Vol.2022-ICS-206 No.2
2022/3/10

マルチモーダルデータを用いたオンラインミーティング参加者の感情推定

鳥羽望海[1]　藤本まなと[2]　諏訪 博彦[1,3]　酒井 元気[4]　酒造 正樹[5]　安本 慶一[1]

概要：労働者の心理状態を知ることは，労働者の不調を上司や産業医が事前に把握し，健全な組織運営をすることに貢献する．一方，昨今の COVID-19 の影響で，労働者がテレワークが取り入れられるようになり，上司や産業医にとってテレワーク中の労働者の心理状態は把握しにくい．我々は費用と労力，感染リスクを減らせるメリットがあるオンラインミーティングサービスに着目し，オンラインミーティングで得られる動画や音声などのマルチモーダルデータを用いて労働者の感情を推定する手法を検討する．オンラインミーティングサービスを用いてグループディスカッションを被験者にしてもらい，そこから映像，心拍など様々なデータをセンシングして参加者の状態を推定する．グループディスカッション中の参加者の心拍，発言録から得られた感情極性，顔のランドマーク座標，他者からの感情アノテーションの 4 種類の指標から，他者への感情アノテーションを推定するために Light GBM (Gradient Boosting Machine)，SVR (Support Vector Regression) の解析を行った結果，平均絶対誤差に関して 1.780 という結果を算出した SVR のほうが良いモデルであることがわかった．

Estimating Emotions of Participants in Online Meeting using Multimodal Data

1. はじめに

2020 年から現在まで，新型コロナウイルス感染症（COVID-19）のパンデミックにより，世界中で外出制限の措置がなされ，仕事におけるオンラインミーティングサービスの利用が急速に広まった．Zoom や Microsoft Teams などのオンラインミーティングサービスの 1 日あたりの参加者数は，2019 年 12 月から 2020 年 4 月で，それぞれ 1,000 万人から 3 億人，2,000 万人から 7,500 万人にまで増加した[*1]．

オンラインミーティングサービスは，費用と労力，感染リスクを減らすことを期待できる一方，オンサイトでのミーティングと比較して，利用者の表情や感情の把握は難しく，ミーティングの質を低下させる恐れがある．これによりオンラインミーティングが非効率的なものになることが危惧されている．また，オンラインミーティングでは，カメラとディスプレイの位置がずれており，アイコンタクトが困難になることや，インターネット回線遅延の影響で相手との会話に時間差が発生することがある．そのため，会話の不一致が発生し，情報伝達がより困難になる可能性がある．

オンラインミーティングにおける参加者の効率性に関連する研究として，Samrose ら [1] は，オンラインミーティングに対する参加者の意識を向上させる参加者ごとにパーソナライズされたダッシュボードの構築を提案している．Cutler ら [2] は，ミーティングの有効性とミーティングの包括性を調査するため，COVID-19 以前に大規模なテクノロジー企業で大規模な電子メール調査を実施し，オンラインミーティングの有効性の多変量モデルを導き出し，電子メールと参加者のミーティングへの貢献度の相関性を示している．ミーティングの効果を向上させることで，企業は時間と費用を大幅に節約でき，組織での作業環境と従業員の定着率も改善すると仮定している．しかし，これらの既存研究では，視線や心拍に関する解析を行っておらず，高精度な感情推定やノンバーバルコミュニケーションの情報

[1] 奈良先端科学技術大学院大学，Nara Institute of Science and Technology
[2] 大阪市立大学，Osaka City University
[3] 理化学研究所, RIKEN
[4] 日本大学， Nihon University
[5] 東京電機大学， Tokyo Denki University

[*1] https://blog.zoom.us/a-message-to-our-users/

を加味できていない．さらに，アノテーションがなく，客観性が不足している．

本研究では，他者の発言に対するアノテーション付きグループディスカッションデータ，発言に関する感情極性データ，グループディスカッション中の参加者の心拍データ，顔のランドマーク座標データといったマルチモーダルなデータを用いて，複数の機械学習アルゴリズムにより感情推定モデルを構築し，その結果を比較検討する．

オンラインミーティングサービスを用いてグループディスカッションを参加者にしてもらい，そこから映像，心拍など様々なデータをセンシングして参加者の他者への感情アノテーションを推定する．グループディスカッション中の参加者の心拍，発言録から得られた感情極性，顔のランドマーク座標，他者からの感情アノテーションの 4 種類の指標を用いて，Light GBM[3]，SVR[4] によるモデル構築を行った．その結果，平均絶対誤差に関して 1.780 という結果を算出した SVR のほうが良いモデルであることを明らかにした．

本研究の結果は，オンラインミーティングに参加する労働者の感情の動きから Workaholic[5] や Burnout[6], [7], [8] を事前に把握し，労働者の健全な管理に貢献する可能性がある．

本論文の構成は，以下の通りである．2 章で関連研究について述べ，3 章で対象とするグループディスカッションデータについて述べる．4 章でグループディスカッション参加者の感情推定システムについて提案する．5 章で実験について説明し，6 章で解析結果を示す．7 章で結論をまとめる．

2. 関連研究

本章では，オンラインミーティングにおける参加者へのフィードバック手法として MeetingCoach を紹介し，その機能と課題を述べる．その上で，課題を解決するための技術として，表現分析手法，音声データ処理手法，心拍を用いた感情推定手法について述べる．

2.1 参加者へのフィードバック

オンラインミーティングにおける参加者へのフィードバックをする手法として，Samrose ら [1] は，参加者ごとにパーソナライズされたダッシュボードである MeetingCoach の構築を提案している．MeetingCoach は，参加者にコンテキストおよび行動のミーティング情報を要約して提供する．また，話者交替，感情などの信号を識別する．MeetingCoach の設計は，実際のオンラインミーティングの参加者によって評価している． 彼らは，4 週間にわたって 8 チームとの電話ミーティングを記録するためのツールを構築した．ダッシュボードを作成することにより，参加者の参加意欲の向上に役立つことを示している．しかしながら，(1) 心拍や視線に関する解析を行っいないこと，(2) アノテーションがなく，客観性が足りないことが課題として挙げられる．

2.2 表情分析手法

表情を分析する手法として，Baltrušaitis ら [9] は，OpenFace を提案している．OpenFace は，顔のランドマーク検出，頭部のポーズ推定，顔の Action Unit の認識，視線の推定が可能なオープンソースツールである．OpenFace はリアルタイム性能を備えており，特別なハードウェアを必要とせず，単純なウェブカメラから実行することが可能である．顔のランドマークの検出と追跡には Conditional Local Neural Fields (CLNF)[10] を使用している．彼らが提案した CLNF モデルでは，顔にある 68 個のランドマークをすべてまとめて検出する．また，顔の Action Unit の認識は，[11], [12] のフレームワークに基づいている．実験では 100 ピクセル以上の顔画像に対して適切な結果が得られることが示されている．

2.3 音声からの話者交代推定手法

音声処理アルゴリズムに関して，Tashev[13] は音声と雑音の異なる確率密度関数に基づいたオフライン音声活動検出器 (Voice Activity Detector: VAD) を提案している．提案された VAD アルゴリズムは，周波数領域で動作し，各オーディオフレームの各周波数に対する音声信号の存在確率と，各フレームの音声存在確率を推定し，さらにフレームごとに二値判定を行っている．これによりグループディスカッション参加者の話者交代を検出することが期待される．

また，音声データからテキストデータを手軽に抽出する技術が開発されている．具体的なサービスとして，Google Speech-to-Text[*2]などがあげられる．これらの技術を用いることにより，誰がどんな発言をしているのかを抽出することが可能となる．

2.4 心拍からの感情推定手法

Harper ら [14] は，単峰性心拍時系列から感情の価値を分類する End-to-end の深層学習モデルを提案している．また，これらの価値予測に対する不確実性をモデル化するためのベイジアンフレームワークを提案している．これらの結果は，非侵襲的なデータ収集と予測の確実性が重視されるヘルスケアなどの実世界の領域における感情コンピューティングの応用の基礎となると述べている．この技術を用いることで，生体データから感情を推定することが可能となる．

[*2] https://cloud.google.com/speech-to-text

3. 対象とするグループディスカッションデータ

本章では，対象とするグループディスカッションデータ（日本大学酒井元気らのグループが提供）について説明する．

このデータセットには，(1) オンラインミーティングサービスを用いた参加者 3 から 4 人が写ったグループディスカッションのギャラリー映像，(2) 発話者のみが写った映像，(3) 各参加者の音声，(4) 音声を用いて発言をテキスト化するサービス Google Speech-to-Text により生成した発言録，(5) 発言録から取得された感情極性データ [15]，(6) ウェアラブルデバイスから取得された各参加者の心拍データ，(7) 他者の発言に対するアノテーションが含まれる．それぞれの詳細は以下の通りである．

(1) ギャラリー映像
オンラインミーティングサービスの表示におけるギャラリーモードを用いて実現している．グループディスカッション実験に参加している参加者全員が表示される映像を MP4 形式で保存している．2021 年 1 月 21 日 13 時 50 分から 14 時 05 分までに行われたオンラインミーティングサービスを用いたグループディスカッションのギャラリー映像から抽出した 1 フレームの例を図 1 に示す．

(2) 発話者のみが写った映像
オンラインミーティングサービスの表示におけるスピーカーモードを用いて実現している．グループディスカッション実験に参加している参加者が発言した際に画面全体に表示される映像を MP4 形式で保存している．

(3) 各参加者の音声
オンラインミーティングサービスの「参加者ごとに個別のオーディオファイルで録音」モードを用いて実現している．グループディスカッション実験に参加している参加者の音声を M4A 形式で保存している．上記グループディスカッションのうち参加者 A の音声の例を図 2 に示す．

(4) 発言録
Google Speech-to-Text により，録音した各参加者の音声をテキスト変換することで実現している．CSV 形式で保存している．カラムインデックスは時間，発話者，発話内容となっており時系列で保存している．

(5) 感情極性データ
意見や態度を含んだ感情をテキストから分析するために重要なリソースとして挙げられる単語の感情極性を用いる [15]． Google Speech-to-Text によりテキストに変換された参加者の発言内容を MeCab[16] により形態素解析を実行したのちに，各単語に関して感情極性データを対応させ，それぞれの文に対して平均を取った．

図 1 映像データ

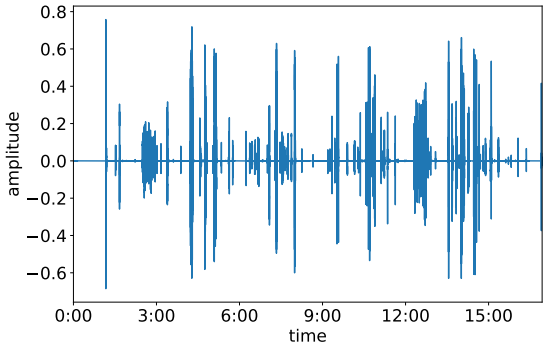

図 2 音声データ

(6) 心拍データ
各参加者にウェアラブルデバイス Fitbit Ionic®[*3]を装着してもらい，そこから心拍を取得することで実現している．Fitbit SDK における Web API を用いて心拍データを抽出し，JSON 形式を変換して CSV 形式で保存している．上記グループディスカッションのうち参加者 A の心拍波形の例を図 3 に示す．

(7) 他者の発言に対するアノテーション
オンラインミーティングサービスによるグループディスカッション収録後に，発話者のみが写った映像を用いて，各参加者がグループディスカッションに同席していた他の参加者の発言に対して，−4 点から 4 点で評価してもらう．悪ければ −4 点，良ければ 4 点の評価をつけることで CSV 形式で実現している．上記グループディスカッションのうち参加者 A による，他者の発言に対するアノテーションの例を表 1 に示す．

4. 感情推定システム

本章では，グループディスカッション参加者の感情推定

[*3] https://www.fitbit.com/

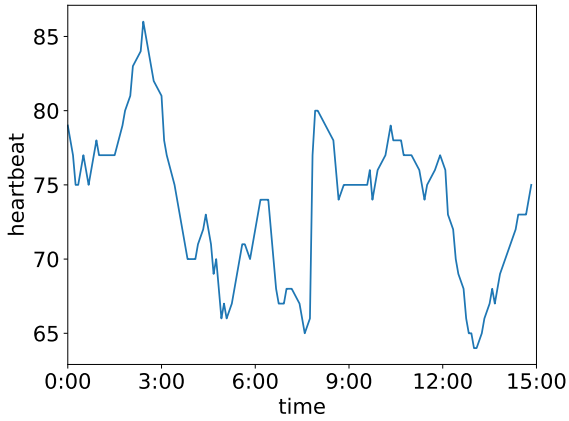

図 3　心拍データ

表 1　参加者 C の感情アノテーションを一例として示す

経過時間	感情ラベル	評価対象の参加者
2:00	3	B
2:19	2	A
4:29	3	A
4:50	1	B
5:45	2	D
6:30	3	B
7:28	3	A
7:50	-1	B
8:48	2	B
9:16	3	D
10:07	4	B
10:26	-2	A
10:55	4	B
12:05	3	B
12:26	3	A
13:16	2	B
13:30	0	D
15:27	3	B

システムについて提案する．提案システムでは，オンラインミーティングサービスやウェアラブルデバイス，グループディスカッション参加者が他者に対して行ったアノテーションから得られたデータから参加者の感情推定を行う．

本システムは，図 4 のような 3 つのステップで構成されている．(1) オンラインミーティングサービスとウェアラブルデバイスを用いてグループディスカッションを実行・記録するステップ，(2) グループディスカッション参加者に他者の発言に対するアノテーションを記録された音声・映像をもとに行なってもらうステップ，(3) 音声・映像・心拍・テキスト解析をすることでマルチモーダルにグループディスカッション参加者の感情を推定するステップである．

我々は，本研究の最終目標として，オンラインミーティングサービスやウェアラブルデバイス，グループディスカッション参加者が他者に対して行ったアノテーションを用いて，自動的に参加者の感情を推定するシステムの実現を目指している．本稿では，その前段階として既存データの分析実験を行い，提案システム実現の可能性を探る．

5. 既存データの分析実験

本章では，提案システムの実現可能性を探るために，既存のグループディスカッションデータからの特徴量の抽出とその特徴量を用いた感情推定モデルの構築を試みる．分析対象は，2021 年 1 月 21 日 13 時 40 分から 14 時 05 分までの 15 分間に行われたオンラインミーティングサービスを用いたグループディスカッションデータである．

5.1 特徴量の抽出

音声データからテキスト抽出をするために，Google Speech-to-Text を用いて，個別に保存されていた参加者 4 人の音声データをテキスト化し，発言した時間に合わせて統合を行った．参加者 A の音声データを図 2 に示す．図 2 より，発話のタイミングが確認することができる．また，この音声データをフーリエ変換するなどすることにより，声質などを取得できるようになると考える．また，4 人の参加者がグループディスカッションをした内容を実際に統合して生成したテキストデータを表 2 に示す．テキストデータを確認することで，与えられたテーマに沿った発言をしているかどうかを判断することや，ポジティブ・ネガティブ分析が可能となり，感情アノテーションと統合して発言の適切性を確認することが可能となる．

映像データから表情を検出するために，FaceLandmarkVidMulti を用いて，複数人の顔のランドマーク検出を行った．4 人の参加者の映像（図 1）を実際に OpenFace を用いて顔のランドマーク検出を行って生成した映像を図 5 に示す．これにより，顔の向きから視線の位置が，ランドマークから表情の検出が可能となっている．

以上の二つの特徴量に加え，心拍データと他者からの感情アノテーションを加えた 4 つを特徴量とする．

5.2 感情推定モデルの構築

グループディスカッション中の発言録から得られた感情極性，顔のランドマーク座標，参加者の心拍，他者からの感情アノテーションの 4 指標から，他者への感情アノテーションを推定するために Light GBM[3] , SVR[4] を用いて感情推定モデルの構築を試みる．モデルの評価は，平均絶対誤差 (Mean Absolute Error: MAE) で行う．MAE は以下の式で表される．

$$M = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|$$

式において，M が MAE，n がフレーム数，F_t がモデルが推定した感情ラベル，A_t が参加者がアノテーションした感情ラベルを表す．また，個人ごとのモデルを検証するた

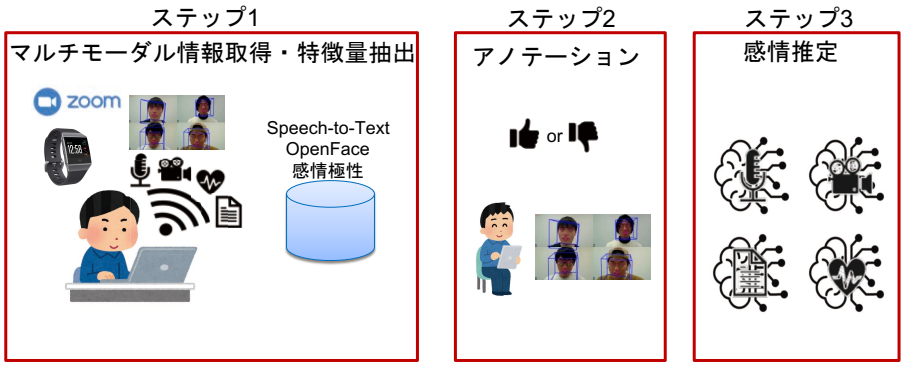

図 4　実験のステップ

表 2　発言録データの一部

時間	話者	内容
2021-01-21 13:53:02	B	まあいいかもしれん勝手に決めたのもありだし親子間でも言うことかわからんけどママ縛りすぎても逆に反発して言うこときかへんくなってなんか気になってもあるし
2021-01-21 13:53:07	B	ちょうどいいルール
2021-01-21 13:53:13	C	じゃいいですか
2021-01-21 13:53:17	A	どうぞ
2021-01-21 13:53:34	C	A さん B さんの意見もすごくいいんですけどなんか僕がスマホを持ち始めた高校の頃は確かにすごく依存してたかもしれない学生になって果たしてそこまでなんか依存するかって言われたらなんか別に
2021-01-21 13:53:50	C	何か使い方分かってきて若干その時間配分とかも決められるようになってるからは小学生がそれができるかっていうかわからへんけどある程度そのスマホに慣れさせるっていう状況を作るっていうのもありなのかなっていう
2021-01-21 13:54:11	C	書いてないから時間がかかるって言うのであってその使い慣れてないものを使い慣れさせるために早いうちから回ってるって言うのはあながちわんちゃん行けるかな
2021-01-21 13:54:15	B	結局ゲームとか LINE だけやから

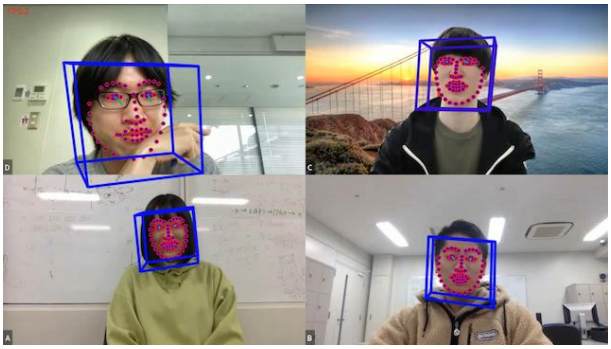

図 5　OpenFace で映像データ解析した結果

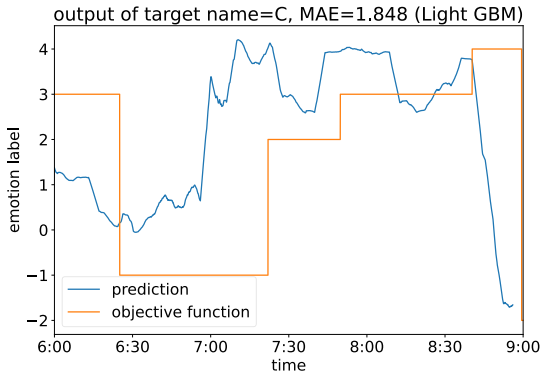

図 6　マルチモーダルなデータを用いた Light GBM による参加者 C の開始 6 分から 9 分までの感情の推定結果

めに，15 分で 25 フレーム毎秒のデータ（全 22500 フレーム，縦 720 ピクセル，横 1280 ピクセル）を 3 分ごとに 5 分割し（開始 0 分から 3 分まで，3 分から 6 分まで，6 分から 9 分まで，9 分から 12 分まで，12 分から終了まで），交差検証をかけた．

6.　分析結果

Light GBM により参加者 C の開始 6 分から 9 分までの実際にアノテーションした感情ラベルと推定した結果を図 6 に示す．縦軸が −4 から 4 点までの感情ラベルの値，横軸が時間，オレンジ線で描かれた矩形波が実際に参加者がアノテーションした感情ラベル，青線で描かれた波形が Light GBM が推定した参加者の感情ラベルを表している．オレンジ線が青線の形状に近ければ近いほど良い推定をし

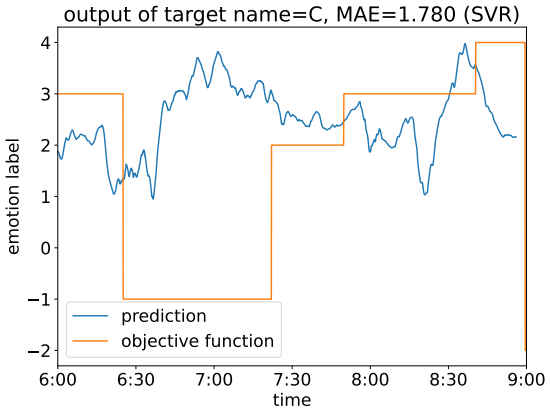

図 7　マルチモーダルなデータを用いた SVR による参加者 C の開始 6 分から 9 分までの感情の推定結果

ているということが言える．MAE で 1.848 を取っていることから，推定した感情ラベルは実際にアノテーションしたものからの乖離が 2 未満であることがわかる．推定した感情ラベルと実際にアノテーションしたものにはおよそ 2 点ほどのずれが生じているものの，おおむね高評価，低評価をしている感情ラベルを推定できている．

一方で 7 分，8 分 30 秒目において，推定した感情ラベルの立ち上がり，立ち下がりが実際にアノテーションしたものと比較して時間的にずれが生じている．まず，7 分目の立ち上がりについて，実際のデータを確認したところ，参加者 B のポジティブな発言に対して参加者 C は「確かに」といった発言やうなずく動作が見られた．このことから Light GBM はアノテーションよりも前に肯定的な推定をしているのではということが示唆される．次に，8 分 30 秒目の立ち下がりについて，同様に実際のデータを確認したところ，16 秒間にわたってグループディスカッション内における発言をする参加者がいなかった．このことから，Light GBM で適切に感情推定ができなかったことが示唆される．これらのずれの問題は，今回行ったグループディスカッションのデータは時系列的である一方で，Light GBM 自体は時系列データを処理するモデルではないことから生じていると考えられる．

次に，SVR により参加者 C の開始 6 分から 9 分までの実際にアノテーションした感情ラベルと推定した結果を図 7 に示す．縦軸が −4 から 4 点までの感情ラベルの値，横軸が時間，オレンジ線で描かれた矩形波が実際に参加者がアノテーションした感情ラベル，青線で描かれた波形が SVR が推定した参加者の感情ラベルを表している．MAE で 1.780 を取っており，これは Light GBM と比較するとより良い結果である．後半部分の感情ラベルの推定はおおむね実際にアノテーションしたものに近いことが図からわかり，適切に推定できていることが示唆される．

次にマルチモーダルなデータと独立した 1 つずつのデータで解析結果との比較をする．顔のランドマーク座標データのみを用いた結果 (図 8，MAE は 1.829)，他者からのアノテーションデータのみを用いた結果 (図 9，MAE は 1.841)，心拍データのみを用いた結果 (図 10，MAE は 1.272) を示す．いずれの場合も MAE はマルチモーダルなデータで解析した場合よりも良い結果を出した．一方で，感情ラベルの目的関数と推定結果のグラフを目視により比較すると，図 6 のほうが時間的には予測のずれがあるものの良い結果に見える．本研究のようなグループディスカッションのような時系列データによる解析を評価するために MAE は十

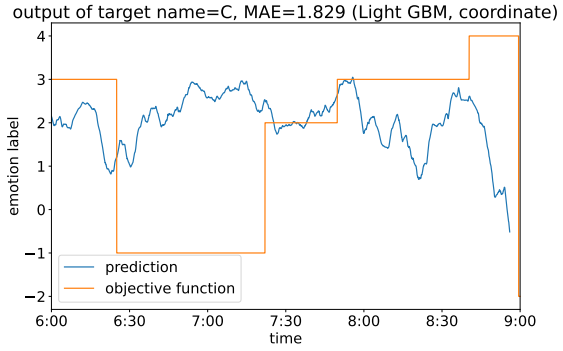

図 8　顔のランドマーク座標データのみを用いた Light GBM による参加者 C の開始 6 分から 9 分までの感情の推定結果

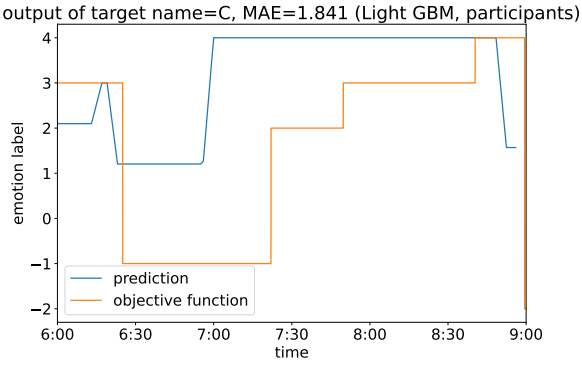

図 9　他者からのアノテーションデータのみを用いた Light GBM による参加者 C の開始 6 分から 9 分までの感情の推定結果

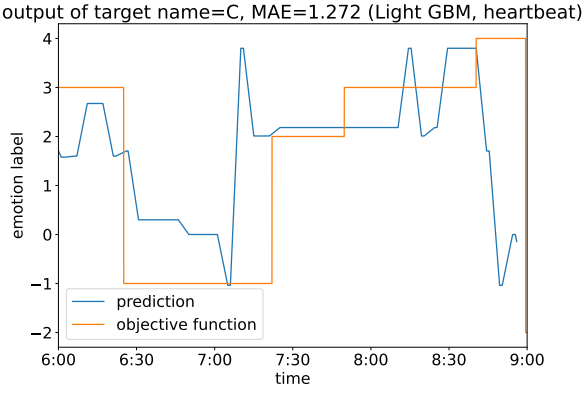

図 10　心拍データのみを用いた Light GBM による参加者 C の開始 6 分から 9 分までの感情の推定結果

分ではないことが示唆される．

7. まとめ

本稿では，オンラインミーティングサービスを用いたグループディスカッションにおける感情推定システムの実現に向けた初期段階として，OpenFace による映像解析，テキスト解析のために音声からの発言録データ作成を行った．また，Light GBM，SVR により感情ラベルを推定した結果，MAE において 1.780 を取った SVR のほうが 1.848 を取った Light GBM よりも良いモデルであることがわかった．一方で，グループディスカッションのマルチモーダルな時系列データを解析するために利用する評価指標は MAE では不完全であるということが判明した．

今後は，他の機械学習モデルや時系列データ特有の問題に対処できる評価指標を用いて，音声・映像・心拍・テキストデータをマルチモーダルに用いてオンラインミーティングサービスを用いたグループディスカッションにおける感情推定システムの実現を目指す．

謝辞

本研究の一部は，Society 5.0 実現化研究拠点支援事業および科研費基盤研究 (B)(No.19H01719) の助成によって行った．

参考文献

[1] Samiha Samrose, Jina Suh, Sean Rintel, Daniel McDuf, Kael Rowan, Kevin Moynihan, Robert Sim, Javier Hernandez, and Mary Czerwinski. Meetingcoach: An intelligent dashboard for supporting effective & inclusive meetings. *Conference on Human Factors in Computing Systems*, Vol. 16, No. 252, pp. 1–13, 2021.

[2] Ross Cutler, Yasaman Hosseinkashi, Jamie Pool, Senja Filipi, Robert Aichner, Yuan Tu, and Johannes Gehrke. Meeting effectiveness and inclusiveness in remote collaboration. *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5, No. 173, pp. 1–29, 2021.

[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, Vol. 30, pp. 3149—-3157, 2017.

[4] Mariette Awad and Rahul Khanna. Support vector regression. In *Efficient learning machines*, pp. 67–80. Springer, 2015.

[5] Janet T. Spence and Ann S. Robbins. Workaholism: Definition, measurement, and preliminary results. *Journal of Personality Assessment*, Vol. 58, No. 1, pp. 160–178, 1992.

[6] Herbert J. Freudenberger. Staff burn-out. *Journal of Social Issues*, Vol. 30, No. 1, pp. 159–165, 1974.

[7] Christina Maslach and Susan E. Jackson. The measurement of experienced burnout. *Journal of Occupational Behaviour*, Vol. 2, pp. 99–113, 1981.

[8] Christina Maslach, Wilmar B. Schaufeli, and Michael P. Leiter. Job burnout. *Annual Review of Psychology*, Vol. 52, pp. 397–422, 2001.

[9] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. *Winter Conference on Applications of Computer Vision*, 2016.

[10] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. *International Conference on Computer Vision Workshops*, 2013.

[11] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specic normalisation for automatic action unit detection.

[12] Jessica L. Tracy and David Matsumoto. The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *the National Academy of Sciences*, Vol. 105, No. 33, pp. 11655–11660, 2008.

[13] Ivan J. Tashev. Offline voice activity detector using speech supergaussianity. *Information Theory and Applications Workshop*, 2015.

[14] Ross Harper and Joshua Southern. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. *IEEE Transactions on Affective Computing*, 2020.

[15] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 133–140, 2005.

[16] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*, 2006.