オンライン複数人会話における参加者の理解度の推定

酒造 正樹[1]　高木 章裕[1]　湯浅 将英[2]　酒井 元気[3]

概要：オンライン上の複数人会話において，映像表示領域の制約から参加者の反応として，他者がどの程度の理解を得ているのか，またどのように感じているのかなどを把握することが対面の場合に比べて困難である．本論文では相手の画像情報から得られる非言語情報をもとに理解度の推定方法について検討を行った．OpenFace を用いて頭部の運動情報，視線，表情の特徴量を抽出し，random forest モデルによる推定結果を交差検証法（leave-one-session-out 法）により確認した．

Estimation of Participants' Comprehension Level in Video Group Communication

Masaki Shuzo[1]　Akihiro Takagi[1]　Masahide Yuasa[2]　Motoki Sakai[3]

1. はじめに

オンラインでの会議が増加した今日，対面では容易に取得できていた他者の状態が伝わりにくくなり，円滑なコミュニケーションが難しくなってきた．本論文において，オンライン上の複数人会話における参加者の反応の把握について議論する．自らの発話に対して，他者はどの程度の理解を得ているのか，そしてどのように感じているかについて知ることは，オンライン会議のみならず対面においても有用である．

相手の反応を得るためには，言語情報のみならず，対面コミュニケーションにおいては様々な非言語情報が役に立っている．相槌，頷き，表情，視線，目の動き，話者に対するの顔向き・距離，体の姿勢・動作などはコミュニケーション研究においてよく議論される情報である．Zoom などに代表されるビデオ会議システムにおいては，顔を中心としたカメラアングルで上半身の一部のみがディスプレイに表示される．よって，必然的に，上述した非言語情報のうち相手の物理的な位置関係に関しては，実際の情報と異なるものが表示されてしまう．

[1] 東京電機大学
Tokyo Denki University
[2] 湘南工科大学
Shonan Institute of Technology
[3] 日本大学
Nihon University

一般に，クライアントソフト上で，表示サイズや位置の調整を行うことができる．本論文においては，全画面モードで全ての参加者の映像を均等サイズに表示するレイアウト（ギャラリービュー）を前提としている．また，12 インチ（小型ノートパソコン）から 24 インチ程度（卓上モニタ）のディスプレイを想定しており，参加者の数が多くなれば，1 人あたりの表示領域が狭くなり，個々の非言語情報の適切な取得が困難になるので，参加者数を 4（2 人× 2 段）または 9（3 人× 3 段）を上限として扱う．

これまで著者らの研究グループでは，良いグループディスカッションのあり方を求め，話し手視点の関連研究と異なり，聞き手の立場の状態推定の研究を行ってきた．本論文では，画像やセンサ情報をもとにした聞き手の感情の研究 [1] に続き，理解度の推定を目的としている．

2. 関連研究

酒井らは，対面で行われた GD 中の話し手の発話に対し，聞き手の感情の状態を推定することでコミュニケーションなどの改善が期待できると考え，加速度計，心電図，筋電図より得られた全 56 個の特徴量を統計的手法（多重比較等）を用いて Ekman の基本 6 感情 [2] における有意差を評価した．また，肯定的または否定的な感情状態の 2 値分類を，support vector machine を用いて行ったところ 76% の精度が得られたと報告がある [3]．また，酒井ら [1] や鳥

羽ら [4] は，オンラインで行われた GD に対する感情推定の研究について，報告している．

平野らは社会的信号処理に基づき感情に注目しユーザの関心，感情，トピックの継続ラベルを発話テキスト，顔の表情，身体及び頭部の動き，音響から得られる特徴量より推定を行った．結果として，マルチモーダルな情報で構成されたディープニューラルネットワークモデルを用いることで推定精度が向上できたと報告がある [5]．河原らはポスターセッション時の聴衆のマルチモーダル情報（目線，相槌）より興味及び理解度の推定を行った．マルチモーダル情報を含めることで，ポスターセッション時の理解度についてナイーブベイズ分類器で 75% の精度が得られたと報告がある [6]．

3. グループディスカッション実験

GD 中の話し手の発言における聞き手の理解度を個人ごとに推定することを目的とした．インターネットサイトのブリタニカの ProCon[*1]から「学生は制服を着なければなりませんか？」「宿題をさせることは有益ですか？」などの 10 個のテーマを選定し，zoom を利用して 15 分の議論を実施した．本実験の参加者は東京電機大学の同一研究室の所属の 4 年生 5 人である．各回の参加は任意とし，人数は 3 または 4 人とした．映像収録の実験詳細は既報 [7] のとおりである．

4. 理解度のアノテーション

収録済みの映像に対して，約 6 ヶ月後に参加者が自分以外の参加者の発言内容に対する理解度のアノテーションを行った．なお，収録に参加した 5 名のうち 1 名はアノテーションに参加していない．参加者の顔が一覧できるギャラリービュー形式で映像を振り返り，議論参加中の自分として評価を行った（図 1）．アノテーションは，他者の発言ごとに 7 段階（1: 非常に理解できない，4: どちらでもない，7: 非常に理解できる）で行った．

入力してもらった理解度アノテーションデータの理解度スコアについて，5〜7 を High（理解度が高い），1〜4 を Low（理解度が低い，またはどちらでもない）と定義し，2 値の理解度ラベルデータを作成した．全体のアノテーション数とセッションごとの High/Low のバランスを，ぞれぞれ表 1，図 2 に示す．なお，表中の括弧内のアルファベット記号は [7] における参加者 ID を示す．

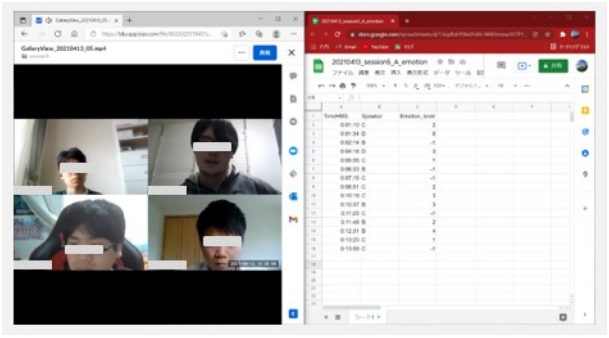

図 1 アノテーション作業の様子

表 1 アノテーション数

Name	Total number	Average
U1(E)	201	20.1
U2(C)	153	19.1
U3(D)	228	28.5
U4(A)	244	40.7

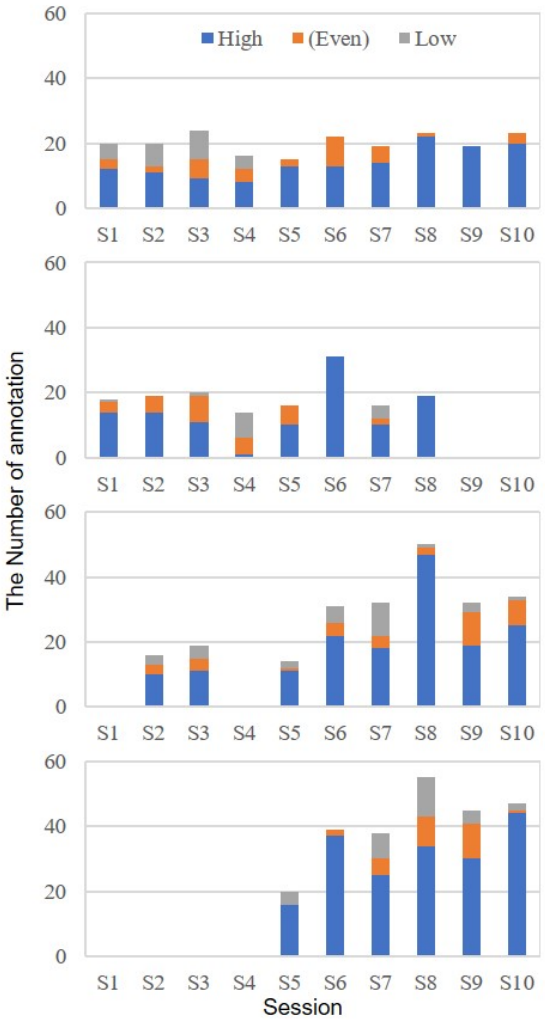

図 2 セッションごとの High/Low のバランスの分布

[*1] https://www.procon.org/

表 2 採用した 17 種の AU

Index	Description
AU01	眉の内側を上げる
AU02	眉の外側を上げる
AU04	眉を上げる
AU05	上瞼を上げる
AU06	頬を持ち上げる
AU07	瞼を緊張させる
AU09	鼻に皺を寄せる
AU10	上唇を上げる
AU12	唇の両端を引き上げる
AU14	笑窪を作る
AU15	唇の両端を下げる
AU17	顎を上げる
AU20	唇の両端を横に引く
AU23	唇を固く閉じる
AU25	顎を下げずに唇を開く
AU26	顎を下げて唇を開く
AU45	瞬きをする

表 3 参加者ごとの理解度推定結果の F1-score

Name	Session →	S1	S2	S3	S4	S5
U1(E)	High	0.64	0.29	0.19	0.47	-
	Low	0.40	0.58	0.37	0.40	-
U2(C)	High	0.88	0.79	0.40	0.18	0.70
	Low	0.67	0.40	0.40	0.47	0.22
U3(D)	High	x	0.67	0.35	-	0.31
	Low	x	0.36	0.48	-	0.40
U4(A)	High	x	x	x	x	0.48
	Low	x	x	x	x	0.13

Name	Session →	S6	S7	S8	S9	S10
U1(E)	High	0.62	0.76	-	-	0.50
	Low	0.27	0.22	-	-	0.22
U2(C)	High	-	0.56	-	x	x
	Low	-	0.43	-	x	x
U3(D)	High	-	0.63	-	0.39	0.47
	Low	-	0.32	-	0.21	0.33
U4(A)	High	-	0.60	0.52	0.46	-
	Low	-	0.34	0.38	0.38	-

5. 理解度推定方法

個人に対応する理解度推定モデルを random forest により構築した．十分量のアノテーションデータを得られなかったことと，データリークを防ぐために，参加したセッションから 1 セッションをテストデータとし，それ以外を学習データとした（leave-one-session-out 法）．

特徴量の抽出については，個人ごとに分割した動画を入力データとして，OpenFace[8] を用いて，頭部の運動情報，視線情報，表情を定義するアクションユニット（AU）[9] 情報を得た．頭部情報からパソコン搭載のカメラから見て横方向を x 軸, 縦方向を y 軸，奥行き方向を z 軸として頭部の x 軸，y 軸，z 軸方向の回転角度（pose_Rx，pose_Ry，pose_Rz）の中央値，分散，10 パーセントタイル値，90 パーセントタイル値を得た．視線情報からパソコン搭載のカメラから見て横方向を x 軸，縦軸を y 軸とし，視線の x 軸，y 軸方向の角度（gaze_angle_x，gaze_angle_y）の中央値，分散，10 パーセントタイル値，90 パーセントタイル値を得た．表情に関しては，表 2 に示す 17 種の AU を採用し，各 AU の平均値，分散，標準偏差，最大値，最小値を得た．以上の全 105 次元の特徴量について，発話区間に対して算出した．

6. 評価結果と考察

アノテーションの協力を得た参加者 4 人（U1 ～ 4）の推定結果の F1-score を表 3 に示す．欠席したセッション，十分量のアノテーション量がなかったセッションについては，それぞれ記号 ‘x’，‘-’ で記した．

次に，全 10 セッションに参加した参加者 U1 の理解度推定結果について述べる．セッションにもよるが概ね High の方が F1-score が高い結果となった．全体的には低い結果となったが，推定にはどのような特徴量が有効的であったのかを確認するため，良好な結果を示していた S1, S6, S7, S10 について特徴量の寄与度の検討を行った．寄与度の高い上位 3 件について，結果を表 4 に示す．

表 4 特徴量の寄与率の上位 3 件

セッション	順位	特徴量
S1	1	gaze_angle_x_var
	2	gaze_angle_y_percentile90
	3	AU20_r_max
S6	1	gaze_angle_y_percentile10
	2	gaze_angle_y_percentile90
	3	AU14_r_mean
S7	1	gaze_angle_y_percentile10
	2	pose_Rx_median
	3	AU45_r_std
S10	1	gaze_angle_y_percentile10
	2	pose_Rx_percentile10
	3	AU45_r_var

この結果，各セッションにおいて視線における特徴量（頭文字 gaze で始まるもの）の寄与度が高くなっている．このことから参加者 U1 が理解度に応じて目線の行動が多く出ていたことが分かる．また，全ての場合において表情の特徴量（頭文字 AU で始まるもの）含まれている．このうち AU45 については S7 と S10 で共通している．このことから瞬きは理解度と関係があるのではないかと考える．以上より理解度推定において，視線や表情のデータの有効性が示唆できた．

7. おわりに

本論文では，オンライン上の複数人会話における参加者の反応として，他者がどの程度の理解を得ているのかの推定を試みた．収録済みの映像データに対して，当時の自らの立場による理解度のアノテーションを実施し，2 値化の理解度データセットを得た．個人ごとの理解度の推定を random forest モデルにより構築し，leave-one-session-out 法により評価し，比較的良好なセッションに対する寄与度の調査を行った．

AI/IoT を活用したグループコミュニケーションにおけるパフォーマンス評価に関して，話者が時系列に入れ替わる場のデータ収録を継続実施中である．このような研究は，個人差が大きく結果に左右するため，試行回数に応じた長時間の拘束や，多くの被験者を要するなどの懸念材料がある．今後さらにデータを増強するためにも，データセットを再利用可能とする研究の枠組みを検討する必要がある．

謝辞

本研究の一部は，JSPS 科研費 JP19H01719 の助成を受けた．

参考文献

[1] 酒井他, “オンライン対話上における聞き手の感情状態の推定,” HCG シンポジウム 2021, I-2-5, 2021

[2] P. Ekman. “Basic Emotions,” in T. Dalgleish and M. J. Power (Eds.), Handbook of cognition and emotion, John Wiley & Sons Ltd., pp. 45–60, 1999.

[3] M. Sakai *et al.*, “Biological and behavioral information-based method of predicting listener emotions toward speaker utterances during group discussion,” in M. Ahad *et al.* (Eds.), Activity and behavior computing, Springer, pp. 189–207, 2021.

[4] 鳥羽他, “オンラインミーティング参加者の感情推定における機械学習モデルと客観的評価との比較,” DICOMO2022, 2022 発表予定.

[5] Y. Hirano *et al.*, “Multitask prediction of exchange-level annotations for multimodal dialogue systems,” in Proc. 2019 International Conference on Multimodal Interaction, pp. 85–94, 2019.

[6] 河原他, “ポスター会話における聴衆のマルチモーダルな振る舞いに基づく興味・理解度の推定,” 情報処理学会研究報告 SLP-97-12, pp. 1–5, 2013.

[7] 酒造他, “グループコミュニケーションの能力向上に向けた長期的な学生間の相互評価の検討,” DICOMO2021, pp. 949–952, 2021.

[8] T. Baltrušaitis, P. Robinson and L. Morency, “OpenFace: An open source facial behavior analysis toolkit,” in Proc. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10, 2016.

[9] J. F. Cohn, Z. Ambadar and P. Ekman, “Observer-based measurement of facial expression with the facial action coding system,” in J. A. Coan & J. J. B. Allen (Eds.), Handbook of emotion elicitation and assessment, pp. 203–221, Oxford University Press, 2007.