−1.57	−0.28	0.87
半濁音		0.36	0.76	0.88	−0.83
ゃ		−0.60	0.51	−0.38	−0.75
ゅ		−0.17	0.54	−0.59	−0.60
ょ		−0.25	0.62	−0.38	−0.60
促音		1.97	−1.34	1.83	0.10

され，「柔らかい-かたい」，「温かい-冷たい」などの 43 項目の形容詞対 [81–84] に対して 5 段階の Likert 尺度で回答が得られた．そして，項目間の相関行列に対してステップワイズ変数選択法 [85] を用いた因子分析を繰り返し施し，独立性が低い因子を除外していくことで，最終的に上記の 4 属性が得られた．この具体的な数値を表 3.4 に示す．提案手法とは異なり，この手法では濁音と半濁音を，対応する清音の子音の値を補正することで表現している．例えば，子音/g/は，子音/k/の値に濁音の補正をかけた値を用いる．なお，子音がない（あ行の）場合は零ベクトルを用いた．この 4 次元ベクトルを結合することで，一つの擬態語から 16 次元のベクトルを得る．ある歩容にアノテーションされた全ての擬

表 3.5　音韻空間の比較結果

音韻ベクトル実装	主観評価値	音韻ベクトル誤差
提案手法（音素単位）	4.751 ± 0.073	0.396
比較手法 A（文字単位）	4.645 ± 0.064	0.750
比較手法 B（秋山ら [33] による数値化）	3.596 ± 0.072	0.367

態語からこのベクトルを算出し，平均をとったものを音韻ベクトルとする．

評価は主観的な尺度と客観的な尺度の両面で行なった．主観的な尺度として，3.3.2 項の Correctness の評価に準ずる主観評価値（7 段階の Likert 尺度），定量的な評価尺度として音韻ベクトルの推定誤差に関する全テストサンプルの平均絶対誤差（Mean Absolute Error; MAE）を用いた．なお，主観評価に際する推定音韻ベクトルの擬態語へ変換は，式 3.5 の $\alpha = 0$ の条件で行なった．推定音韻ベクトルはテストサンプルと同じ数だけ出力されるが，本実験では主観評価の設問数削減のため，同一の歩容から切り出されたサンプルから得られた複数の推定音韻ベクトルの平均をとり，その平均ベクトルを擬態語に変換したものを評価に用いた．また，次元数が異なる音韻ベクトル実装の誤差の値を公平に比較するために，各音韻ベクトル誤差の値は，各音韻ベクトル真値の平均 L1 ノルム（提案手法は 4，比較手法 A は 2，比較手法 B は 0.0793）で正規化した．

評価結果を表 3.5 に示す．表中の主観評価値は Likert 尺度の平均 ± 標準誤差である．主観評価実験の評価者は全部で 16 名であり，1 問あたり 5 名が回答するようにした．評価者は全員日本語を母語とする大学生であった．

この結果から，比較手法 A は音韻ベクトルの誤差で，比較手法 B は主観評価値で，それぞれ提案手法よりも評価が低下しており，提案手法は主観的な尺度，客観的な尺度の双方の評価を，比較的高い水準で両立していると言える．また，筆者の先行研究 [86] 上で実施した評価実験（実験手順は本実験の主観評価と同様，被験者数 9 名）において，歩容映像と，その歩容とは無関係な，無作為な音素の組み合わせにより作成した擬態語の組を用いた主観評価結果が 4.014 ± 0.081 となっており，提案手法と比較手法 A はこの水準を上回っている．以上から，提案手法は，少なくとも無作為な音素の組み合わせにより作成した擬態語よりは正確な擬態語を出力できていることを確認できた．比較手法 B は，主

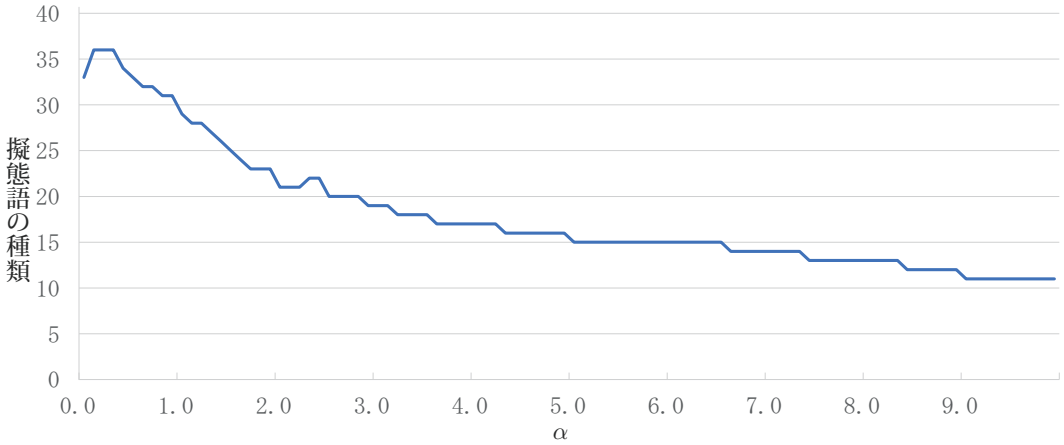


図 3.8　α を徐々に変化させた時の出力擬態語の種類の変化

観評価において無作為な音素の組み合わせにより作成した擬態語を出力する条件（4.014）よりも低いという問題がある．比較手法 A は，主観評価において提案手法に匹敵する結果を示しているが，実際に出力された擬態語を見てみると，全 146 本の歩容に関して，比較手法 A が 13 種類の擬態語を出力したのに対し，提案手法は 16 種類の擬態語を出力している．出力擬態語の自然さは，音韻ベクトルから擬態語に変換する段で制御したいことを考慮すると，音韻ベクトルの時点ではなるべく多様な出力が得られていることが望ましく，提案手法のほうが望ましい性質を持っていると言える．また，比較手法 A の音韻ベクトル誤差が提案手法や比較手法 B と比較して大きくなっているのは，提案手法の音韻ベクトルが 41 次元，比較手法 B の音韻ベクトルが 16 次元であるのに対して，比較手法 A の音韻ベクトルが 151 次元と大きく，次元の呪いの影響を受けている可能性が考えられる．

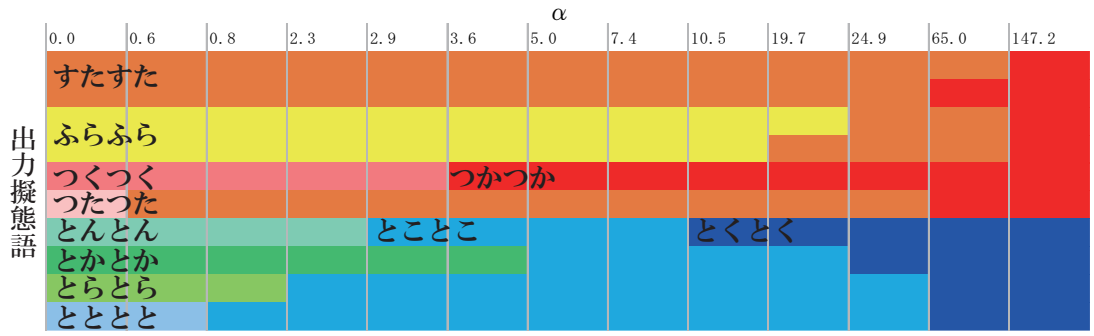
3.4　考察

3.4.1　罰則項の重みによる記述結果の変化について

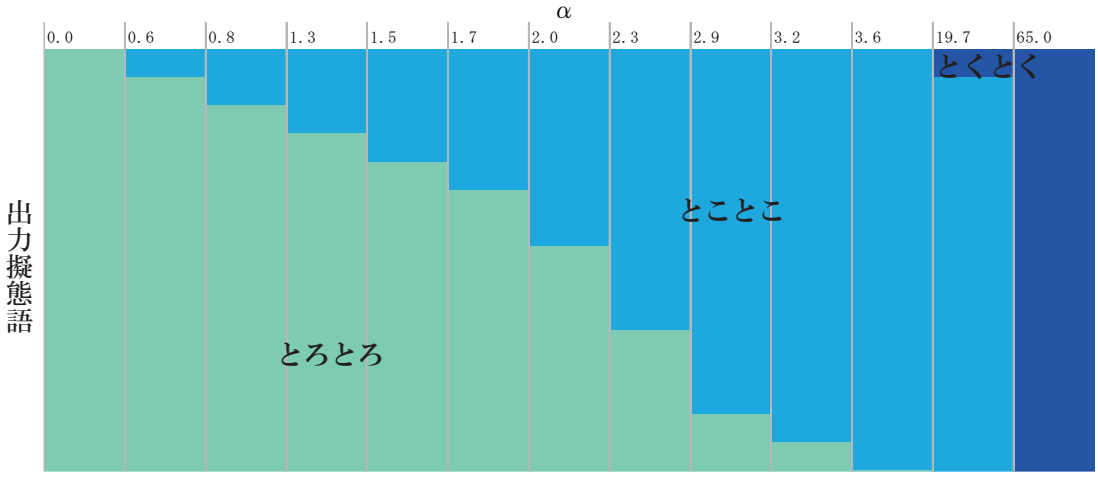
提案手法には，自然さの考慮具合を調整するハイパパラメータ α が存在する．3.3.2 項の評価実験においては，主観評価の手間を考慮し，$\alpha = 0, 1, 3, 6$ とした場合の出力擬態語に絞って実験を行なったが，ここでは，この α を徐々に変化させた時の、出力擬態語の推移について調べ，適切な α の値や，提案手法の特性，問題点などについて考察する．具体的には，α を 0 から 0.1 刻みで変化させて擬態語を出力し，その推移を調べた．

まず，α を徐々に変化させた時の出力擬態語の種類数の変化を図 3.8 に示す．ここで図

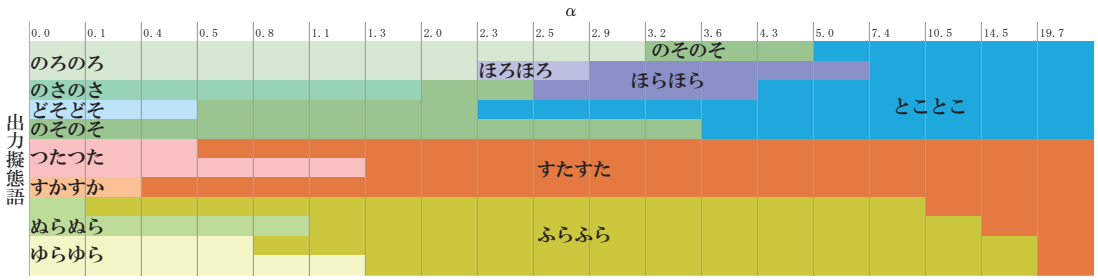
の横軸は α の値，縦軸が出力擬態語の種類である．$\alpha = 0$ のとき擬態語は 33 種類であり，そこから α が大きくなるにつれて，基本的には擬態語の種類は減少していく傾向が見て取れる．$\alpha = 1.0$ では 29 種類，$\alpha = 3.0$ では 19 種類，$\alpha = 6.0$ では 15 種類であった，最終的には $\alpha = 223.7$ で 3 種類まで減り，これ以後変化しなかった．最終的に残った 3 種類の擬態語は「つくつく」,「とくとく」,「たくたく」であった．この 3 種類はいずれも第 1 子音/t/，第 2 子音/k/，第 2 母音/u/の組み合わせからなっており，第 1 母音のみが異なっている．これは提案手法で用いた自然さに関する罰則項に第 1 母音に関する制約が含まれていないためである（式 3.8 を参照）．具体的に，α の変化に応じた擬態語の変化の様子の例を図 3.9 に示す．縦軸がサンプル（歩容），横軸が α の変化であり，横軸は各サンプルの出力に変化があった α のみを抽出して表記している．図 3.9(a) の例を見ると，「すたすた」,「ふらふら」など，元から一般的な擬態語は，α が増大しても変化しにくいことがわかる．図 3.9(b) の例では $\alpha = 0$ の条件下で「とろとろ」と記述されるサンプル群に着目しており，これらは α が増大するにつれて出力擬態語が「とことこ」へと変化していくが，変化するタイミングはサンプルによって異なることがわかる．これは $\alpha = 0$ で「とろとろ」とされたサンプルの中でも，元から「とことこ」に近い音韻ベクトルを持っていたサンプルほど早い段階で「とことこ」に変化することを示しており，式 3.5 の $\mathcal{L}_d$（音韻ベクトルの距離）と $\mathcal{L}_c$（自然さに関する罰則項）の双方が正しく機能していることが窺える．図 3.9(c) は，比較的 α が小さいうちから出力が変化するサンプルを中心に取り上げている．これを見ると，$\alpha = 1$ 前後で「つたつた」,「ぬらぬら」といった語が「すたすた」,「ふらふら」といった一般的な語に変化すること，$\alpha = 2$ から 3 にかけて「のろのろ」が「ほろほろ」,「ほらほら」等に変化することがある．オノマトペ辞典 [10] や表 2.4 で示したアノテーション実験での回答結果からもわかるように，「のろのろ」は一般的な語であり，「ほろほろ」,「ほらほら」は一般的な語ではないと考えられる．提案手法では，元から一般的な語である「のろのろ」が「ほろほろ」,「ほらほら」に変化するという，自然さという観点では望ましくない変化が一部含まれることが確認され，この点の改善は今後の課題である．改善方法としては，最適化（式 3.5）の際に，2 音素の組み合わせによる罰則項 $\mathcal{L}_c$ のほかに，一般的な語は出力されやすくする（損失を軽減する）ような補正項を追加することなどが考えられる．どのような語を一般的とするかに関しては，表 2.4 のような統計でよく出現する擬態語や，辞典 [10] に掲載されている擬態語を一般的な語と定義するなどの方法が考えられる．このような望ましくない変化は $\alpha = 2$ を超えたあた



(a) 一般的な語が変化しにくいことを示す例



(b) $\alpha = 0$ で「とろとろ」と記述されるサンプルの推移の例



(c) 比較的 α が小さいうちから出力が変化するサンプルに着目した例

図 3.9　α を徐々に変化させた時の出力擬態語の変化の例

りで起こることが多いことから，この問題の影響を低減するためには，$\alpha = 1$ 前後の，比較的小さい値を採用するのが安全であると考えられる．

3.5　まとめ

本章では，第 1 章で述べたような背景に基づき，人間の直感に近い歩容のモデル化を実現するために音韻ベクトルを利用し，擬態語により歩容を記述する手法を提案した．提案手法では，歩容モーションを入力として推定音韻ベクトルを求めるモジュールと，推定音韻ベクトルを擬態語に変換するモジュールの 2 段階に分けて歩容を擬態語で記述した．そして，評価実験により，提案手法は歩容の記述の正確さを損なわずに，より自然な擬態語を出力できることを確認した．また，提案手法は無作為に音素を選んで生成した擬態語よりも正確な擬態語を出力できていることを確認した．更に，提案手法に新奇的な語を含む多様な擬態語を出力する能力があること，自然さの考慮度合いを制御するハイパパラメータ α を調節することで，用途に応じて擬態語の新奇性を調節できることを確認した．これらにより，擬態語の音象徴性を利用した，人間の直感に近い歩容の記述が実現できたと考えられる．

第 4 章

擬態語による歩容の生成

第 1 章で述べたように，本論文では擬態語による歩容の記述および，擬態語からの歩容の生成の二つの手法を提案するが，本章ではこのうち，擬態語から歩容を生成する手法について述べる．以降，まず 4.1 節で，本研究の問題設定と，提案手法のアプローチについて述べる．続いて 4.2 節で，提案手法のアナロジーとなったスタイル変換の概念について説明する．そして 4.3 節で，提案手法について詳述する．更に 4.4 節で，提案手法の有効性を確認するための評価実験について報告し，4.5 節で考察する．最後に 4.6 節で，本章をまとめる．

4.1　はじめに

本章でも第 3 章と同様に，ABAB 型の形式の擬態語を取り扱う．学習に用いるデータも第 3 章と同一のものであり，一つの歩容に対して複数の擬態語がアノテーションされたものである．提案手法では，入力（クエリ）として単一の擬態語を想定し，(1) 擬態語クエリを音韻ベクトルに変換するモジュールと，(2) 音韻ベクトルに応じて適当な歩容を生成するモジュールの 2 段階に分けて歩容を生成する．モジュール (1) は，第 3 章で提案した，音韻ベクトルを擬態語に変換するモジュールの逆処理にあたるが，音素の組み合わせという離散的な表現である擬態語から，連続値をもつ密な特徴表現である音韻ベクトルへと変換する必要がある．その際に，不足している情報を補うために，強調音韻ベクトルという概念を新たに導入する．モジュール (2) は，生成問題をスタイル変換問題 [87] として解釈し，スタイル変換モデルを学習する．そして，スタイルの指定に音韻ベクトルを用いることで，音韻ベクトルと歩容との関係性を獲得する．なお，二つのモジュールが分

離していることから，クエリとして音韻ベクトルを直接指定し，モジュール（2）のみを用いることでも歩容の生成は可能である（この用法は 4.4.3 項の精度評価で用いる）．

4.2　スタイル変換

スタイル変換とは，Gatys らが提唱した画像変換のタスク [87] である．このタスクは，画像がスタイルとコンテンツの独立した二つの要素に分離できると考える．そして，元画像，スタイル画像という 2 枚の画像を入力とし，元画像のコンテンツを維持したまま，スタイルのみをスタイル画像と同一のものに変更するというタスクである．Gatys らの手法においては，スタイルは画像のテクスチャ，すなわち空間不変な要素であると定義されている．そこで，スタイル画像から空間不変な特徴量を算出し，これがスタイルの記述子であると考える．そして，元画像を少しずつ改変し，改変した画像から同特徴量を抽出する作業を何度も繰り返すことによって，改変画像の特徴量をスタイル画像から算出した特徴量に近づけていく．この際，空間不変でない要素（コンテンツ）ができるだけ変化しないようにしながら改変を繰り返すことにより，元画像のコンテンツと，スタイル画像と同様のスタイルをもつ新たな画像が生成される．Gatys らの手法は，あくまで画像 2 枚を入力として与えて変換画像を得るための手法であり，スタイル変換を行なうモデルを獲得するわけではないため，変換を行なうために都度最適化処理が必要であるという問題点がある．

スタイル変換を行なうモデルを学習する手法の一つとして，Adaptive Instance Normalization（AdaIN）[88] を用いる手法が提案されている．このモデルは Encoder-Decoder モデルとなっており，Encoder を通して得られた画像の中間表現に対し，その画素値の平均と分散をスタイル画像のそれに単純に置き換えたものを Decoder に通すことで変換画像を得るモデルである．すなわち，画像の中間表現の画素値の平均と分散を空間不変な特徴量として利用している．

以上の手法では，変換先のスタイルを，クエリとして具体的なスタイル画像を与えることで指定している．すなわち，変換画像を得るためには，スタイルを指定するための画像を与える必要がある．これに対し，Karras らは，彼らが提案している StyleGAN という画像生成手法 [89] において，AdaIN の前段に Latent transform というサブネットワークを導入することで，入力として画像を与えることなく，AdaIN が置き換えに用いるベ

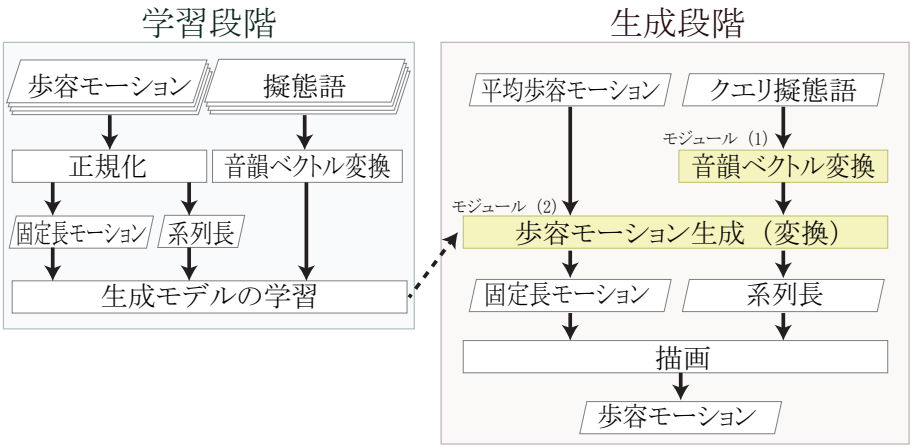


図 4.1　擬態語からの歩容生成手法の処理手順

きスタイルの平均と分散の値を自動的に算出することができることを示している．すなわち，Latent transform を用いることで，画像ではない（ドメインが異なる任意の）スタイルクエリを入力として，AdaIN が利用可能となる．

以上をふまえ，提案手法では，スタイル変換のアナロジーとして，歩容の印象は時間不変であると仮定し，これが歩容のスタイルであると考える．人間の歩行動作は周期的な行動であることから，「つまずく」，「話しかけられる」などといった突発的な別のイベントが発生しない限りは，歩容の印象が急に変化することは考えづらく，短時間であれば歩容の印象は時間不変であるという仮定が成立すると考えられる．また，スタイル変換問題として定式化する上では，スタイルのほかに，コンテンツに相当する概念を定義する必要があるが，歩容は個人認証に利用できるほど個人差が大きい [90] ことから，この個人差が歩容の印象と独立であると仮定し，歩容の個人差を歩容のコンテンツとみなす．そして，AdaIN と Latent transform を併用することで，音韻ベクトルによるスタイルの指定を可能にする．

4.3　擬態語からの歩容の生成手法

本節では，音象徴性に基づく歩容生成手法を提案する．提案手法の処理手順を図 4.1 に示す．学習段階では，まず 4.3.1 項で述べるようにして擬態語から音韻ベクトルを算出する．次に，4.3.2 項で述べるようにして歩容を生成するモデルを学習する．生成段階では，4.3.3 項で述べるようにして学習されたモデルから歩容モーションを生成する．なお，各

処理の詳細な実装に関しては 4.4.1 項で述べる．

4.3.1　擬態語の音韻ベクトル化

学習に用いる音韻ベクトル $\mathbf{p}$ は，3.2.1 項で定義した音韻ベクトルと同じである．これは複数の擬態語から算出された音韻ベクトルであるため，本章では以降これを混合音韻ベクトルと呼ぶ．混合音韻ベクトルは，対応する歩容を表現する密な特徴表現であり，モデルを学習する上では好ましい表現である一方で，混合音韻ベクトルを算出するにはクエリとして複数の擬態語を与える必要があるため，学習済みのモデルを用いて生成を行なう際には，クエリを用意するのが面倒であるという問題点がある．実用上はクエリに単一の擬態語を与えて歩容を生成したいという要望が多いと考えられるが，混合音韻ベクトルの考え方で単一の擬態語 M' を音韻ベクトルに強引に変換すると，その構成要素 C'_j に対応する部分音韻ベクトル $\mathbf{p}_j = \mathbf{q}(C'_j)$（式 3.1 において $n = 1$ の場合）となり，$\mathbf{p}$ は第 1 子音・第 1 母音・第 2 子音・第 2 母音の変数に一つずつ 1 が入った 4-hot ベクトルとなり，学習時にはこのような疎な音韻ベクトルをもった学習サンプルが存在しないことから，このままではうまく生成ができない．そこで，単一の擬態語 M' をクエリとして使用できるようにするために，単一の擬態語を密な音韻ベクトルの形に変換することを考える．そのために，学習データセットに含まれる，アノテーションされた全ての擬態語から集計された音素出現頻度を用いる．具体的には，学習データセット全体から部分音韻ベクトルの平均 $\bar{\mathbf{p}}_j$ と標準偏差 $\boldsymbol{\sigma}_j$ を求めておき，クエリ語を構成する音素以外の変数は平均の値をそのまま採用し，構成する音素の変数は平均から定数 a と標準偏差 $\boldsymbol{\sigma}_j$ の分だけ強調することで，単一の擬態語を密な音韻ベクトルに変換する．すなわち，クエリとなる単一の擬態語 M' の構成要素 C'_j に対応する部分音韻ベクトル $\mathbf{p}'_j$ は，

$$\mathbf{p}'_j = \bar{\mathbf{p}}_j + a\boldsymbol{\sigma}_j \circ \mathbf{q}(C'_j) \tag{4.1}$$

とする．ここで $\circ$ は要素積である．以下ではこのようにして得られた音韻ベクトル $\mathbf{p}'$ を強調音韻ベクトルと呼ぶ．例えば，「のろのろ」というクエリが与えられた場合の強調音韻ベクトルは図 4.2 のようになる．ここで，グラフの水色の部分が平均の値，赤色の部分が標準偏差で強調された分を表す．

実際に用いた HOYO データセットの各歩容に対応する音韻ベクトルの平均と，その各要素の標準偏差の値は，表 2.6 に示したものである．なお，強調音韻ベクトルの計算に用

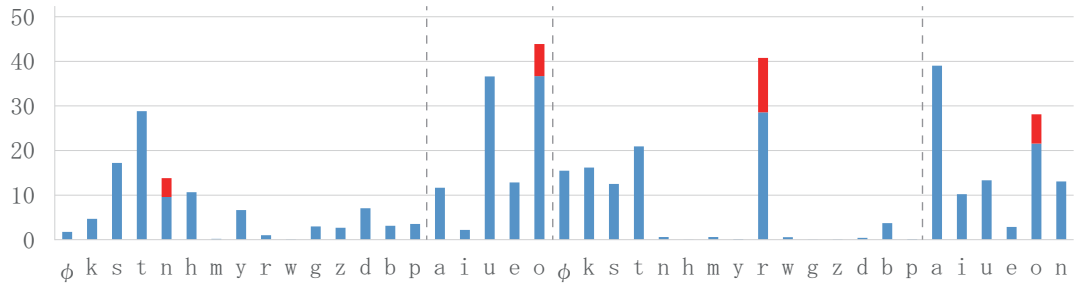


図 4.2　「のろのろ」の強調音韻ベクトル

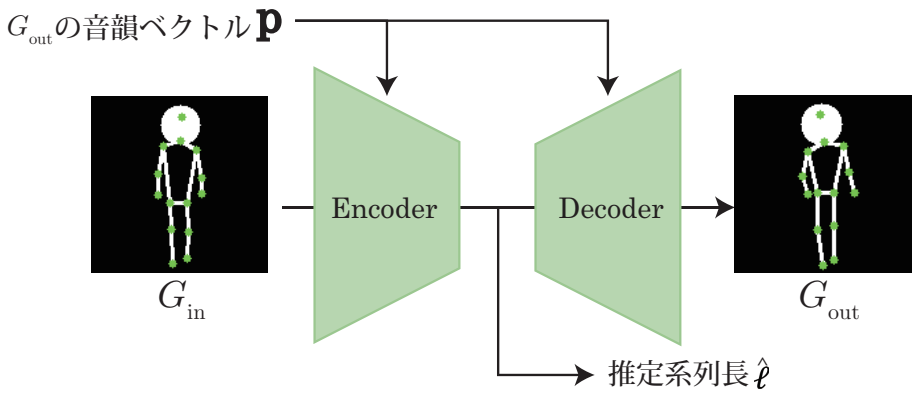


図 4.3　歩容を生成する提案モデルの概略図

いる定数 a は経験的に 0.5 とした．

4.3.2　歩容を生成するモデルの学習

本研究で提案する，歩容を生成するモデルの概略図を図 4.3 に示す．前述の通り，本手法では歩容の生成をスタイル変換 [87] のタスクとみなしてモデルの学習を行なう．本研究においては，歩容の印象は歩容 1 周期の間では時間不変であると仮定する．これにより，スタイル変換のアナロジーとして，歩容の印象がスタイルに相当すると解釈できる．

提案手法では歩容を生成するモデルとして Encoder-Decoder モデルを採用する．モデル上で扱う歩容 G_{in}，G_{out} は，歩容モーションのサンプルであり，$D \times \tau$ の行列として表現されている．ここで，τ は系列長である．D はチャネル数であり，これは歩容モーションの部位数 × 座標系の次元数である．出力歩容 G_{out} のフレーム長 $\hat{\ell}$ は Encoder 出力部から分岐した別のユニットから推定される構造になっている．ここで，フレーム長 $\hat{\ell}$ は生成歩容の速さの逆数を表している．スタイルに相当する音韻ベクトル $\mathbf{p}$ は AdaIN [88] に

よって入力する．ここで用いる音韻ベクトルは，3.2.1 項で述べた混合音韻ベクトルである．4.2 節で述べた通り，歩容は個人認証に利用できるほど個人差が大きい動作として知られていることから，提案手法では，この個人差がスタイル変換におけるコンテンツに相当すると解釈し，Encoder の入力 G_{in} には教師 G_{out} と同一の歩行者 ID をもつサンプルを学習データセット内から無作為に選択したものを用いる．

4.3.3　歩容の生成と事後処理

生成時には，Encoder の入力 G_{in} として全学習サンプルの平均を用いる．これは，歩容生成タスクにおいては入力歩容（個人差）には興味がないためである．入力音韻ベクトルとしては，4.3.1 項で述べた強調音韻ベクトル $\mathbf{p}'$ を用いる．

ただし，後述の評価実験のうち，4.4.3 項の精度評価においては，Encoder の入力 G_{in} に特定の人物の歩容を，入力音韻ベクトルとして 3.2.1 項で述べた混合音韻ベクトル $\mathbf{p}$ を用いる．これは，学習した提案モデルを，モデル本来の用途，すなわちスタイル変換タスクに利用する場合の入出力であり，4.4.3 項の精度評価においては，未知歩容モーションの推定精度を評価するためにこのような入出力を用いる．

最後に事後処理として，出力された固定長の生成歩容 $\hat{G}_{\mathrm{out}}$ と，推定フレーム長 $\hat{\ell}$ を用いて生成歩容 $\hat{G}_{\mathrm{out}}$ の系列長が $\hat{\ell}$ になるよう補間・再サンプリングすることで最終的な出力歩容を得る．

4.4　評価実験

本節では，4.3 節で提案したモデルの評価実験について報告する．まず 4.4.1 項で，評価実験を行なった際の提案手法の実装の詳細について述べる．次に 4.4.2 項で，提案手法で生成した歩容の妥当性を評価する主観評価実験について報告する．続いて 4.4.3 項で，提案手法における音韻空間の定義方法について評価する実験について報告する．

4.4.1　実装

4.4.1.1　サンプルの作成

本実験では，モデルに入力する学習サンプルとして，データセット中の歩容モーションを学習しやすいサンプル単位に分割・変換したものを用いた．HOYO データセット中の

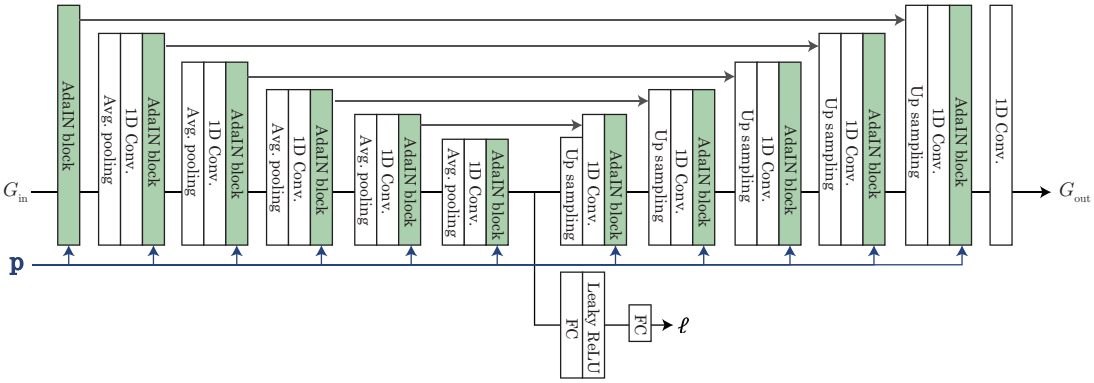
歩容モーションは系列長が一定ではないが，これらには右足および左足が接地した瞬間の時刻がアノテーションされている．本実験ではこれを利用して，元の歩容データから歩容 1 周期（2 歩）分だけを切り出した．この際，右足が接地する瞬間を時刻 0 とし，サンプル間で位相が一致するようにした．さらに，切り出した 1 周期分の歩容を平滑化スプライン補間で 128 フレームの固定長（$\tau = 128$）に変換し，元のフレーム長 l は別に保持した．このようにして得られた固定長の 1 周期分の歩容 G と元のフレーム長 ℓ の組をサンプルとし，学習に利用した．元の歩容データの長さが 2 周期以上ある場合，一つの歩容データから複数のサンプルが得られることになるが，この場合はどちらのサンプルにも同じ音韻ベクトルを付与した．さらに，データ拡張のために，サンプルの開始時刻を ±1 フレームずらして切り出したものも学習に利用した．なお，歩容モーションの部位数は 14，座標系の次元数は 2 であるため，チャネル数は $D = 28$ である．実際に学習に使用したサンプル数は 1,472 個であった．

4.4.1.2　モデルアーキテクチャ

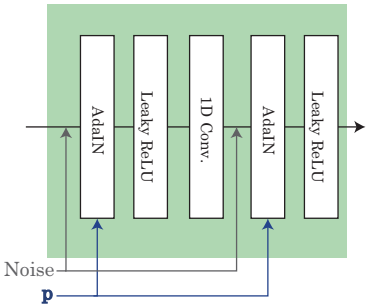
本実験では，Encoder-Decoder モデルの具体的な構造として，1 次元の畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）を採用した．これは時間方向に畳み込みを行なう CNN である．Encoder の入力層においては，部位座標系列それぞれをチャネルとみなして，ユニット数 128，チャネル数 28 の入力を受け付ける．Encoder 出力部から分岐してフレーム長を推定するネットワークが存在しており，これは中間層を一つもつ全結合ネットワークである．また，モデルには Skip-connection が存在している．詳細なモデルの構造を図 4.4 に示す．ここで，図中の $\mathbf{p}$ は音韻ベクトルである．

4.3.2 項で述べた通り，学習時の Encoder の入力 G_{in} には，教師として用いる学習サンプル G_{out} と同一人物の学習サンプルを無作為に選択して用いる．本実験では，学習エポック開始時に歩行者一人あたり 25 サンプルを無作為に選択し，Encoder の入力に用いた．すなわち一つの G_{out} に対し，Encoder の入力値を変えながら 25 回学習を行なった．場合によっては $G_{\mathrm{in}} = G_{\mathrm{out}}$ となる可能性もある．この選択はエポックごとに変更する．また，データセット中の歩行者は 10 人であるため，1 エポックあたりの学習回数を $\sum_{k=0}^{9} 25 \times N_k$ とし，具体的には 36,800 回であった．ここで N_k は歩行者 ID が k である学習サンプルの数である．

これらの深層学習モデルの実装には Keras [80] を用い，パラメータは実験的に設定



(a) モデル全体



(b) AdaIN [88] block の詳細

図 4.4　実験で使用した提案手法のモデルアーキテクチャ

した．

4.4.1.3　座標の正規化

本実験では，事前処理としてサンプルの座標値の正規化を行なった．まず，フレーム単位で全部位の座標平均を算出し，これが原点となるように平行移動させた．すなわち，各部位の座標値が人体の中心座標からの相対位置を表すようにした．

また，2.1 節で述べた通り，学習データには異なる 10 人の人物による歩容モーションが収録されているが，人物によって身長が異なるため，これを揃えずに学習すると，中心点から離れた頭や足などの座標の学習が不安定になる問題があった．そこで，学習データ（サンプルを切り出す前の系列）単位で，y 座標の最大値（頭の y 座標の値）と最小値（地についている方の足の y 座標の値）を算出し，これの差を用いて系列の座標値を拡縮した．すなわち，各サンプルにおいて歩行者の身長がほぼ 1 となるようにスケールを調整した．なお，縦横比が保たれるように x 軸方向も同じスケールで拡縮した．

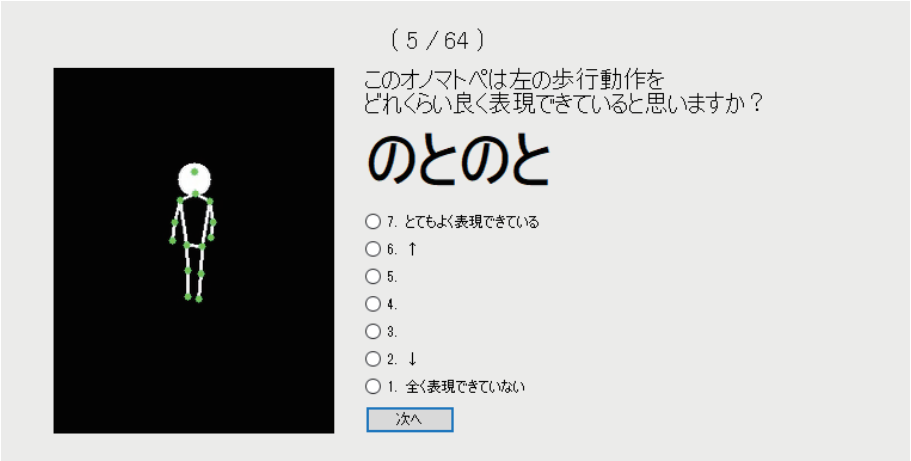


図 4.5　主観評価実験に用いたインタフェース

4.4.1.4　事後処理

生成した歩容の動きを滑らかにするため，出力された固定長の生成歩容 $\hat{G}_{\mathrm{out}}$ に対して，5 フレーム窓の平滑化処理を施した．また，推定フレーム長 $\hat{\ell}$ を用いて生成歩容 $\hat{G}_{\mathrm{out}}$ の系列長を $\hat{\ell}$ に伸縮する処理には平滑化スプラインを利用した．最後に，伸縮した推定歩容は，座標正規化の都合で，そのままでは宙に浮いているように見えてしまうため，これを接地させる処理を行なった．具体的には，フレームごとに右足の座標値と左足の座標値を比較し，下側にある方を基準として全部位の y 座標を補正した．

4.4.2　生成した歩容の主観評価

提案手法の有効性を検証するために，クエリとなる擬態語 M に対応した歩容モーションを適切に生成できたか調べる主観評価実験を行なった．本研究では，一般的な擬態語を扱う場合と，新奇的な擬態語を扱う場合の二つの条件で実験を行なった．

本実験では，生成された歩容モーションと擬態語のペアを評価者に提示し，このオノマトペが歩行動作をどのくらい良く表現できていると思うか質問した．回答方法は 7 段階の Likert 尺度を用いた．主観評価実験で用いたインタフェースを図 4.5 に示す．歩容モーションは 60 fps の動画像で提示した．評価に用いた生成歩容モーションの例を図 4.6 に示す．この図では，左から右に向かって 1 周期分の歩容を時系列順に並べてある．提示した擬態語 M_{show}（以下，これを提示語と呼ぶ）は 27 種類であり，この提示語をクエ

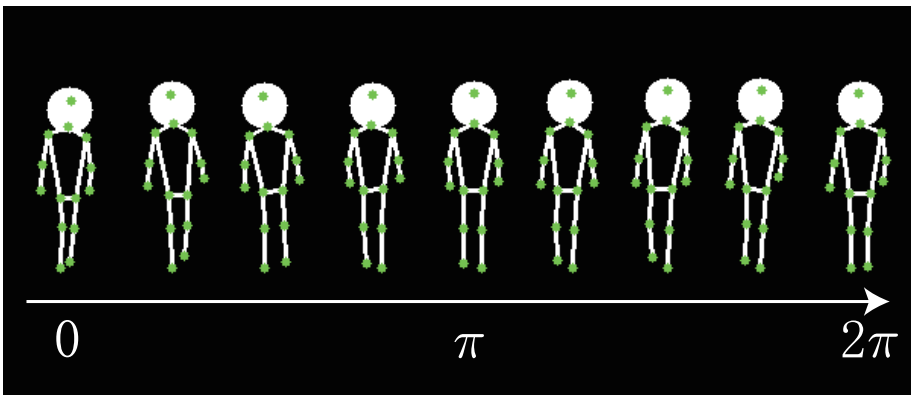


図 4.6　生成した歩容モーションの例

リとして生成した歩容モーションと提示語のペア（$M_{\mathrm{show}} = M'$）を正例ペア，提示語とは異なる無作為に選んだ擬態語をクエリとして生成した歩容モーションと提示語のペア（$M_{\mathrm{show}} \neq M'$）を負例ペアと定義する．正例ペアと負例ペアを混ぜ，この全 54 問を無作為な順番で提示することで評価した．ここで負例ペアはベースラインであり，正例ペアが負例ペアと比較して高評価であるほど提案手法による生成が優れていると言える．

ここで，提案手法が一般的な語を扱った場合の性能と，新奇的な語を扱った場合の性能を評価するために，提示語 M_{show} として，一般的な擬態語を用いる場合と，新奇的な擬態語を用いる場合の二つの条件で実験を行なった．前者の条件では，提示語 M_{show} として，HOYO データセットのアノテーションに含まれる擬態語のうち，3 本以上の歩容に対してアノテーションされているものを用いた．一方，後者の条件では，提示語 M_{show} として，無作為に四つの音素を選択して生成した ABAB 型の擬態語 27 語を用いた．表 4.1 に使用した擬態語の一覧を示す．評価者はどちらの場合も日本語を母語とする大学生 22 名であった．ただし，両実験間では評価者が一部異なる．

まず，一般的な擬態語を扱った評価の結果，正例ペアの評価値平均は 4.396，負例ペアの評価値平均は 4.047 となり，正例ペアの評価値の方が高くなった．なお，正例ペアにおける評価者間の各設問に対する評価値の相関係数は 0.32 となっており，回答の傾向には弱めの相関しか見られなかった．強い相関がないのは，回答者によって擬態語や歩容の感じ方にある程度の個人差があるためだと考えられる．また，正例ペアの評価値と負例ペアの評価値の有意差を検定した．Carifio らの議論 [91] を参考とし，評価者ごとに回答の値を平均したものを標本とした．両分布に正規性が認められなかったため，Wilcoxon の符号順位検定 [92] を適用した．各群の標本数は 22，正例ペア評価値の中央値は 4.389，負例

表 4.1　主観評価実験で用いた擬態語

一般的な擬態語	よたよた，どしどし，つらつら，たかたか，そろそろ，よろよろ，ぐわぐわ，ゆらゆら，てとてと，ぐらぐら，だらだら，のしのし，ずかずか，どかどか，すいすい，すらすら，くらくら，のそのそ，とろとろ，のろのろ，とぼとぼ，ぶらぶら，たらたら，とことこ，ふらふら，すたすた，てくてく
新奇的な擬態語	ぎそぎそ，せごせご，もよもよ，へうへう，でぃがでぃが，ぱひぱひ，ねふねふ，おつおつ，らぷらぷ，ちぽちぽ，にけにけ，ぺうぺう，さうさう，ぴるぴる，よすよす，でをでを，によによ，やへやへ，ぎかぎか，わふわふ，われわれ，やすやす，るくるく，きうきう，がじがじ，どゆどゆ，じぎじぎ

ペア評価値の中央値は 4.019 であり，検定の結果 $p < 0.002$ で有意差が認められた．ここから，提案手法によって生成した歩容モーションはクエリとなる擬態語をより適切に反映したものとなっていることがわかる．すなわち，提案手法が擬態語をクエリとして，比較的適切な歩容を生成できることを確認できた．

次に，新奇的な擬態語を扱った評価の結果，正例ペアの評価値平均は 3.273，負例ペアの評価値平均は 3.202 となり，正例ペアの評価値の方が高くなった．ただし，新奇的な擬態語を用いた場合，一般的な擬態語を用いた場合よりも正例ペア，負例ペア間の差が小さくなっており，有意差は確認できなかった．これは一般的な擬態語からの歩容の生成よりも，新奇的な擬態語からの歩容の生成の方が難しいタスクであるためだと考えられる．さらなる手法の改良や，学習データセットの拡充により，新奇的な擬態語を用いた場合でも十分な生成能力を発揮できるようにすることは今後の課題である．

4.4.3　音韻空間の評価

3.3.3 項と同様に，本研究で用いている音韻ベクトルの定義が妥当であることを確認するために，異なる定義による音韻ベクトルを用いて評価実験を行なった．使用した比較手法は 3.3.3 項と同一であり，比較手法 A は擬態語を文字単位で分解したもの，比較手法 B は擬態語を音素単位で分解したのち，秋山ら [33] が提案する音素定量化手法に則って数

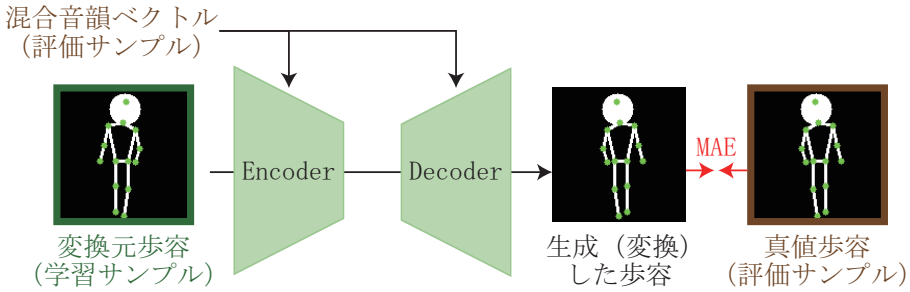


図 4.7　精度評価の概念図

値化したものである．この評価のために，4.4.2 項の実験に準ずる主観評価による比較と，未知の評価用データを用いた精度評価の二つの実験を行なった．

主観評価実験の基本的な手順は 4.4.2 項と同様である．ただし，主観評価実験の実施コスト上，実験条件としては，一般的な擬態語を用いる場合の正例ペアに限って評価を行なった．すなわち，生成された歩容モーションと提示語の正例ペアを評価者に提示し，4.4.2 項の実験と同じ図 4.5 のインタフェースを用いて，提示したオノマトペが歩行動作をどのくらい良く表現できていると思うかを 7 段階の Likert 尺度で評価させた．提示語には表 4.1 に示した擬態語のうち，一般的な擬態語 27 語を用い，全 27 問を無作為な順番で提示した．評価者は日本語を母語とする大学生 22 名であった．ただし，4.4.2 項の実験とは評価者が一部異なる．

また，主観評価実験に加えて，精度評価，すなわち，学習した提案モデルを，未知の評価用データを用いて，スタイル変換モデル本来の用途で用いた場合の性能を評価する実験を行なった．この精度評価方法の概念図を図 4.7 に示す．ここでは，HOYO データセットに含まれる歩容のうち，学習時には用いなかったものを評価用データとして用いた．具体的には，右足が前に出ている状態から開始する歩容モーションは全て学習に利用しているため，左足が前に出ている状態から開始する歩容モーションを左右反転させることで，疑似的に右足が前に出ている状態から開始する未知歩容モーションを作成した．

4.3.3 項で述べた通り，本実験に限っては，学習時と同様に，評価用歩容モーションにアノテーションされた擬態語群を混合音韻ベクトルに変換し，これを入力として歩容を生成した．また，生成は歩行者（データセット上の人物 ID）ごとに行ない，生成時の Encoder 部の入力 G_{in} には，その歩行者の学習用データを無作為に選択して用いた．そして，生成した歩容モーションと，評価用歩容モーションの真値との誤差を計測した．

表 4.2　音韻空間の比較評価

音韻ベクトル実装	主観評価値	座標推定誤差	系列長推定誤差
提案手法（音素単位）	4.396 ± 0.023	0.539（0.916 m 相当）	11.27 フレーム
比較手法 A（文字単位）	4.337 ± 0.037	1.040（1.767 m 相当）	22.88 フレーム
比較手法 B（秋山ら [33] による数値化）	3.428 ± 0.048	0.479（0.814 m 相当）	11.19 フレーム

評価指標には 1 フレームあたりの座標値の平均絶対誤差（Mean Absolute Error; MAE）を用いた．なお，評価サンプル数は 156 であった．

両実験の評価結果を表 4.2 に示す．ただし，提案手法の主観評価の数値は 4.4.2 項で報告した結果の再掲である．座標推定誤差の括弧内の数値は，歩行者の身長を 1.7 m とした場合の換算値（L1 距離）である．比較手法 A は精度評価（推定誤差）で，比較手法 B は主観評価値で，それぞれ提案手法よりも評価が低下しており，提案手法は主観評価，精度評価の双方を，比較的高い水準で両立していると言える．なお，入力された混合音韻ベクトルと，最も L1 距離が近い音韻ベクトルをもつ学習セット内の歩容モーションを出力する最近傍モデルを用いて精度評価を実施すると，座標推定誤差は 0.577，系列長推定誤差は 19.82 フレームとなり，提案手法と比較手法 B はこの水準を上回っている．また，比較手法 A の座標推定誤差および系列長推定誤差が，提案手法や比較手法 B と比較して倍程度に大きくなっているのは，3.3.3 項の実験結果と同様，提案手法の音韻ベクトルが 41 次元，比較手法 B の音韻ベクトルが 16 次元であるのに対して，比較手法 A の音韻ベクトルが 151 次元と大きいことから，次元の呪いの影響を受けている可能性が考えられる．

なお，座標推定誤差は 14 部位の誤差の合計値であり，提案手法の座標推定誤差は 1 部位あたりでは 6.5 cm の誤差に相当する．また，学習データの平均系列長は約 79 フレームであり，提案手法の系列長推定誤差 11.27 フレームは約 14 % の誤差であった．

4.5　考察

本節では，評価実験の結果について考察する．

4.5.1　提示語の一般度と評価値の関連性について

主観評価実験においては，一般的な提示語のグループと，新奇的な提示語のグループを用いて評価を行なったが，新奇的な提示語のグループの方が正例ペア，負例ペアともに評

表 4.3　より一般度が高い提示語のみを集計した主観評価結果

一般度	正例ペア	負例ペア
3 以上	4.396	4.047
4 以上	4.466	4.045
5 以上	4.555	4.098
7 以上	4.565	3.964
10 以上	4.705	4.080

価値が低くなった．この結果は，一般的でない語が提示語であった場合に，評価値に対して負のバイアスがかかることを示唆している．

4.4.2 項の主観評価実験のうち，一般的な語を提示語として用いた場合の実験では，提示語として HOYO データセットのアノテーションに含まれる擬態語のうち，3 本以上の歩容に対してアノテーションされているものを用いた．この提示語 27 語の中でも，アノテーションされている歩容の本数が最も多かった語は「てくてく」であり，40 本の歩容にアノテーションされている．このように，提示語の中でもアノテーションされた歩容の数（以下，これを語の一般度と呼ぶ）には多寡があるため，これと主観評価実験における評価値との関係性を調査した．具体的には，主観評価実験結果のうち，より一般度が高い語に限って集計をした場合の結果の変化を調べた．その結果を表 4.3 に示す．ここから，一般度が高い語に限るほど正例ペアの評価値が高くなる傾向が見てとれる．しかし，負例ペアの結果を見ると，一般度が異なってもほぼ変化がないことから，擬態語自体が評価値に影響を与えている（当たり障りのない擬態語を出力すれば安定して高評価が見込める）わけではないこともわかる．すなわち，この結果は 4.4.2 項の新奇的な語の実験結果と同様に，一般度が低い語をクエリにした場合ほど生成がうまくいきにくいことを示している．これは，一般度が高い語ほど，その音韻ベクトルに近い音韻ベクトルをもつ学習サンプルが多いためだと考えられる．これは逆に言えば，より大規模なデータセットを用いて学習することで提案手法の性能が向上する余地があること，特に，追加のデータを収集する際には，一般度が低い語に対応するような歩容を重点的に収集するのが効果的であることを示唆している．

表 4.4　歩容の系列長が主観評価に与える影響について

	動きの推定結果を利用	速さの推定結果を利用	主観評価値
正例ペア	✓	✓	4.389
(a)	✓	–	4.130
(b)	–	✓	4.014
負例ペア	–	–	4.019

4.5.2　歩容の系列長が主観評価に与える影響について

提案手法では，固定長の歩容モーションと速さ（系列長）の二つに分けて歩容を生成している．この二つの要素が評価結果に及ぼす影響を調べるために，追加で主観評価実験を実施し，4.4.2 項の実験結果と比較した．具体的には，4.4.2 項の主観評価実験と同じ手順で，

(a) 正例ペアにおいて，生成した歩容モーションの系列長を 79 フレームに固定したもの

(b) 負例ペアにおいて，生成した歩容モーションの系列長を正例ペアの系列長で置き換えたもの

を用意して評価した．なお，79 フレームは学習データの平均フレーム長である．正例ペアは動きと系列長の両方の推定結果を生成に利用するもの，負例ペアはどちらの推定結果も生成に利用しないものとみなすことができるが，(a) の条件は，動きの推定結果のみを生成に利用し，系列長の推定結果を利用しないもの，(b) の条件は，系列長の推定結果のみを生成に利用し，動きの推定結果を利用しないものであると言える．

本実験の評価者は日本語を母語とする大学生 8 名であった．比較条件と評価の結果をまとめたものを表 4.4 に示す．ただし，正例ペア，負例ペアの主観評価値は 4.4.2 項で実施した実験結果の再掲であり，本追加実験とは被験者数が異なる．被験者数が異なるため厳密な比較はできないが，正例ペアが (a) の条件よりも高く評価されていることから，速さの推定結果を利用することは有効であると考えられる．また，(b) の条件が負例ペアとほぼ同水準であるということは，速さが正しくても動きの情報を正しく生成できなければ生成結果は正しく見えないことを示唆している．

4.5.3　真値データを用いた主観評価について

4.4.2 項において提案手法の評価値平均が 4.389 であることを報告した．この値の良し悪しを考察するための基準を得るために，データセットに含まれる歩容モーション（真値）を用いて，4.4.2 項と同様の主観評価実験を行なった．

4.4.2 項の主観評価実験では歩容モーションと単一の擬態語のペアを用いたが，HOYO データセットの歩容には複数の擬態語がアノテーションされているため，そのままでは実験に利用できない．そこで本実験では，HOYO データセットに含まれる歩容のうち，自由記述アノテーションにおいてアノテータ 15 人中 13 人以上が同一の擬態語を回答したものとその擬態語のペアを実験に用いた．具体的には，「ふらふら」，「すたすた」の 2 種類の歩容を各 2 サンプルずつ，計 4 サンプルを用いて評価を行なった．また，4 サンプルだけでは数が少ないと思われることから，閾値を下げて，アノテータ 15 人中 6 人以上が同一の擬態語を回答したものを実験に用いた場合の結果も併記する．こちらは「ふらふら」，「すたすた」に加えて「とぼとぼ」，「てくてく」，「のろのろ」の 5 種類を各 2 歩容ずつ，計 10 サンプルを用いた．本実験の評価者は日本語を母語とする大学生 8 名であった．

評価の結果，4 サンプルによる評価値平均は 5.281，10 サンプルによる評価値平均は 5.088 となった．この値は，提案手法が目指せる評価値の上限とみなすことができる．本研究の主観評価実験では 7 段階の Likert 尺度を利用したが，一連の主観評価実験で得られた評価値平均が概ね 4 から 5 の間に集まっていることをふまえると，より細かい尺度を用いて評価を行なう方が望ましい可能性があり，この点は今後の課題である．

なお，筆者の先行研究 [86] において，本実験と同様の実験が実映像と擬態語のペアに対して実施されており，そこでは真値を用いた主観評価結果が，7 段階の Likert 尺度で平均 6.0 という結果が得られている．これと比較して，本実験における真値の評価結果が低いことは，棒人間により表示される歩容モーションと実映像の間にはギャップが存在することを示唆している．目的を実映像の生成などに拡張し，このギャップをふまえた評価を実施することも今後の課題である．

4.6　まとめ

本章では，第 1 章で述べたような背景に基づき，人間の直感に近い歩容のモデル化を実現するために音韻ベクトルを利用して，第 3 章で述べた歩容の記述とは逆に，擬態語から歩容モーションを生成する手法を提案した．また，提案手法では，歩容の生成をスタイル変換問題として捉え，音韻ベクトルをスタイルとして入力し，歩容を生成するモデルを構築した．そして，評価実験により，入力に一般的な擬態語を用いた場合，クエリとした擬態語と生成された歩容モーションが比較的正しく対応していることを確認した．これにより，擬態語の音象徴性を利用した，人間の直感に近い歩容の生成が実現できたと考えられる．ただし，入力に新奇的な擬態語を用いた場合の生成能力は評価実験で有意差がみられなかったことから，さらなる手法の改良や，学習データセットの拡充により，新奇的な擬態語を用いた場合でも十分な生成能力を発揮できるようにすることは今後の課題である．

第 5 章

むすび

本章では，本論文の内容を総括し，今後の展望について述べる．

5.1　総括

第 1 章で述べたように，歩行動作は人間にとって最もなじみが深い動作の一つである．人間の歩行動作の様子は歩容と呼ばれ，見慣れた動作であるがゆえに，人間は人間の歩容の細かな違いを弁別することができ，「強そう」，「軽やか」など，様々な印象を感じ取ることができる．しかし，歩容認証や行動認識などのタスクが盛んに研究される一方で，従来は歩容を記述するための適切なラベルが提案されていないことを大きな要因として，歩容そのものの直感的で精緻な記述・生成を試みた研究は少なく，限定的なのが現状である．そこで，本研究では擬態語を利用することで，このような歩容の記述・生成を試みた．擬態語には音象徴性という性質があり，擬態語を構成する音素の音響的印象が事象の様態と対応するとされている．この性質をふまえると，例えば「とことこ」歩く，「どこどこ」歩く，というように，擬態語を構成する音素の一部を入れ替えることによって，微妙な印象の違いを表現することが可能である．言い換えれば，この性質は，動作の細かい印象の基底として擬態語の音素を利用可能であることを示唆している．

本論文では，以上の仮説に基づき，擬態語を「音韻空間」上で表現したベクトルである「音韻ベクトル」と歩容との対応関係を獲得することにより，歩容の印象の細かな違いを表現し分けることが可能な，人間の直感に近いモデルを構築し，これをもって（1）擬態語により歩容を記述する手法，および（2）擬態語から歩容を生成する手法の二つを提案した．

まず第 2 章で，本研究で用いるために新たに構築した，擬態語がアノテーションされた歩容データセット「HOYO」について紹介した．本研究では歩容の細かな違いをモデルに学習させる必要があることから，歩容自体もできるだけ多様なものが収録されているデータセットを用いるのが望ましい．そこで，既存の歩容認証データセットにアノテーションを追加するのではなく，歩容の動画像も新規に撮影した．この際，歩行者に対して擬態語による動作の教示を行なうことにより，歩容そのものの多様性を豊かにした．しかし，歩行者が自身のイメージ通りに体を動かせるとは限らないため，得られた歩容は客観的に見て，教示された擬態語を表現できているとは限らない．そこで，第三者が歩容を見た際に想起する擬態語を，歩容を表現する真の擬態語と定義し，第三者による評価に基づいて，改めて歩容に対する擬態語のアノテーションを行なった．ここで，本データセットでは各歩容に 2 種類の擬態語アノテーションを付与した．一つは選択式アノテーションであり，主観ラベルと同様のカテゴリカルな擬態語ラベルを付与するものである．もう一つは自由記述アノテーションであり，アノテータに複数の擬態語を自由に回答させるものである．構築したデータセットは Web 上で公開しており，擬態語がアノテーションされた唯一無二の公開歩容データセットとして，感性工学，行動認識，異常検出などのさまざまなタスクへの応用が期待される．

続く第 3 章では，擬態語により歩容を記述する，すわわち，擬態語を入力として歩容を出力する手法を提案した．音韻ベクトルは，音象徴性に基づく表現であるため，人間の直感をよく反映した形式となっていることが期待できる．そこで，提案手法では歩容から End-to-End に擬態語を求めるのではなく，(1) 歩容を入力として推定音韻ベクトルを求めるモジュールと，(2) 推定音韻ベクトルを擬態語に変換するモジュールの 2 段階に分けて歩容を記述した．そして，評価実験により，提案手法は歩容の記述の正確さを損なわずに，より自然な擬態語を出力できることを確認した．また，提案手法は無作為に音素を選んで生成した擬態語よりも正確な擬態語を出力できていることを確認した．更に，提案手法に新奇的な語を含む多様な擬態語を出力する能力があること，自然さの考慮度合いを制御するハイパパラメータ α を調節することで，用途に応じて擬態語の新奇性を調節できることを確認した．しかし，提案手法では，α を大きくしていくと，元々一般的な語である「のろのろ」が「ほろほろ」,「ほらほら」に変化するなど，自然さの観点では望ましくない変化が一部含まれることが確認され，この点の改善は今後の課題である．改善方法としては，最適化（式 3.5）の際に，2 音素の組み合わせによる罰則項 $\mathcal{L}_c$ のほかに，一般的

な語が出力されやすくする（損失を軽減する）ような補正項を追加することなどが考えられる．

そして第 4 章では，第 3 章とは逆に，擬態語から歩容を生成する手法を提案した．提案手法では，歩容の生成をスタイル変換問題として捉え，音韻ベクトルをスタイルとして入力し，歩容を生成するモデルを構築した．ここで，本来複数の擬態語から算出される音韻ベクトルを単一の擬態語クエリから算出するために，HOYO データセットの擬態語アノテーションの統計情報を利用した強調音韻ベクトルを提案し，利用した．そして，評価実験により，入力に一般的な擬態語を用いた場合，クエリとした擬態語と生成された歩容モーションが比較的正しく対応することを確認した．一方，入力に新奇的な擬態語を用いた場合の生成能力は評価実験で有意差がみられなかったことから，さらなる手法の改良や，学習データセットの拡充により，新奇的な擬態語を用いた場合でも十分な生成能力を発揮できるようにすることが今後の課題である．

以上の成果をもって，本研究では，歩容の印象の細かな違いを表現し分けることが可能な，人間の直感に近いモデルを構築し，また，これを用いた擬態語による記述と生成のための方法論を確立し，各々について一定の水準で実現できたと言える．

5.2　今後の展望

今後本研究が発展し，より正確に擬態語による歩容の記述と生成が可能となることは，計算機が人間の感覚をより良く理解できるようになることを意味しており，ここから，例えば対話型インタフェースなどといった，より人間味がある人工知能，人間の生活に寄り添う人工知能の開発へとつながることが期待される．このほかにも，例えば，記述手法は，歩容という動きの情報を，擬態語という文字列の形にエンコードしていると解釈できることから，動画像のクリップをサムネイル画像と擬態語で表現するなどして，超高圧縮率な映像要約に応用することができる可能性がある．また，記述手法と生成手法を組み合わせることにより，例えば歩容を擬態語に変換し，それを歩容に再変換することで，歩容の印象を保持したまま歩容を匿名化する等の応用も考えられる．

いずれの応用を目指すにせよ，提案手法では，記述・生成ともに，主観評価での評価値が 7 段階の Likert 尺度の 4 点台に留まっており，さらなる性能向上が望まれる．データセットの拡充，深層学習モデルのアーキテクチャの再検討などにより，性能を向上させる

ことは今後の課題である．このような単純な性能改善のほかにも，本研究は様々な発展の方向性が考えられる．以下，本節では 3 次元骨格情報の利用，歩行者の見た目の印象の考慮，多様性がある歩容生成の 3 点について，今後の展望を述べる．

5.2.1　3 次元骨格情報の利用

本研究で利用した 2 次元骨格は 3 次元骨格よりも取得が容易であり，例えば記述タスクの応用として，監視カメラ映像に映る遠方歩行者の歩容を記述するような状況を想定すると，低解像度な映像からでも比較的取得が容易な 2 次元骨格を利用しているのは利点である．一方，3 次元骨格が得られれば，それを 2 次元骨格に変換（射影）するのは容易であるため，生成タスクにおいては 3 次元骨格を生成できる手法の方が優れていると言える．

本研究では，データセット構築時の制約から，3 次元骨格情報を取得するのが困難であったため，モデルの入出力に用いる歩容モーションには 2 次元骨格情報を用いた．しかし，提案手法の考え方は歩容モーションに 3 次元骨格情報を用いる場合にもそのまま適用可能であると考えられる．モデルの自由度が増すことから，学習に際してより多くのサンプルが必要となる可能性があるが，3 次元骨格であれば，関節間の距離など，人体の形状に関する知識を制約として用いることが可能なため，モデルの自由度の問題はある程度低減可能であると考えられる．

このように，提案手法で 3 次元骨格情報を利用するためには，擬態語がアノテーションされた 3 次元骨格情報による歩容モーションが必要であり，3 次元骨格情報を含むデータセットを確保することが最大の課題である．第 2 章でも述べた通り，本研究で用いるようなデータセットを構築する上で問題となるのは，(1) 歩容に対して擬態語をアノテーションする必要があること，(2) 動きそのものに多様性がある歩容を収集する必要性から，既存の歩容データセットの歩容を転用するのが困難であることの 2 点である．このうち前者に関しては，アノテーションを付与したい 3 次元歩容モーションを 2 次元歩容モーションに変換し，本研究で提案した記述手法を適用することで自動化することができる可能性がある．

5.2.2　歩行者の見た目の印象の考慮

本研究では擬態語と歩容モーションの関係に着目した．歩容モーションは歩容から動きの情報だけを抽出したものであり，歩行者の見た目の情報は除去されている．しかし，歩行動作を行なっている人物の見た目の印象，例えば体形や性別なども擬態語との関係に影響を与えると考えられる．記述タスクにおいては，提案手法で行なっているような動きの印象推定に加えて，見た目の印象推定を並列に行ない，両出力（音韻ベクトル）を統合することで，歩行者の見た目まで考慮した，擬態語による歩容の記述が可能になると考えられる．

生成タスクにおいては，例えばコンピュータグラフィックス（Computer Graphics; CG）のモーション生成に利用する場合，生成したモーションに付与する外見の属性を考慮したうえで，その変化を逆算してモーションを生成する必要がある．提案手法の場合，モデルの再学習等を行なわずとも，入力音韻ベクトルに対して直接補正をかけることで対応が可能であると考えられる，

5.2.3　多様性がある歩容生成

歩容は個人差が大きいという点からもわかるように，一般に音韻ベクトルから歩容への変換は 1 対多変換であると考えられる．ゆえに擬態語からの歩容生成においても，出力結果には多様性があることが望ましい．そのためには，より多様性が大きなデータセットを構築することに加えて，陽に生成結果の多様性を課す評価関数を導入するのが望ましい．これは，例えば敵対的生成ネットワーク（Generative Adversarial Network; GAN）[93, 94] の枠組みを利用し，Discriminator に歩行者の ID や属性などの推定タスクを追加するなどの方法が考えられる．

謝辞

本論文の主査であり，指導教員である，名古屋大学大学院 情報学研究科 井手一郎 教授に深く感謝いたします．学部時代に旧村瀬研究室メディアグループに配属されて以来，長い間熱心にご指導いただきました．研究の大まかな方向性に関する御助言，論文校正時の鋭いご指摘など，大小様々な場面で助けていただきました．

副査であり，本研究が始まった頃から最も近くで多くの助言をいただきました，人間環境大学 人間環境学部 平山高嗣 客員教授に深く感謝いたします．近年はご多忙の中にもかかわらず，多くの時間を割いていただき，研究方針や原稿に関する相談に乗っていただきました．

副査として，そして旧村瀬研究室時代に，多大な御指導，御鞭撻を賜りました，名古屋大学大学院 情報学研究科 村瀬洋 特任教授に深く感謝いたします．当時はご多忙の中，節目節目に的確なご助言をいただき，大いに研究の指針となりました．また，充実した研究環境を提供していただき，快適に研究活動をさせていただきました．

副査として，そして旧村瀬研究室時代に研究環境の面で大変お世話になりました，名古屋大学大学院 情報学研究科 出口大輔 准教授に深く感謝いたします．研究室再編時の環境移行の際に，しばらく以前の環境を残していただくなど，多くのご配慮をいただきました．

副査としてお世話になりました，名古屋大学大学院 情報学研究科 戸田智基 教授に深く感謝いたします．副査をお願いする以前にも，修士時代の節目の発表等で第三者の視点から鋭いご指摘を多々いただき，大いに研究の助けとなりました．

文献調査や研究環境面でお世話になりました，理化学研究所 情報統合本部 ガーディアンロボットプロジェクト 川西康友 客員准教授に深く感謝いたします．深層学習関連の環境構築をしていただき，手法の実装を滞りなく進めることができました．

データセット構築，および評価実験に際して多大な貢献をしていただきました，中京大学 工学部 道満恵介 准教授に深く感謝いたします．予定にない評価実験を急遽お願いした

際にも，快く，そして手早く実験参加者を集めていただき，大変助かりました．

コンテンツ科学研究室再編後からお世話になりました，名古屋大学数理・データ科学教育研究センター 駒水孝裕 准教授に深く感謝いたします．本研究のこれまでの経過を知らないフラットな目線からいただいた助言は，本論文をまとめるうえで助けとなりました．

種々の手続きでお世話になりました秘書の蒲文代さん，田中弘美さんに深く感謝いたします．こちらの不手際からご迷惑をおかけしたこともありましたが，いつも万事滞りなく手続きを進めていただき，研究に集中することができました．

日々の研究生活やディスカッション，ミーティングなどで多くの御意見，御助言を頂きましたコンテンツ科学研究室，および旧村瀬研究室の諸氏に深く感謝いたします．中でも，データセットの撮影，アノテーション実験および主観評価実験に御協力頂いた方々には，突然のお願いに快く応じて頂きました．深く感謝いたします．また，データセット撮影に協力頂きました中京大学 目加田・道満研究室の諸氏に深く感謝いたします．

参考文献

[1] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pp.4724–4732, 2016.

[2] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Skeleton-based action recognition with directed graph neural networks,” Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition, pp.7912–7921, 2019.

[3] U. Bhattacharya, C. Roncal, T. Mittal, R. Chandra, A. Bera, and D. Manocha, “Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping,” Proc. 16th European Conf. on Computer Vision, Part X, Lecture Notes in Computer Science, vol.12355, pp.145–163, 2020.

[4] O. Temuroglu, Y. Kawanishi, D. Deguchi, T. Hirayama, I. Ide, H. Murase, M. Iwasaki, and A. Tsukada, “Occlusion-aware skeleton trajectory representation for abnormal behavior detection,” Proc. 26th Int. Workshop on Frontiers of Computer Vision, Communication in Computer and Information Science, vol.1212, pp.108–121, 2020.

[5] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, “A richly annotated dataset for pedestrian attribute recognition,” arXiv preprint, arXiv:1603.07054, 2016.

[6] 川西康友，新村文郷，出口大輔，村瀬　洋，“サーベイ論文：画像からの歩行者属性認識，” 電子情報通信学会技術研究報告，PRMU2015-112，2015.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, pp.7291–7299, 2017.

[8] “Open pose,” https://github.com/CMU-Perceptual-Computing-Lab/

openpose/ [2021/12/15 閲覧].

[9] L. Hinton, J. Nichols, and J.J. Ohala, Sound Symbolism, Cambridge University Press, Cambridge, UK, 1994.

[10] 小野正弘，擬音語・擬態語日本語 4500 オノマトペ辞典，小学館，東京，2007.

[11] C.M. Doke, Bantu Linguistic Terminology, Longmans, Harlow, UK, 1935.

[12] M. Dingemanse, "The meaning and use of ideophones in Siwu," PhD thesis, Nijmegen and Radboud University, Nijmegen, The Netherlands, 2011.

[13] K. Akita, "A grammar of sound-symbolic words in Japanese: Theoretical approaches to iconic and lexical properties of mimetics," PhD thesis, Kobe University, Kobe, Japan, 2009.

[14] S. Hamano, The Sound-Symbolic System of Japanese, CSLI Publications, Stanford, CA, USA, 1998.

[15] 田守育啓，ローレンススコウラップ，オノマトペ—形態と意味—，くろしお出版，東京，1999.

[16] 藤野良孝，井上康生，吉川政夫，仁科エミ，山田恒夫，"運動学習のためのスポーツオノマトペデータベース，" 日本教育工学会論文誌，vol.29，no.Suppl，pp.5–8，2005.

[17] 神原啓介，塚田浩二，"オノマトペン，" インタラクティブシステムとソフトウェアに関するワークショップ（WISS）2008 予稿集，pp.79–84，2008.

[18] W. Köhler, Gestalt Psychology, Horace Liveright, New York, NY, USA, 1929.

[19] V.S. Ramachandran and E.M. Hubbard, "Synaesthesia —A window into perception, thought and language," J. of Consciousness Studies, vol.8, no.12, pp.3–34, 2001.

[20] 小倉慶郎，"日英オノマトペの考察：日英擬音語・擬態語の全体像を概観する，" 大阪大学日本語日本文化教育センター授業研究，vol.14，pp.23–33，2016.

[21] 呂　佳蓉，"英語のオノマトペの象徴メカニズム，" 言語科学論集，vol.10，pp.99–116，2004.

[22] 石原一志，坪田　康，奥乃　博，"日本語の音節構造に着目した環境音の擬音語への変換，" 電子情報通信学会技術研究報告，SP2003-38，2003.

[23] 比屋根一雄，澤部直太，飯尾　淳，"単発音のスペクトル構造とその擬音語表現に関する検討，" 電子情報通信学会技術研究報告，SP97-125，1998.

[24] S. Sundaram and S. Narayanan, “Analysis of audio clustering using word descriptions,” Proc. 2007 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol.2, pp.769–772, 2007.

[25] S. Sundaram and S. Narayanan, “Classification of sound clips by two schemes: Using onomatopoeia and semantic labels,” Proc. 2008 IEEE Int. Conf. on Multimedia and Expo, pp.1341–1344, 2008.

[26] T. Fukusato and S. Morishima, “Automatic depiction of onomatopoeia in animation considering physical phenomena,” Proc. 7th ACM Int. Conf. on Motion in Games, pp.161–169, 2014.

[27] 權　眞煥，川嶋卓也，下田　和，坂本真樹，“DCNN を用いた画像の質感認知—音象徴性からのアプローチ—，” 第 31 回人工知能学会全国大会，2L3-OS-09b-1，2017.

[28] W. Shimoda and K. Yanai, “A visual analysis on recognizability and discriminability of onomatopoeia words with DCNN features,” Proc. 2015 IEEE Int. Conf. on Multimedia and Expo, pp.1–6, 2015.

[29] M. Sakamoto, Y. Ueda, R. Doizaki, and Y. Shimizu, “Communication support system between Japanese patients and foreign doctors using onomatopoeia to express pain symptoms,” J. of Advanced Computational Intelligence and Intelligent Informatics, vol.18, no.6, pp.1020–1025, 2014.

[30] 清水祐一郎，土斐崎龍一，坂本真樹，“オノマトペごとの微細な印象を推定するシステム，” 人工知能学会論文誌，vol.29，no.1，pp.41–52，2014.

[31] R. Doizaki, J. Watanabe, and M. Sakamoto, “Automatic estimation of multidimensional ratings from a single sound-symbolic word and word-based visualization of tactile perceptual space,” IEEE Trans. on Haptics, vol.10, no.2, pp.173–182, 2017.

[32] 戸本裕太郎，中村剛士，加納政芳，小松孝徳，“音素特徴に基づくオノマトペの可視化，” 日本感性工学論文誌，vol.11，no.4，pp.545–552，2012.

[33] 秋山広美，小松孝徳，清河幸子，“オノマトペから感じる印象の客観的数値化方法の提案，” 情報処理学会研究報告，2011-HCI-142（23），2011.

[34] T. Komatsu, “Quantifying Japanese onomatopoeias: Toward augmenting creative activities with onomatopoeias,” Proc. 3rd ACM Int. Conf. on Augmented

Humans, no.15, pp.1–4, 2012.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol.86, no.11, pp.2278–2324, 1998.

[36] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, vol.25, pp.1097–1105, 2012.

[37] 鍵谷龍樹，白川由貴，土斐崎龍一，渡邊淳司，丸谷和史，河邉隆寛，坂本真樹，“動画と静止画から受ける粘性印象に関する音象徴性の検討，” 人工知能学会論文誌，vol.30，no.1，pp.237–245，2015.

[38] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.35, no.1, pp.221–231, 2013.

[39] L. Sharan, R. Rosenholtz, and E.H. Adelson, “Accuracy and speed of material categorization in real-world images,” J. of Vision, vol.14, no.10, pp.1–24, 2014.

[40] Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, “Joint intensity and spatial metric learning for robust gait recognition,” Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, pp.5705–5715, 2017.

[41] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, “Joint intensity transformer network for gait recognition robust against clothing and carrying status,” IEEE Trans. on Information Forensics and Security, vol.14, no.12, pp.3102–3115, 2019.

[42] A. Sakata, N. Takemura, and Y. Yagi, “Gait-based age estimation using multi-stage convolutional neural network,” IPSJ Trans. on Computer Vision and Applications, vol.11, no.1, pp.1–10, 2019.

[43] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol.1, pp.886–893, 2005.

[44] L. Cao, M. Dikmen, Y. Fu, and T.S. Huang, “Gender recognition from body,” Proc. 16th ACM Int. Conf. on Multimedia, pp.725–728, 2008.

[45] D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. of Computer Vision, vol.60, no.2, pp.91–110, 2004.

[46] Y. Ge, J. Lu, X. Feng, and D. Yang, “Body-based human age estimation at a distance,” Proc. 2013 IEEE Int. Conf. on Multimedia and Expo Workshops, pp.1–4, 2013.

[47] J.W. Davis, “Visual categorization of children and adult walking styles,” Proc. 3rd Int. Conf. on Audio- and Video-based Biometric Person Authentication, pp.295–300, 2001.

[48] J. Man and B. Bhanu, “Individual recognition using gait energy image,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.28, no.2, pp.316–322, 2006.

[49] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, “A study on gait-based gender classification,” IEEE Trans. on Image Processing, vol.18, no.8, pp.1905–1910, 2009.

[50] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, pp.6299–6308, 2017.

[51] T.N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” arXiv preprint, arXiv:1609.02907, 2016.

[52] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, “Disentangling and unifying graph convolutions for skeleton-based action recognition,” Proc. 2020 IEEE Conf. on Computer Vision and Pattern Recognition, pp.143–152, 2020.

[53] J.J. Gibson, The Perception of the Visual World, Houghton Mifflin, Boston, MA, USA, 1950.

[54] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” arXiv preprint, arXiv:1406.2199, 2014.

[55] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” Proc. 2007 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp.1–8, 2007.

[56] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, “Dense trajectories and motion boundary descriptors for action recognition,” Int. J. of Computer Vision, vol.103, no.1, pp.60–79, 2013.

[57] G.E. Hinton and R.R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol.313, no.5786, pp.504–507, 2006.

[58] A. Hernandez, J. Gall, and F. Moreno-Noguer, “Human motion prediction via spatio-temporal inpainting,” Proc. 17th IEEE Int. Conf. on Computer Vision, pp.7134–7143, 2019.

[59] U. Bhattacharya, N. Rewkowski, P. Guhan, N.L. Williams, T. Mittal, A. Bera, and D. Manocha, “Generating emotive gaits for virtual agents using affect-based autoregression,” Proc. 2020 IEEE Int. Symposium on Mixed and Augmented Reality, pp.24–35, 2020.

[60] L. Kovar, M. Gleicher, and F. Pighin, “Motion graphs,” Proc. 29th ACM Annual Conf. on Computer Graphics and Interactive Techniques, pp.473–482, 2002.

[61] Y. Zhang, Y. Makihara, D. Muramatsu, J. Zhang, L. Niu, L. Zhang, and Y. Yagi, “Learn to walk across ages: Motion augmented multi-age group gait video translation,” IEEE Access, vol.9, pp.40550–40559, 2021.

[62] D. Holden, J. Saito, and T. Komura, “A deep learning framework for character motion synthesis and editing,” ACM Trans. on Graphics, vol.35, no.4, pp.1–11, 2016.

[63] K. Chen, Z. Tan, J. Lei, S.-H. Zhang, Y.-C. Guo, W. Zhang, and S.-M. Hu, “Choreomaster: choreography-oriented music-driven dance synthesis,” ACM Trans. on Graphics, vol.40, no.4, pp.1–13, 2021.

[64] M.Z. Uddin, T.T. Ngo, Y. Makihara, N. Takemura, X. Li, D. Muramatsu, and Y. Yagi, “The OU-ISIR large population gait database with real-life carried object and its performance evaluation,” IPSJ Trans. on Computer Vision and Applications, vol.10, no.1, pp.1–11, 2018.

[65] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, “The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits,” J. of Visual Communication and Image Representation, vol.25, no.1, pp.195–206, 2014.

[66] S. Yu, D. Tan, and T. Tan, “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,” Proc. 18th Int. Conf. on

Pattern Recognition, vol.4, pp.441–444, 2006.

[67] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer, “The HumanID gait challenge problem: Data sets, performance, and analysis,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.27, no.2, pp.162–177, 2005.

[68] J.D. Shutler, M.G. Grant, M.S. Nixon, and J.N. Carter, “On a large sequence-based human gait database,” Applications and Science in Soft Computing, Advances in Soft Computing, vol.24, pp.339–346, 2004.

[69] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” arXiv preprint, arXiv:1705.06950, 2017.

[70] “Flea3 USB3 | Teledyne FLIR,” https://www.flir.jp/products/flea3-usb3/ [2021/12/15 閲覧].

[71] 伊藤惇貴，加納政芳，中村剛士，小松孝徳，“オノマトペの音象徴属性値の調整のための一手法，” 人工知能学会論文誌，vol.30，no.1，pp.364–371，2015.

[72] 小松孝徳，秋山広美，“ユーザの直感的表現を支援するオノマトペ表現システム，” 電子情報通信学会論文誌（A），vol.J92-A，no.11，pp.752–763，2009.

[73] 黒川伊保子，怪獣の名はなぜガギグゲゴなのか，新潮社，東京，2004.

[74] 黒川伊保子，“商標評価手法の—考察—ことばの感性評価，” 知財管理，vol.56，no.5，pp.745–752，2006.

[75] V. Ramakrishna, D. Munoz, M. Hebert, J.A. Bagnell, and Y. Sheikh, “Pose machines: Articulated pose estimation via inference machines,” Proc. 13th European Conf. on Computer Vision, Part II, Lecture Notes in Computer Science, vol.8690, pp.33–47, 2014.

[76] Q. Li, Y. Wang, A. Sharf, Y. Cao, C. Tu, B. Chen, and S. Yu, “Classification of gait anomalies from Kinect,” The Visual Computer, vol.34, no.2, pp.229–241, 2018.

[77] 杉山雄紀，近藤敏之，“ロボットの歩行動作設計によるオノマトペ・情報表現の共通理解，” 第 25 回人工知能学会全国大会，1C1-OS4a-4，2011.

[78] L.L. Thurstone, “A law of comparative judgment,” Psychological Review, vol.34,

no.4, pp.273–286, 1927.

[79] C. Spearman, “The proof and measurement of association between two things,” American J. of Psychology, vol.15, no.1, pp.72–101, 1904.

[80] “Keras documentation,” https://keras.io/ [2021/12/15 閲覧].

[81] 井上正明，小林利宣，“日本における SD 法による研究分野とその形容詞対尺度構成の概観，” 教育心理学研究，vol.33，no.3，pp.253–260，1985.

[82] 佐藤　順，森島繁生，“音声に込められた感情の意味次元に関する検討，” 電子情報通信学会技術研究報告，HCS97-99，1997.

[83] 大山　正，瀧本　誓，岩澤秀紀，“セマンティック・ディファレンシャル法を用いた共感覚性の研究，” 行動計量学，vol.20，no.2，pp.55–64，1993.

[84] 北村音壱，“再生音の心理的評価について，” 電気通信学会電気音響研究会専門委員会資料，1962.

[85] Y. Kano and A. Harada, “Stepwise variable selection in factor analysis,” Psychometrika, vol.65, no.1, pp.7–22, 2000.

[86] 加藤大貴，平山高嗣，道満恵介，川西康友，井手一郎，出口大輔，村瀬　洋，“音象徴性を利用したオノマトペによる歩容の記述，” 人工知能学会論文誌，vol.33，no.4，pp.B–HC2_1–9，2018.

[87] L.A. Gatys, A.S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pp.2414–2423, 2016.

[88] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” Proc. 16th IEEE Int. Conf. on Computer Vision, pp.1501–1510, 2017.

[89] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition, pp.4401–4410, 2019.

[90] M. Okumura, H. Iwama, Y. Makihara, and Y. Yagi, “Performance evaluation of vision-based gait recognition using a very large-scale gait database,” Proc. 2010 IEEE Int. Conf. on Biometrics: Theory, Applications and Systems, pp.1–6, 2010.

[91] J. Carifio and R. Perla, “Resolving the 50-year debate around using and misusing

Likert scales," Medical Education, vol.42, no.12, pp.1150–1152, 2008.

[92] F. Wilcoxon, "Individual comparisons by ranking methods," Breakthroughs in statistics, vol.2, pp.196–202, Springer-Verlag, New York, NY, USA, 1992.

[93] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," arXiv preprint, arXiv:1406.2661, 2014.

[94] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," Proc. 34th Int. Conf. on Machine Learning, pp.2642–2651, 2017.

研究業績

学術雑誌

- 加藤大貴，平山高嗣，道満恵介，川西康友，井手一郎，出口大輔，村瀬　洋，“音象徴性を利用したオノマトペによる歩容の記述，” vol.33，no.4，p.B-HC2_1–9，人工知能学会論文誌，July 2018.
- 加藤大貴，平山高嗣，道満恵介，川西康友，井手一郎，出口大輔，村瀬　洋，“音象徴性に基づく擬態語からの歩容の生成，” vol.36，no.5，p.D-KC7_1–10，人工知能学会論文誌，Sept. 2021.

国際会議

- Hirotaka Kato, Takatsugu Hirayama, Keisuke Doman, Yasutomo Kawanishi, Ichiro Ide, Daisuke Deguchi, and Hiroshi Murase, “Toward describing human gaits by onomatopoeias,” Proc. 2017 IEEE Int. Conf. on Computer Vision (ICCV2017) Workshop on Analysis and Modeling of Faces and Gestures, pp.1573–1580, Oct. 2017.
- Hirotaka Kato, Takatsugu Hirayama, Keisuke Doman, Yasutomo Kawanishi, Ichiro Ide, Daisuke Deguchi, and Hiroshi Murase, “More-natural mimetic words generation for fine-grained gait description,” Proc. 26th Int. Conf. on Multimedia Modeling (MMM), pp.214–225, Jan. 2020.

研究会・シンポジウム

- 加藤大貴，平山高嗣，川西康友，道満恵介，井手一郎，出口大輔，村瀬　洋，“人体部位の相対的位置関係を利用したオノマトペ歩容映像の識別に関する検討，” 情報処理学会研究報告, 2016-CVIM-202-25，May 2016.
- 加藤大貴，平山高嗣，川西康友，道満恵介，井手一郎，出口大輔，村瀬　洋，“オノマトペにより歩容を記述するための音韻空間と人体部位の動きの関係性，” HCG シンポジウム 2016，A-3-5，Dec. 2016.
- 加藤大貴，平山高嗣，川西康友，道満恵介，井手一郎，出口大輔，村瀬　洋，“音韻と人体部位の動きの関係に着目したオノマトペによる歩容の記述に向けて，” 電子情報通信学会技術研究報告，PRMU2017-25，June 2017.
- 加藤大貴，平山高嗣，道満恵介，川西康友，井手一郎，出口大輔，村瀬　洋，“オノマトペによる歩容の記述の高精度化に向けたデータセットの構築，” 第 21 回画像の認識・理解シンポジウム（MIRU2018），PS3-44，Aug. 2018.
- 加藤大貴, 平山高嗣, 道満恵介, 川西康友, 井手一郎, 出口大輔, 村瀬　洋, “HOYO : オノマトペを付与した歩容データセット，” 電子情報通信学会技術研究報告，MVE2018-21，Sept. 2018.

全国大会・支部大会

- 加藤大貴，平山高嗣，川西康友，井手一郎，出口大輔，村瀬　洋，“オノマトペ表現に対応した歩容映像の識別に関する検討，” 平成 27 年度電子情報通信学会東海支部卒業研究発表会，D-1-3，Mar. 2016.
- 加藤大貴，平山高嗣，川西康友，井手一郎，出口大輔，村瀬　洋，“音韻と人体部位の動きの関係に着目したオノマトペによる歩容の zero-shot 記述に向けた検討，” 電気・電子・情報関係学会東海支部連合大会，B2-6，Sept. 2017.