発話情報を考慮した頷き認識に関する研究

TP17020 中川茉耶　　指導教授：藤吉弘亘

1. はじめに

ロボットが人間の生活環境で協調して動作するには，人間とロボットのコミュニケーションが重要である．このとき，ロボットは言語表現だけでなく，頷きやジェスチャ等の非言語表現を理解することが期待されている．そこで本研究では，聞き手の頷きが話し手の発話から誘発されることに着目し，聞き手の顔器官点位置の変化と話し手の発話抑揚を同時に学習することで，聞き手の頷き認識の高精度化を図る．

2. 非言語表現の必要性

円滑なコミュニケーションには，言語表現だけでなく，頷き，まばたきなどの身体動作といった言葉によらない非言語表現が重要である [1]．これら非言語表現を含めた身体全体を介したコミュニケーションは，身体的コミュニケーションと呼ばれる．原初的コミュニケーションである乳児と母親においては，この身体的コミュニケーションが主体であり，それに基づく言語表現と非言語表現の関係形成による認知・言語発達など，人間コミュニケーションにとって普遍的であり，本質的重要性をもっていると考えられる．よって，ロボットにおいて非言語表現の理解は，人間との自然なコミュニケーションにつながる．

3. 行動認識手法

人物行動認識の研究では，Deep Neural Network(DNN) を用いた手法が多く提案されている [2]．これらの手法では，行動ラベル付き動画データを用いて DNN を学習し，未知の動画に対する行動ラベルを予測する．DNN による人物行動認識には，画像認識で使われる 2 次元の Convolutional Neural Network(2DCNN) をフレーム毎に適用して特徴量を抽出し，RNN に入力して行動ラベルを出力する手法がある．動画を入力とする人物行動認識は，大量の動画を用いて学習を行うことが予測性能の向上に繋がる．したがって，目的に合わせて認識したい行動を決定し，その行動が映る動画を大量に集める必要がある．さらに，顔に着目した画像や動画を入力とする場合は，顔の見えに関する情報を捉えることが考えられるため複数人物のデータが必要である．

4. 提案手法

本研究では，カメラから取得できる聞き手の動きに関する情報に加えて，話し手の発話を考慮した頷き認識手法を提案する．提案手法では，認識対象の顔の動きに関する情報及び，話し手の発話に関する情報を抽出する．動画像の各フレームで得たこれらの情報を LSTM に逐次入力し，各時刻の頷き認識結果を出力する．

4.1 動き情報の取得

動きに関する情報として，現時刻と前時刻の顔器官点の差分 (変化量) を用いる．顔画像を用いた場合，顔の見えに関する特徴が抽出されるため，頷き認識に必要となる動き情報だけを捉えることができない．そこで，顔器官点の動き差分を用いることで動き情報を捉え，さらに見えに関する不要な情報を除くことができる．本研究では，動画内の顔に対して Dlib[3] により全 68 点の顔器官点を検出する．顔器官点 i の時刻 t における x 方向の移動量 $d_{x,i}(t)$ と y 方向の移動量 $d_{y,i}(t)$ は式 (1) となる．ここで，顔器官点は 68 点あるため，動き情報の特徴量は $d_{x,i}(t)$ と $d_{y,i}(t)$ を合わせて 136 次元となる．

$$\begin{cases} d_{x,i}(t) &= x_i(t-5)-x_i(t) \\ d_{y,i}(t) &= y_i(t-5)-y_i(t) \end{cases} \tag{1}$$

4.2 話し手の発話情報

聞き手の頷きは話し手の発話に誘発されることから，発話情報を用いることで聞き手の頷くタイミングが理解できると考えられる．そこで，本研究では，話し手の発話情報が聞き手の頷き認識精度向上に有効と考え，入力データに用いる．発話情報として，発話有無と発話抑揚の 2 種類を検討する．発話有無は，話し手が発話しているかいないかを 2 値で表現する．また，発話抑揚は，音声波形データをもとに算出する．音声波形データは，音声データを 16bit の離散データとして記録したものである．発話の抑揚 (音声波形データの音量) は，音声波形データにおいて振幅の大きさで表現される．しかし，音声波形データは負の値もとるため，前処理として音声波形データの音声パワーを算出する．また，動画と音声データではサンプリング周波数が異なるため，時刻 t の前後 10 サンプルの音声パワーを総和した値を発話情報の特徴量 $S(t)$ として，式 (2) で求める．

$$S(t)=\sum_{i=-10}^{10} s(t+i)^2 \tag{2}$$

4.3 頷きラベル

一般的に頷きには，相手の発話に対する頷きと自分の発話に伴う頷きがある．本研究では，相手の発話に対する頷きを認識することが目的である．したがって，相手の発話に対する頷きを You，自分の発話に伴う頷きを Me と表現し，You のみを学習および評価実験に用いる．

4.4 提案手法のネットワーク

図 3(a) に発話情報のみを入力データとする場合のネットワーク構造，図 3(b) に顔器官点の変化量のみを入力データとする場合のネットワーク構造を示す．また，図 3(c) に提案手法の顔器官点の変化量と発話情報を 0〜1 に正規化して入力し，入力データ連結後に LSTM2 層に与えて頷き認識を行うネットワーク構造を示す．concatenate 後の全結合層は，入力データごとの適切な重みを学習し頷き認識精度を向上させる目的で導入する．モデルの学習には，Optimizer に RMSprop，学習回数は 300epoch，入力フレーム数は 30，バッチサイズは 32 とする．

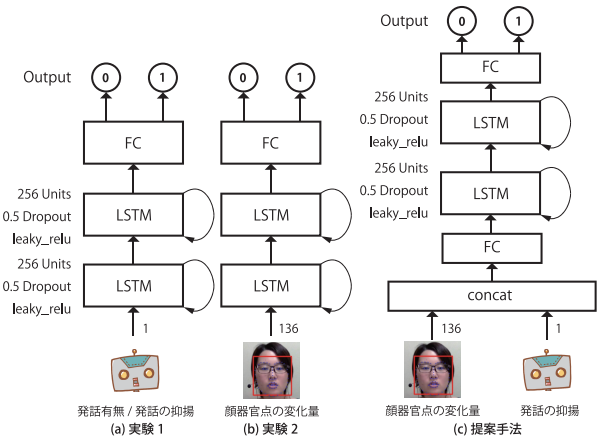

図 3：ネットワーク構成

5. 評価実験

評価実験として，3 つの実験を行う．実験 1 では話し手の発話情報として適した発話ラベルを調査するため，図 3(a) のネットワーク構造を用いて実験を行う．実験 2 では顔器官点の移動量に関する調査として，図 3(b) のネットワーク構造を用いて変化量を算出するフレーム間隔に関する実験を行う．実験 3 では発話抑揚と顔器官点の変化量を同時に学習することによる頷き認識精度向上を示すため，図 3(c) の構造を用いて実験を行う．

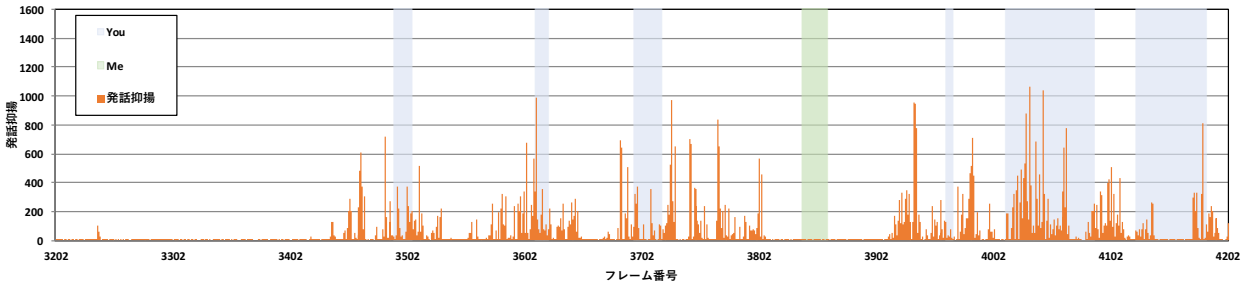

図 1： 発話抑揚と頷きの可視化

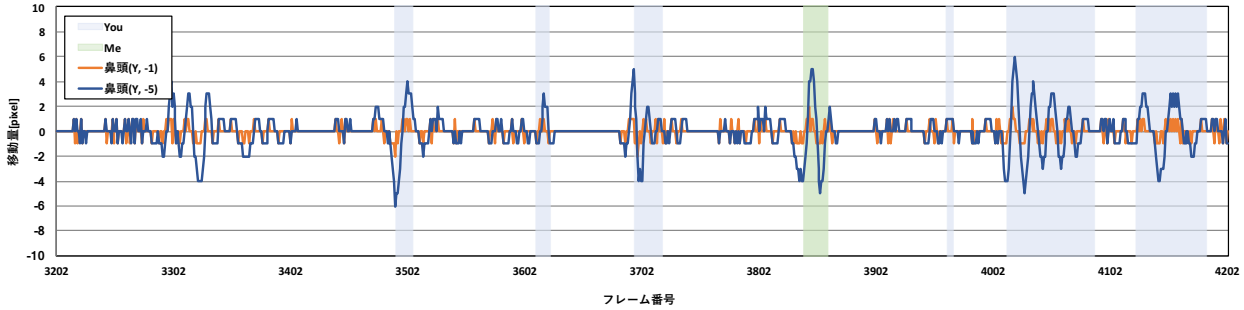

図 2： 顔器官点の変化量と頷きの可視化

6. データセットの構築

本研究では，頷きに着目したデータセットを構築し，使用する．データセットは，人間同士のテレビ電話を介した対話を約 10 分間収録した動画で構成されている．また，動画には「発話内容」,「表情」,「頷き」,「頷きの対象」などのアノテーションが付与されている．学習サンプルは頷きが 9572 フレーム，頷き以外が 9572 フレームである．評価サンプルは頷きが 2140 フレーム，頷き以外が 4402 フレームである．

6.1 実験 1： 発話抑揚と頷きの関係

図 1 に発話抑揚と頷きを可視化した例を示す．図 1 から，話し手の発話中に You が発生しやすい．また，発話抑揚は 1 文発話すると山なりになる傾向がある．これより，発話抑揚から頷き認識が可能か検討する．評価結果を表 1 に示す．表 1 より，発話抑揚を入力した場合に認識率の平均が約 63.6%となっている．発話有無を入力した場合には全く認識できなかった Nod を，発話抑揚を入力した場合では約 42.0%認識率できている．これより，話し手の発話抑揚が聞き手の頷き認識に重要であることがわかる．

表 1： 発話ラベルによる認識率 [%]

発話ラベル	Nod(頷き)	Other	平均
発話有無 (発話区間全域に付与)	0.0	100.0	50.0
発話抑揚	42.0	85.1	63.6

6.2 実験 2： 顔器官点の変化量と頷きの関係

頷き認識の精度向上を目的として発話抑揚とともに顔器官点の変化量を入力する．顔器官点の変化量は，現時点の顔器官点と数フレーム前の顔器官点の差とする．まず，顔の移動量に関する調査として，顔器官点の変化量を算出するフレーム間隔を変更し実験する．図 2 に顔器官点の変化量と頷きの可視化例を示す．評価結果を表 2 に示す．表 2 より，5 フレーム前との変化量を入力した場合に最も精度が良いとき約 83.1%の認識率となっている．しかし，1 フレーム前との変化量を入力した場合は 5 フレーム前との変化量を入力した場合と比較すると，Nod の認識率が向上し，Other の認識率が低下したことが平均の認識精度向上につながっている．実験 3 では，顔器官点の変化量を算出するフレーム間隔は 5 フレームで実験を行う．

表 2： 顔器官点の変化量による認識率 [%]

変化量	Nod(頷き)	Other	平均
1 フレーム前	81.7	84.4	83.1
5 フレーム前	72.6	91.7	82.2

6.3 実験 3： 顔器官点の変化量と発話抑揚の同時学習の効果

発話抑揚と顔器官点の変化量を同時に学習することで頷き認識の精度向上が期待できる．評価結果を表 3 に示す．表 3 より，顔器官点の変化量のみを学習した場合と顔器官点の変化量と発話抑揚を同時に学習した場合に最も精度が良く約 84.4%の認識率である．また，表 3 に示すように，同時学習の場合は顔器官点変化量のみを入力した場合と比較し，Other の認識率が約 4.3%低下している．一方で Nod の認識率は約 8.8%向上している．このことから，発話抑揚を同時に学習することは頷き認識に有効であると言える．

表 3： 入力情報による認識率 [%]

入力情報	Nod(頷き)	Other	平均
発話抑揚のみ	42.0	85.1	63.6
顔器官点変化量のみ	72.6	91.7	82.2
発話抑揚&顔器官点変化量	81.4	87.4	84.4

7. おわりに

本研究では，話し手の発話抑揚と聞き手の顔器官点の変化量から頷きを認識する手法を提案した．発話抑揚と顔器官点の変化量を同時に学習することで，顔器官点のみを学習した場合と比較し頷き認識精度を 8.8%向上させることができた．今後の展望として，複数人物への対応や 3DCNN の認識結果を統合する End-to-end のネットワーク，頷き以外の非言語表現の認識が挙げられる．

参考文献

[1] 渡辺富夫, “ コミュニケーションにおける引き込みと身体性 ”, Neonatal Care, 1999.

[2] K. Simonyan, *et al*, “ Two-Stream Convolutional Networks for Action Recognition in Videos ”, NIPS, 2014.

[3] D. E.King, “ Dlib-ml: A machine learning toolkit ”, Journal of Machine Learning Research, 2009.

研究業績

[1] 中川茉耶 等, “ Deep Convolutional Neural Network による認知症リスクの推定 ”, ViEW, 2016.