Vol.2019-HCI-184 No.1
2019/7/22

# ジェスチャ自動生成に向けたLSTMを用いたレストフェーズの判定

日和　航大[1,a)]　長谷川　大[2]　金子　直史[3]　白川　真一[4]　荒木　健治[5]

**概要**：データドリブンアプローチにより音声信号から発話に伴うジェスチャモーションを生成する試みが行われており，よりリッチなデータセットが求められている．本稿では，既存データセットの一部に人手によるジェスチャフェーズのアノテーションを行い，DNN を用いて残りのデータセットに対するジェスチャフェーズの推定を行う方法を提案する．提案手法では，LSTM を含む 3 層ネットワークにより，ジェスチャモーションとジェスチャフェーズの関連性を学習し，レスト状態と非レスト状態の識別モデルを作成した．提案手法の有効性を確認するため，速度閾値による判定手法との比較実験を行った．実験の結果，提案手法によって速度閾値による判定手法を上回る判定精度（適合率 0.822，再現率 0.853，F 値 0.837）が得られることを確認した．

## 1. はじめに

人間のコミュニケーションにおいて，非言語的要素は重要な役割を担う．近年，人間のような外観をもつ擬人化エージェントや Embodied Conversational Agent（ECA）[1] と呼ばれるインタフェースを導入することで，ジェスチャをはじめとする非言語的要素を利用し，人対コンピュータの様々なインタラクションに改善をもたらす試みが数多くみられる [2], [3].

これらの人型インタフェースにおいて，現状ではジェスチャを生成する一般的な手段として，人手でジェスチャを作成する，もしくはモーションキャプチャから実際のジェスチャを取得する方法が採用されている．しかし，これらの手法はソフトウェアの適用や適したジェスチャの生成に専門的な知識や技術を要することや費用が掛かるといった点から，コストの高い作業となっている．そのため，音声またはテキストからジェスチャを自動的に生成するための研究が行われている．

Chiu ら [4] は Deep Conditional Neural Field(DCNF) モデルを提案し，発話音声や基本周波数などの韻律的特徴とテキストを入力に，14 種類のジェスチャのうち最も適するジェスチャを時系列の適切な時点に割り振る予測タスクを行った．高い予測精度が報告されているが，事前に定義したジェスチャ以外を出力できないため多様な形態で産出されうるジェスチャのバリエーションに対応することが困難であると考えられる．Chiu ら [5] は，Hierarchical Factored Conditional Restricted Boltzmann Machine (HFCRBM) を改良したモデルにより，発話音声の大きさとピッチによる韻律的特徴から，発話の強調やリズムに関連のあるジェスチャモーションの生成を試みている．しかしながら，本手法では音素認識に必要な音韻的特徴が考慮されておらず，Iconic，Deictic，Metaphoric などの発話の意味的内容に関連の深いジェスチャの生成を行うことが難しい．

これらの手法と比較して，Hasegawa ら [6] は，発話の音韻的特徴を入力にした Bi-directional LSTM を含む 5 層のネットワークを用いることで，発話音声の音韻的特徴に基づくジェスチャの生成法の検討を行っている．本手法により生成されたジェスチャはランダムに割り当てられたジェスチャよりも，発話音声に対して自然なジェスチャと知覚されることが確認できているが，単語の意味との関連性など意味的に整合性のあるジェスチャの生成には至っていない．このことの一因として，以下に述べるようなデータセットの問題が挙げられる．Hasegawa ら [6] は，210 分の

[1] 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University
[2] 北海学園大学工学部
Faculty of Engineering, Hokkai Gakuen University
[3] 青山学院大学理工学部情報テクノロジー学科
Department of Integrated Information Technology, Aoyama Gakuin University
[4] 横浜国立大学大学院環境情報研究院
Faculty of Environment and Information Sciences, Yokohama National University
[5] 北海道大学大学院情報科学研究院
Graduate School of Information Science and Technology, Hokkaido University
[a)] kururi@eis.hokudai.ac.jp

データセットを用いているが，音韻的特徴から十分な量とはいえない．しかし，音声とジェスチャをペアとしたデータ作成にはモーションキャプチャなどの機材を必要とするため，データセットの拡張は容易ではない．そのため，既存データセットのモーションデータを分析しデータセットに有益な情報を与えることには，自動生成のジェスチャを改善する可能性があると考えられる．本来は全てのデータに対して人手でアノテーションを行うことが望ましいが，データセットが大きくかつ緻密な作業を要する場合，作業量が非常に大きくなる．そこで本稿では，既存データセットのモーションデータの一部に人手によるアノテーションを行い，各フレームを動作の停止状態やジェスチャに含まれない小さな動作を示すレストフレームまたはジェスチャとなる動作を示す非レストフレームの 2 種類のフレームタイプに分類した後に，残りのモーションデータのフレームタイプの推定を行う方法を提案する．判定方法として，速度閾値による判定手法とニューラルネットワークを用いた判定手法の 2 つのアプローチを検討した．ニューラルネットワーク手法では，時系列考慮が可能な RNN，GRU，LSTM の 3 種類のモデル毎にネットワークを構築した．比較実験の結果，LSTM を含む 3 層ネットワークにおいて速度閾値による手法の結果を上回り，その有効性が確認された．

## 2. 関連研究

Naguri ら [7] は両手のモーションデータの座標位置と速度を入力にした LSTM ネットワークを用いて，モーションデータのフレーム毎にジェスチャフレームと非ジェスチャフレームの 2 種類のフレームタイプの分類を行った．時系列の考慮は現時点のフレームと約 0.15 秒前のフレームの計 2 フレームと少なく，ネットワークの構造は 1 層の LSTM ネットワークと浅い構造でありながらも，ジェスチャフレームの判定精度において高精度な結果を報告している．しかしながら本稿で扱うデータはモーションキャプチャから獲得した全身のデータであるため，より複雑なモーションデータに適応した識別モデルの構築が必要である．そこで，ネットワークへの入力時により多くの時系列情報の追加やネットワークの多層化を行うことで，識別モデルの判定精度の向上を検討する．

ジェスチャの最中に身体の動きが一時中断することがあり，そのような停止状態にはジェスチャの強調や次のジェスチャの用意を表現する機能がある．Hasegawa らの手法で生成されたジェスチャは，絶え間なく腕を一定の範囲で上下に動かす動作が目立ち停止状態を表現できていなく，今後のジェスチャの生成段階では，ジェスチャフレームよりもレストフレームのタグ情報の考慮が必要であると考えられる．そのため，本稿の実験での判定精度の評価対象はジェスチャフレームではなくレストフレームとした．

## 3. アノテーション

### 3.1 既存データセット

Takeuchi ら [8] が作成したデータセットを対象に，アノテーションを行う．当データセットには，合計 1,047 センテンス (約 210 分) の mp3 形式の発話の音声データと付随する BioVisionHierarchy(BVH) 形式のモーションデータが記録されている．データの記録は，話者があるトピックについて聴者に一方的に話すスピーチ形式で行われた．モーションデータは，全身を 64 関節の xyz 座標位置とオイラー回転角で表現した時系列データである．

### 3.2 ジェスチャタイプとジェスチャフェーズ

ジェスチャには，その動作の構成単位となるジェスチャフェーズとジェスチャの種類を示すジェスチャタイプが存在する．McNeill[9] はジェスチャタイプとジェスチャフェーズを以下のように分類している．

- ジェスチャタイプ
  - Iconic(図像的) ジェスチャ：
    具体的な事物の視覚的イメージを伝える動作
  - Beat(拍子的) ジェスチャ：
    発話の強弱時に手でリズムを取るような動作
  - Deictic(指示的) ジェスチャ：
    話者の周辺環境にある事物を指し示す動作
  - Metaphoric(隠喩的) ジェスチャ：
    抽象的な事物を 3 次元空間に写像して表現する動作
- ジェスチャフェーズ
  - Rest：休止状態
  - Preparation(Prep)：ジェスチャの準備段階の動作
  - Pre-stroke Hold(PreHold)：Stroke 前の停止状態
  - Stroke：ジェスチャの中心動作
  - Post-stroke Hold(PostHold)：Stroke 後の停止状態
  - Retraction：休止状態に移る動作

Stroke はジェスチャの核となる動作を示す．他のフェーズは Stroke の前後に組織され，ジェスチャによっては省略されることがある．

### 3.3 アノテーション方法

アノテーションデータとして，Takeuchi ら [8] のデータセット内の 240 センテンス分のモーションキャプチャデータを対象にした．動画分析ツール Anvil[*1]を用い，25fps の粒度でフレーム毎に 4 種類のジェスチャタイプと 7 種類のジェスチャフェーズのタグ付けを行った．また発話内容を記録し，Stroke と関連性が高い発話内容については Stroke 別に記録した．これはジェスチャに対する複数のラベリン

[*1] https://www.anvil-software.org

グを行うことは，今後のジェスチャ解析に有効なためである．図 1 にアノテーションの一例を示す．


図 1　アノテーションの例

アノテーション後のジェスチャタイプの数を表 1 に，ジェスチャフェーズの数を表 2 に示す．表 1，表 2 に示すように，ジェスチャタイプは Metaphoric ジェスチャ，ジェスチャフェーズは Stroke の出現割合が多くなっている．

表 1　ジェスチャタイプの数

| Iconic | Beat | Deictic | Metaphoric |
|---|---|---|---|
| 104 | 144 | 77 | 521 |

表 2　ジェスチャフェーズの数

| Rest | Prep | PreHold | Stroke | PostHold | Retraction |
|---|---|---|---|---|---|
| 411 | 575 | 145 | 832 | 397 | 362 |

### 3.4　正解データ

フレームに付与されたジェスチャフェーズのラベルに応じてフレームタイプを決定し，240 センテンスの正解データを作成する．ジェスチャフェーズの内，停止状態の Rest，PreHold，PostHold をレストフレームとし，動作が生じている状態の Prep，Stroke，Retraction を非レストフレームとした．またジェスチャに含まれない小さな動作についてはレストフレームとした．フレームタイプを決定する際，フレームの粒度を 25fps から 20fps に変更する処理を行う．これは，Hasegawa らの自動生成ジェスチャが 20fps の粒度で生成されているためである．正解データの内訳を表 3 に示す．

表 3　正解データの数

| レストフレーム | 非レストフレーム |
|---|---|
| 25,355 | 28,703 |

## 4.　速度閾値による判定手法

ここでは，本稿でのベースラインとなる速度閾値による判定手法の概要について示す．

### 4.1　概要

ジェスチャが発生する場合とそうではない場合や，ジェスチャフェーズによって両手の速度に変化が生じることに着目した判定手法である．速度は座標位置の数値の 1 フレーム分の差分から導出した．速度の単位は cm/frame である．速度閾値による判定条件は，片手もしくは両手の対象部位の速度が閾値以下のフレームをレストフレームとし，閾値を上回ったフレームを非レストフレームとした．なお速度閾値による判定の対象部位として，本稿では仮に手首と中指の指先に設定した．

### 4.2　予備実験

速度閾値の最適なパラメータを獲得するために予備実験を行った．アノテーション対象の 240 センテンスのうち 216 センテンス（90%）に対して，レストフレーム判定の F 値が最高値となるパラメータの値を検討した．速度閾値は 0.0 から 3.0 の範囲とし，対象となる身体部位の速度毎に 0.1 ずつ値を変えていく．また対象部位の速度の補正として速度の移動平均値，速度の中央値を速度との比較対象にした．これは，モーションキャプチャから獲得したモーションには多少の揺れが生じているが，それによる不必要な速度の変化を均すための補正である．移動平均値と中央値の窓幅は 3 から 19 の範囲で奇数のみを対象にする．予備実験の評価指標として，システムが出力したレストフレームの適合率，再現率，F 値を求めた．F 値は全センテンスの適合率と再現率の平均値を用いて導出した．実験結果から得た最適なパラメータの値を表 4 に示す．

表 4　速度閾値の最適なパラメータ

| 条件 | 窓幅 | 指先 | 手首 | 適合率 | 再現率 | F 値 |
|---|---|---|---|---|---|---|
| 移動平均 | 9 | 0.7 | 0.6 | 0.691 | 0.859 | 0.734 |

## 5.　提案手法

ここでは，提案手法となるニューラルネットワーク手法の概要について示す．

### 5.1　概要

ニューラルネットワークを用いてジェスチャモーションとジェスチャフェーズの関連性を学習し，レスト状態と非

レスト状態の識別モデルを作成する手法である．ネットワークは TensorFlow[10] をバックエンドとし Keras[11] で実装した．ネットワークの構造は図 2 に示すように 3 層の隠れ層からなる．

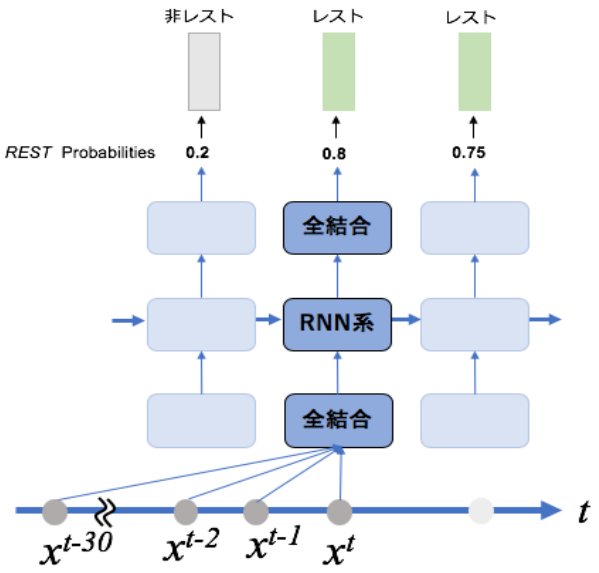


図 2　3 層ネットワークの構造

1 層目 $h^{(1)}$ と 3 層目 $h^{(3)}$ は再帰ではない全結合層であり，2 層目 $h^{(2)}$ に過去の時系列の考慮が可能な RNN 系の層を用いる．1 層目では，過去の時系列情報を十分に考慮できるように過去 30 ステップ分のモーションデータを結合したベクトル $x(t)$ を入力として受け取る．$h^{(1)}$ は式 (1) のように計算できる．

$$h_t^{(1)} = g(W^{(1)}h_t^{(0)} + b^{(1)}) \tag{1}$$

ここで，$g(z)$ は Rectified-Linear Unit(ReLU) の活性化関数，$W^{(l)}$ と $b^{(l)}$ はそれぞれ $h^{(l)}$ の重みの行列とバイアスパラメータを示す．

2 層目は RNN 系レイヤの層になっている．この層は，$t = 0$ から順に再帰を行う前向きの再帰型ユニット $h^{(f)}$ が含まれる．$h^{(2)}$ は式 (2) のように計算できる．なお，用いる RNN 系レイヤのモデル毎の概要については **5.2** で述べる．

$$h_t^{(2)} = g(W^{(2)}h_t^{(1)} + W_r^{(f)}h_{t-1}^{(f)} + b^{(2)}) \tag{2}$$

3 層目は出力層となっており，前の層の過去の情報を考慮した再帰ユニットからの出力を入力に受け取り，重みとバイアスを合わせて演算した後，Sigmoid 関数 $j(z)$ を用いて活性化を行う．$h^{(3)}$ は式 (3) のように計算できる．

$$h_t^{(3)} = j(W^{(3)}h_t^{(2)} + b^{(3)}) \tag{3}$$

出力はレスト状態の確率値となる．出力結果に対して binary cross entropy を算出し，Adadelta[12] を用いて最適化を行う．フレームタイプを決定するために，出力値が 0.5 以上の時はレストフレーム，0.5 を下回る時は非レストフレームとした．最終的な出力のベクトル形式は，タイムステップ数 × 1 となる．

### 5.2　実験で用いるニューラルネットワークの時系列モデル

**5.1** で示すように，ネットワークの 2 層目に過去の時系列考慮が可能なニューラルネットワークモデルとして RNN，GRU，LSTM をネットワークを適用しその影響を調べる．以下，モデル毎に概要を示す．

Recurrent Neural Networks (RNN)[13] は前時刻の中間層の出力を扱うことで，時系列の影響を考慮した学習を可能にする．しかし，実際には誤差逆伝搬による学習を行う際，勾配が消失する勾配消失問題が存在するため，長期依存の系列の学習が正しく行われない欠点があった．Long short-term memory(LSTM)[14] は，勾配消失の問題を緩和するために提案されたネットワークの 1 つである．このメモリセルにより，LSTM は通常の RNN よりも記憶容量が拡張し，長期記憶が可能になっている．Gated Reccurent Unit (GRU)[15] は LSTM 内部の構造の簡潔化に成功したモデルである．LSTM と比較して高速に処理を行うことが可能であり，かつ同程度の表現力を保持している．

### 5.3　データの前処理

入力の座標位置と速度では，取りうる値の範囲が大きく異なるため標準化によるスケーリングを行った．標準化はデータ $X$ を平均 0，分散 1 の $Y$ に変換する正規化法である．これにより異なるスケールの特徴量でも，その大小を比較することが可能となる．標準化は以下の式 (4) で表される．なお $\mu$ は X の平均，$\sigma$ は X の分散とする．

$$Y = \frac{X - \mu}{\sigma} \tag{4}$$

216 センテンスの正解データを用いて特徴量毎に平均値と標準偏差のパラメータを獲得した．

また本手法には 2 種類の入力形式がある．まず Naguri らの手法と同様のモーションデータの座標位置と速度である．64 関節の座標位置と速度の xyz 座標を扱うため，モーションの特徴量は 384 次元である．入力のベクトル形式は，タイムステップ数 × $(30+1)$ × 384 となる．次にモーションデータの座標位置と速度に加え，フレーム毎に速度閾値手法の結果を含めた入力である．速度閾値による判定手法を考慮しつつ学習を行うことで，識別モデルの表現力が増すため判定精度が向上すると考えた．速度閾値の判定結果は 1 次元のバイナリ値であるため，入力のベクトル形式はタイムステップ数 × $(30+1)$ × 385 となる．

## 6.　性能評価実験

### 6.1　実験方法

ここでは，速度閾値による判定手法とニューラルネッ

トワーク手法の比較実験の概要を示す．ニューラルネットワーク手法では **5.2** の 3 種類の再帰型ニューラルネットワークと，**5.3** の 2 種類の入力形式がある．そのため，Naguri らの手法の 1 層 LSTM，速度閾値手法，座標位置と速度である 3 層 LSTM，3 層 GRU，3 層 RNN，入力形式が座標位置と速度と速度閾値の判定結果である 3 層 LSTM，3 層 GRU，3 層 RNN の計 8 種類が比較対象となる．

アノテーションによって作成された 240 センテンスの正解データのうち，24 センテンス（10%）をテストデータとした．テストデータは全部で 6,934 フレームであり，レストフレームが 3,412 フレームで非レストフレームが 3,522 フレームである．速度閾値による判定手法では表 4 で示した最適なパラメータで，テストデータによる評価を行う．ニューラルネットワーク手法では残りの 216 センテンスは学習に用い，**5.3** で示したように，入力データのスケーリングを行う．学習データの特徴量から標準化を行うための平均値と分散のパラメータを獲得し，テストデータの特徴量に適用する．

Naguri らの手法を含むニューラルネットワーク手法の全 7 手法のネットワークのエポック数は 500 に設定した．比較実験の評価指標として，システムが出力したレストフレームの適合率，再現率，F 値を求めた．F 値は全センテンスの適合率と再現率の平均値を用いて導出した．

### 6.2 実験結果

表 5 に比較実験の結果を示す．入力にモーションデータの座標位置と速度に加えて各フレームに速度閾値手法の結果を含めた 3 層 LSTM(3 層 LSTM+速度閾値) において，F 値の最高値 0.837 となった．ニューラルネットワーク手法における全ての提案手法で Naguri らの手法と比較して両側 t 検定を行ったところ有意差を示した ($p < 0.001$)．ゆえに過去の時系列情報の追加，多層化の有効性が確認できた．また，ニューラルネットワーク手法の提案手法が速度閾値による判定手法の F 値を 0.135 ポイント上回り，有意差を示した ($p < 0.001$)．

**表 5** 比較実験の結果，$^{***}p < 0.001$

| 手法 | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| 速度閾値手法 | 0.573 | 0.906 | 0.702 |
| Naguri ら [7] の手法 | 0.786 | 0.734 | 0.759 |
| 3 層 RNN | 0.835 | 0.791 | 0.812*** |
| 3 層 GRU | 0.795 | 0.824 | 0.810*** |
| 3 層 LSTM | 0.792 | 0.851 | 0.820*** |
| 3 層 RNN+速度閾値 | 0.816 | 0.850 | 0.833*** |
| 3 層 GRU+速度閾値 | 0.806 | 0.842 | 0.824*** |
| **3 層 LSTM+速度閾値** | **0.822** | **0.853** | **0.837***** |

## 7. 考察

図 3 は 3 層 LSTM+速度閾値による識別モデルの出力値と正解データを示す．ファイル id=297 の 310 ステップ付近や id=315 の 150 ステップ付近のように，突発的にフレームタイプが変化したために正解データとフレームタイプが異なるフレームが見受けられた．新たなデータセットの作成時にノイズになる可能性があるため，そのようなフレームを修正する後処理が必要である．そこで，突発的にフレームタイプが変化したフレームにおいて一定の前後幅でフレームタイプに変化がない場合，該当フレームのフレームタイプを変更する後処理を検討している．「突発的なフレーム」と判定するフレーム幅や，フレームの前後幅といった最適なパラメータの値は今後明らかにしていきたい．

ファイル id=297 の 20 から 50 ステップ付近のように突発的なフレームタイプの変化が連続的に起こる場合，上記の後処理の適用は難しい．突発的なフレーム変化が連続して起こる範囲を特定し前後のフレームを考慮することで，適切なフレームタイプの修正を行う後処理の条件を定めていきたい．

## 8. まとめと展望

本稿では，既存データセットの一部に人手によるジェスチャフェーズのアノテーションを行い，残りのデータセットに対するジェスチャフレームの推定を行う方法として，速度閾値による判定手法とニューラルネットワーク手法の 2 つのアプローチを検討した．速度閾値による判定手法では，予備実験から獲得した手首と中指の指先の速度の閾値に応じてフレームタイプを判定した．ニューラルネットワーク手法では，ジェスチャモーションとジェスチャフェーズの関連性を学習し，レスト状態と非レスト状態のフレームタイプの識別モデルを作成した．結果は，速度閾値の判定結果を含めた 3 層 LSTM は速度閾値による判定手法の F 値を 0.135 ポイント上回り，また Naguri らの手法と比較して手法の有効性が確認された．

今後は，識別モデルの出力に対し後処理を行い，新たに作成したデータセットを用いてジェスチャの生成段階に入る．後処理については，突発的にフレームタイプが変化するフレームの修正を行っていきたい．ジェスチャの生成法については，アノテーション結果を考慮した学習が可能なマルチタスクラーニングを検討している．ネットワークは Hasegawa らの手法を参考にしつつ，新たに各フレームのフレームタイプを強調したモーションデータを出力する損失関数を設計することで，ジェスチャの改善が見込まれると考えられる．改善されたジェスチャと従来手法のジェスチャで印象評価を行い，ジェスチャの質が従来手法を上回ることを示していきたい．

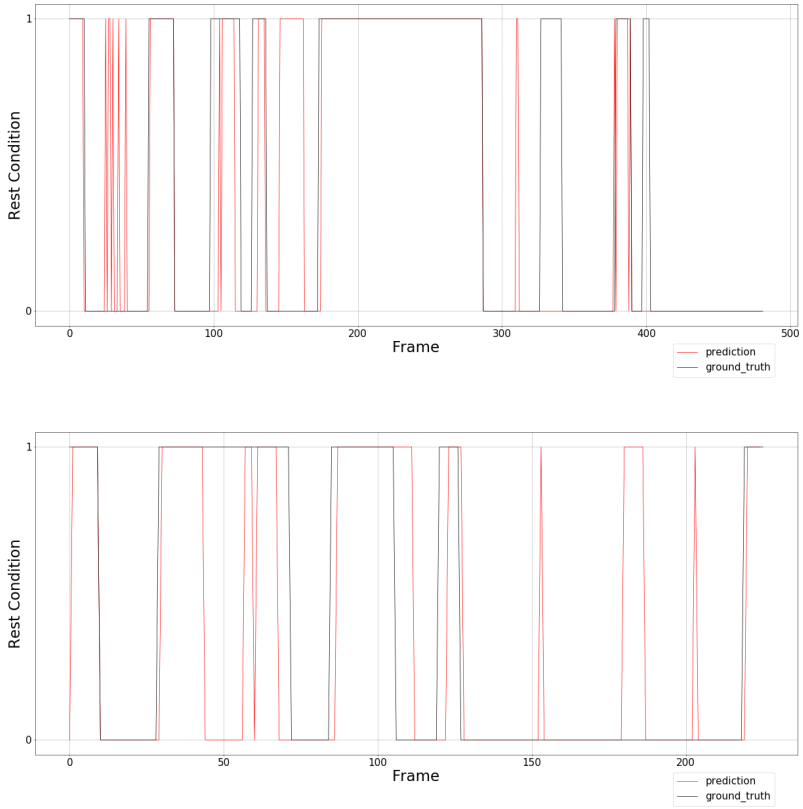


**図 3** id=297(上図) と id=315(下図) でのシステムと正解データの出力結果の比較
(y 軸の 1 はレスト，0 は非レストを表す)

## 参考文献

[1] Justine Cassell : Embodied conversational agents. MIT press(2000).
[2] Timothy W Bickmore, Laura M Pfeifer, Donna Byron, Shaula Forsythe, Lori E Henault, Brian W Jack, Rebecca Silliman, and Michael K Paasche-Orlow: Usability of conversational agents by patients with inadequate health literacy: Evidence from two clinical trials. Journal of Health Communication 15, S2, pp.197–210(2010).
[3] RuiFang, MalcolmDoering, and JoyceYChai: Embodied collaborative referring expression generation in situated human-robot interaction. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction(HRI), pp.271-278(2015).
[4] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella: Predicting co-verbal gestures: a deep and temporal modeling approach. In International Conference on Intelligent Virtual Agents, pp.152-166(2015).
[5] Chung-Cheng Chiu, and Stacy Marsella: How to train your avatar: A data driven approach to gesture generation. In Intelligent Virtual Agents, pp.127-140(2011).
[6] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi: Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. In Proceedings of the 18th International Conference on Intelligent Virtual Agents, pp.79-86(2018).
[7] Chinmaya R. Naguri, and Razvan C. Bunescu: Recognition of Dynamic Hand Gestures from 3D Motion Data Using LSTM and CNN Architectures. In IEEE International Conference on Machine Learning and Applications, pp.1130-1133(2017).
[8] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta: Creating a gesture-speech dataset for speech-based automatic gesture generation. In Proceedings of the International Conference on Human-Computer Interaction (HCI), pp.198-202(2017).
[9] David McNeil: Hand and mind: What gestures reveal about thought. University of Chicago press(1992).
[10] Martin Abadi, et al: TensorFlow: Large-scale machine learning on heterogeneous systems(2015).
[11] François Chollet, et al.: Keras(2015), URL https://github.com/fchollet/keras.
[12] Matthew D. Zeiler: ADADELTA: an adaptive learning rate method, CoRR, vol. abs/1212.5701 (2012), URL http://arxiv.org/abs/1212.5701
[13] Elman, Jeffrey L.: Finding structure in time. Cognitive science 14.2, pp.179-211(1990).
[14] Sepp Hochreiter and Jürgen Schmidhuber: Long short-term memory.Neural computation 9.8, pp.1735-1780(1997).
[15] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation(2014). URL https://arxiv.org/abs/1406.1078v3