映像と音声を用いた議論への関与姿勢や肯定的・否定的態度の推定方式の検討

金岡 翼[1] 上原 佑太郎[1] 原 直[1] 阿部 匡伸[1]

概要：本報告では，議論の映像と音声データを用いて，議論への関与姿勢の推定方式と発話の肯定的・否定的態度の推定方式を検討する．また議論の分析を行うにあたって，模擬会議の収録を行う．そこでの収録データにラベル付与を行い、教師データとして扱う．議論への関与姿勢の推定では，関与姿勢のラベルを 2 値分類としてランダムフォレストを用いて推定を行った．セッションごとでの交差検証においての最大の F 尺度として 0.75 となった．発話態度の認識実験では，SVM を用いて認識を行った．4 つのうち 3 つのセッションでは音声と映像を組み合わせた特徴量が音声と映像単独で得られる特徴量よりも高い F 尺度が得られた．

Examination of a method for estimating the attitude of participation in discussions and positive and negative attitudes by video and audio

KANAOKA TSUBASA[1] YUTARO UEHARA[1] SUNAO HARA[1] MASANOBU ABE[1]

1. はじめに

学習指導要領の改訂により，アクティブラーニング型授業の一環として議論を行う場面が増加している．すべての学生が積極的に議論に参加していることが望ましいが，議論内容を考えていなかったり，傍観してしまうなど，議論に加わることができない学生も存在する．そこで，議論の映像や音声から発話態度を推定したり議論への関与姿勢を推定することができれば，効率よく，教員から学生らの議論の扶助が可能となる．

コミュニケーションには，言語情報と非言語情報が含まれている．言語情報は言葉で表すことができる情報で，非言語情報はジェスチャや声の抑揚など，言葉以外の情報である．非言語情報の分析では，声の抑揚を表す基本周波数を用いて会話の盛り上がりを検出する研究 [1] が行われている．その際，各話者の発話に基づいた分析が行われている．また，発話の頻度や長さから議論の状態を推定する研究 [2] では盛り上がった際に現れる会話の衝突など，会話で起きるシーンに基づいた分析を行っている．また，発話態度の認識としては，音声だけではなく映像も活用した研究がなされている．例えば，Kobayashi ら [3] は，1 対 1 の対話における応答発話をカメラとマイクで収録し, その応答発話の非言語情報から発話態度を認識している．非言語情報としては，表情を表す, 輪郭のような顔のキーポイントと声の抑揚を表す基本周波数が用いられている．また,Fujie ら [4] は，音声対話システムに対する人の応答発話をカメラとマイクで収録し, 音声の基本周波数と，顔のキーポイントから判別した，頷き, 首をかしげる, 頭を横に振るという 3 つの頭の動きから, 発話態度認識を行い，音声対話システムに適用する研究が行われている．

本稿では，議論の映像データと音声データから抽出した特徴量を用いて，議論への関与姿勢と肯定・否定的態度を推定する方式について検討する．映像と音声を併用することで，議論参加者のふるまいを詳細に捉えられると考えられる．映像の特徴量としては，openpose を用いて，骨格や顔の特徴点の動作を分析する．音声の特徴量としては，opensmile を用いて，多角的な観点から音声の特徴を分析する．推定には，Support Vector Machine や ランダムフォレストなどの一般的な機械学習アルゴリズムを用いて，議論への関与姿勢と肯定・否定的態度の推定実験を行う．

[1] 岡山大学 大学院ヘルスシステム統合科学研究科

2. 提案方式

議論の参加者は，「発言者」としての役割と，「聴取者」としての役割を交互に担う．したがって，議論中の参加者の様子を分析するためには，発言者の振る舞いだけではなく，聴取者の振る舞いにも注目する必要がある．ただし，聴取者の役割を担っている時，その者は発言をしていないことが一般的であると考えられる．そこで，議論を精緻に分析するためには，発話中の音声だけではなく，映像や雑音等の音も併用することが重要だと考えられる．

本報告では，「発言者のふるまい」と「聴取者のふるまい」に注目して，分析をおこなう．議論中の発言者のふるまいとしては，提案としての発言，提案に対する肯定・否定などの態度を伴う発言，顔や上半身の動きによるジェスチャ等を交えた話し方，などの観点に注目することが考えられる．また，聴取者のふるまいとしては，顔の向き，椅子への座り方や座り直し，メモを取る際の雑音，体を動かした際の雑音，などが分析のための情報源になると考えられる．

そこで，音に関する特徴量と，映像に関する特徴量をそれぞれ抽出し，議論中の参加者の行動が検出されるという仮定の下，より高次の情報の推定実験を行う．本稿では，発言者の肯定・否定の態度推定と，聴取者含む参加者の議論への関与姿勢の推定を行う．

発話に関する特徴量は各話者に装着したピンマイクで収録した音声を用いて抽出する．議論の振る舞いは，各話者の正面に取り付けたカメラから特徴量を抽出する．

2.1 音響特徴量

音響特徴量を抽出するために OpenSMILE[5] を用いる．OpenSMILE は，音声認識や感情認識などで用いられる特徴量を抽出できるツールキットであり，議論音声の分析にてしばしば用いられている [6][7]．音響特徴量は，40 msec 程度の短い時間の区切り (フレーム) ごとで計算された Low Level Descriptor(LLD) に対して，基本統計量を計算することで得られる．

2.2 映像特徴量

映像特徴量を抽出するために，OpenPose[8][9] を用いる．OpenPose とは, CVPR2017 で発表された単眼カメラを用いたスケルトン検出アルゴリズムを実装したプログラムである．単一画像内の複数の人物の姿勢をリアルタイムで検出し, 人物を表す主要な特徴点を抽出する．抽出可能な特徴点を図 1 ，図 2 に示す．図 1 は人体の関節をつないだ骨格を表している．図 2 は顔のパーツ (目, 鼻, 口, 輪郭等) を表している．

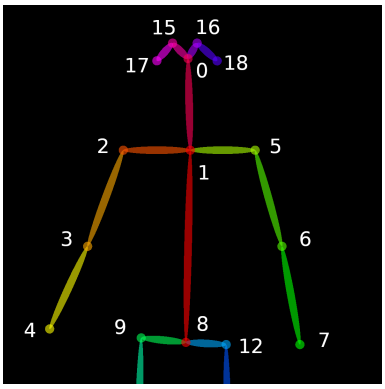

図 1 OpenPose で取得する骨格

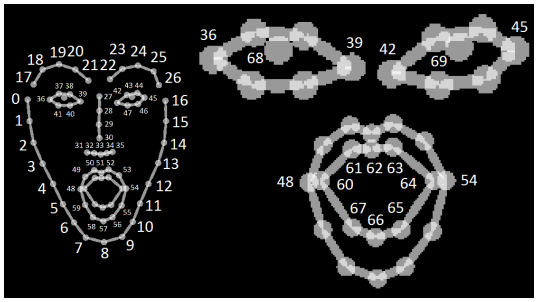

図 2 OpenPose で取得する顔の特徴点

2.2.1 上半身の動き

話者の振る舞いや顔の動きを特徴量としてとらえるために,OpenPose を用いて抽出した座標の時間差分を特徴量として使用する．抽出する座標は図 1 の骨格および関節の座標 [0-8, 15-18]13 点と，図 2 の顔のパーツを表す座標 70 点である．

時間差分の計算は式 (1) とする．なお $p(t)$ は二次元座標の点を表し，$\Delta p(t)$ は x 軸 y 軸それぞれの差分 ($\Delta x(t)$, $\Delta y(t)$) を持つ．

$$\Delta \mathbf{p}(t) = \mathbf{p}(t) - \mathbf{p}(t-1) \qquad (1)$$

計算した $\Delta x(t)$，$\Delta y(t)$ を映像の LLD とした．

全フレーム間での時間差分を，12 種類の基本統計量で計算した値を映像特徴量とする．

3. 会議音声・映像収録によるデータセット構築

3.1 模擬会議の収録

議論映像の収録参加者は著者と同じ研究室内の学生である．議論には 4 人参加しており，そのうちの 1 人は全ての議題においてファシリテーターを務めている．

図 4 に収録環境を示す．参加者全員の中央には，会議全体の音を収録するための会議用 USB マイクロフォン（YAMAHA PJP-20UR）を設置した．また，各参加者には，全指向性のラベリアマイクロフォン（SHURE MX187）を胸元に装着してもらった．各話者に装着したラベリアマイクの音は，USB オーディオデバイス（steinberg UR-44）

表 1　議論映像の構成

	議題	時間	関与姿勢のラベル数
議題 s1	今この瞬間トイレにいる人の数	16 分 55 秒	68
議題 s2	スクールバスにゴルフボールが何個入るか	12 分 00 秒	48
議題 s3	日本にあるマンホールの数	11 分 24 秒	48
議題 s4	日本にあるスターバックスの数	4 分 29 秒	20

を介して，Macbook Air 上の Audacity[*1] により収録した．映像の記録に関しては, どちらも本体内蔵のメモリに記録した．

機材の収録条件は，データセットとして取り扱いやすいものとするため, できる限り同じ値とした．音声ファイルの量子化ビット数は，ラベリアマイクロフォンと会議用 USB マイクロフォンの音声は 32 bit，個人映像用カメラは 24 bit である．音声のチャンネル数は個人映像用カメラが 2 チャンネル, それ以外の機材では 1 チャンネルである．映像用カメラはすべて 29.97 fps である． 音声データのヘッダーは RIFF 形式 (.wav) であり, 圧縮符号化方式は Linear-PCM である．また, 映像データのフォーマットは mp4 であり, 圧縮符号化形式は H.264 である．

議論映像の内容は表 1 の構成である．議題は 4 つあるが，議題 4 については，他の議題よりも収録時間が短くなっている．各議題は，フェルミ推定に基づく数量の予測問題であり，グループディスカッションの結果として，参加者全員の合意が取れた解答に辿り着くまで，議論をおこなう．フェルミ推定は，前提知識をあまり必要としないため，参加者全員が問題に関して取り組み易くなっている．

図 3 に議論での会話例を示す．また，m1-4 はそれぞれ議論参加者で，m1 はファシリテーターである．議論の序盤はファシリテーターから問題の提起があり，議論でのゴールが定まる．そこから議論参加者同士で，前提条件の合意形成を行う．中盤では，序盤で定めた前提条件をもとに問題を解く議論を行う．終盤では，中盤で立てた仮説の検証を行い，答えの合意形成を取る．

3.2　議論参加者の関与姿勢に関する評価ラベルの付与

「聴取者」の視点からの評価を目的として，議論参加への関与姿勢を第三者に評価ラベルを付与させた．

収録した映像データは，大まかに 1 分ごとの映像に分割した．この 1 分の映像を，シーンと呼ぶ．そして，各シーンを対象に，各議論参加者の議論への関与姿勢を，3 名の評価者によって評価した．

評価者は，図 5 に示すような動画を視聴しながら，評価ラベルを付与する．評価用の動画は，図 4 に示した全体カメラ 2 台の映像を結合し，各話者が話している声や振る舞いが全て確認できるように編集している．評価ラベルは，1 分のシーンに映る議論参加者 4 人に対して，関与姿勢が悪い (-1)，普通 (0)，良い (1) の 3 段階を付与する．1 シーン視聴後には，評価ラベルを考える時間を 1 分 20 秒設け

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

話者	発話内容
m1)	「日本のスタバの数を考えましょう.」
m2,3,4)	「おぉ (感嘆)」
m3)	「まず岡山規準で考えてみる?」
m1)	「でも都会にはいっぱいあるし」
m2)	「鳥取には 1 個しかない」
m4)	「山口県にはようやく 2 つ目ができて大騒ぎしていた」

(a) 議論 s4-序盤：議論のはじまり

話者	発話内容
m3)	「じゃあ 20 × 47 ですか?」
m2)	「これは答え出そうやな」
m4)	「じゃあ 940 個あると」
m2)	「でも大丈夫か?このガバガバ計算」
m1)	「日本全国に 940?」
m2)	「少ない気がしますよね」
m3)	「うん」

(b) 議論 s4-中盤：岡山県内にある数を 20 件であると合意形成した後

話者	発話内容
m2)	「なんか東京・大阪・名古屋に 300，300，300 あって，残りは分け合うみたいな」
m3)	「本当に都会圏で数百あって，他で分け合ってる感はある」
m2)	「って考えたら妥当な気がしてきた」
m1)	「じゃあ，これ (スタバの数) はこれ (970) で」
m2,3,4)	「はい」

(c) 議論 s4-終盤：全国のスターバックスの数を 970 件であると仮定した後

図 3　議論での対話例

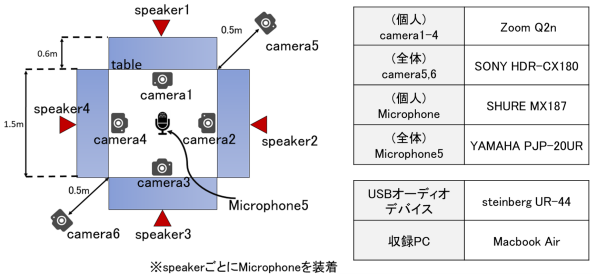

図 4　議論映像の収録環境

た．議論ごとに議論開始から終了までのシーンを順番に視聴し，評価を行った．ただし，視聴順によって評価者の基準が変わる可能性を考慮して，議論の提示順序は評価者ごとで変えている．

なお，評価値の付与にあたって，事前に，議論時間が短い議題 4 の動画 (5 シーン) を用いて評価練習させた．その後，評価者間の評価基準の統制を行った．その際に，下記の 3 点について話し合わせた。

- どのような点が関与姿勢が悪く見えたか
- どのような点が関与姿勢が良く見えたか
- 評価する上でどのような所を注視したか

評価者別にした評価ラベルの点数ごとの分布を表 2 に示す．また，議論映像の議題別にした評価ラベルの点数ごとの分布を表 3 に示す．

3.3　発話者の肯定的・否定的態度のラベル付与

「発言者」の視点からの評価を目的として、発言者自身

表 2　評価者別評価ラベルの点数ごとの分布

	評価者 m1	評価者 m2	評価者 m3	評価者 m4
1(良い)	93	104	30	75
0(普通)	36	59	98	77
-1(悪い)	35	1	36	32
-1 の割合	0.21	0.01	0.22	0.20

表 3　議題別評価ラベルの点数ごとの分布

	議題 s1	議題 s2	議題 s3
1(良い)	130	103	69
0(普通)	105	60	85
-1(悪い)	37	29	38
-1 の割合	0.14	0.15	0.20

図 5　評価用動画の一例

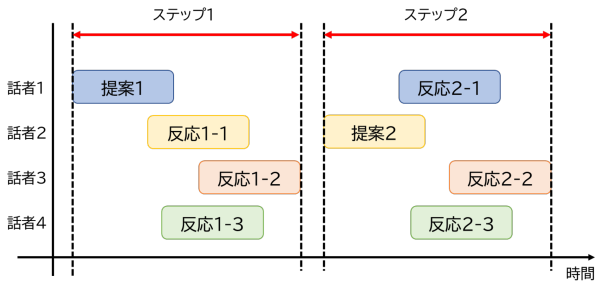

図 6　想定する議論の形式

によって、発言内容の肯定否定について評価ラベルを付与させた。なお，このラベル付与においては，議論中の会話は，図 6 のような提案と応答の繰り返しにより成り立つと考え，このような発話単位で行った．図 6 では，話者 1 の提案に対し，話者 2-4 が応答している．この提案-応答の一連の流れをステップ 1 とする．同様にステップ 2 では話者 2 の提案に対して，話者 1, 3, 4 が応答している．なお，発話区間の抽出は参加者 4 名のピンマイク音声データの用いて，始まりと終わりの時刻を決定した．

ラベルは，動画解析ツールの ELAN[*2] で会議音声と映像を見ながら付与させた．ラベラーが付与するラベルは以下

[*2] https://archive.mpi.nl/tla/elan

表 4　会議実験データ

	参加者 m1	参加者 m2	参加者 m3	参加者 m4
発話時間 [s]	916.6	926.9	930.1	374.5
発話区間ラベル数	395	389	654	234
平均ラベル長 [s]	2.321	2.382	1.422	1.600
ラベル長標準偏差 [s]	0.973	1.772	0.966	0.999

表 5　参加者ごとの発話態度ラベルデータ

	PA	NA	Neu.	Oth.
参加者 m1	134	43	257	0
参加者 m2	73	28	246	40
参加者 m3	89	37	459	71
参加者 m4	40	19	99	88

表 6　議題ごとの発話態度ラベルデータ

	PA	NA	Neu.	Oth.
議題 s1	141	48	401	83
議題 s2	117	28	305	58
議題 s3	47	35	231	53
議題 s4	31	16	124	5

の 4 つである．

- Positive Attitude(PA): 他者の提案に対して同意するなどの肯定的な態度を伴う応答
- Negative Attitude(NA): 他者の提案に対して拒否するなどの否定的な態度を伴う応答
- Neutral: 提案 (第 1 部分)
- Other: 上記 3 つ以外の, 独り言などの発話

PA，NA ラベルは，発話者本人の基準でラベル付けを行った．

表 4 に各参加者の全セッションにおける発話時間データをまとめる．模擬会議時間のうちの発話時間が参加者 m4 で約 6 分, その他の参加者で約 15 分となった．

発話態度ラベルの付与によって得られたそれぞれのラベルの数を話者別の合計した数を表 5，議題別に合計した数を表 6 にまとめる．それぞれの枠の数値は [PA, NA, Neutral, Other] ラベルの順番になっている．ラベル数を参加者, セッションごとにラベルを集計した．PA ラベルは全参加者で 303 個, NA ラベルは 125 個, Neutral ラベルは 1055 個, Other ラベルは 199 個となっている．

4. 評価実験

4.1 議論への関与姿勢の推定実験

3.2 で作成したラベルを用いて，音響特徴量と映像特徴量からランダムフォレスト法により，関与姿勢を推定する実験を行った．関与姿勢が悪いかどうかの判断を評価するため，本実験では，評価者 4 人の合計値が −1 以下の場合を関与姿勢が悪い，0 以上の場合を関与姿勢が悪くないとする 2 値分類とした．2 値化した後のデータ数を表 7 に示す．

本実験では，各議題 1 つをテストデータとして，用いて他を学習データとして分類を行うことで交差検証を行う．

表 7　2 値化した後の評価ラベルの数

	議題 1	議題 2	議題 3	合計
関与姿勢が悪い (評価者 4 人の合計値 ≦-1)	15	9	15	39
関与姿勢が悪くない (評価者 4 人の合計値 ≧0)	53	39	33	125
関与姿勢が悪いラベルの割合	0.22	0.19	0.31	0.24

表 8　関与姿勢の推定で使用した特徴量

音響特徴量 (152 次元)	
LLD	音声確率 (Probability)
	音の大きさ (Loudness)
	音の強さ (Intensity)
	零交差率 (Zero-Crossing Rate)
	各 LLD の動的特徴量
基本統計量	最大値/最小値とその範囲
	最大値/最小値のあるフレーム位置算術平均
	標準偏差
	線形近似における勾配と切片
	線形近似における真値との二乗誤差
	尖度
	歪度
	各四分位数
	各四分位数ごとの範囲

映像特徴量 (1992 次元)	
LLD	座標の時間差分 ($\Delta x(t)$，$\Delta y(t)$)
基本統計量	最大値/最小値とその範囲
	最大値/最小値のあるフレーム位置
	算術平均
	標準偏差
	線形近似における勾配と切片
	線形近似における真値との二乗誤差
	尖度
	歪度

表 9　ランダムフォレストに使用するハイパーパラメータ

分岐に使用する特徴量の数	18
分割基準	情報利得
決定木の最大の深さ	None
決定木の数	1000

議題ごとの評価ラベル数を表 1 に示す．

4.1.1　実験条件

使用する音響特徴量および映像特徴量を表 8 に示す．ランダムフォレストは Python の機械学習ライブラリ scikit-learn[*3]を用いて実装した．ランダムフォレストのハイパーパラメータはグリッドサーチを用いて選択した．本実験で使用したランダムフォレストのハイパーパラメータを表 9 に示す．評価指標として各議題ごとでの交差検証の結果で生じた平均 F 尺度を用いる．F 尺度の計算の際は，関与姿勢が悪い方を正例として計算を行う．

4.1.2　実験結果

3 つの議題ごとでの交差検証での分類結果を表 10 に表す．表 10 より，各議題ごとの F 尺度は，議題 s1 の値がもっとも低く，議題 s2 での F 尺度が最も高い結果となった．また，議題 s1 や議題 s3 において F 尺度が悪い理由として Recall が悪いことが原因であることが分かる．

[*3] https://scikit-learn.org

表 10　議題ごとの交差検証の結果

テストデータ	議題 1	議題 2	議題 3	平均
Precision	0.50	0.78	0.75	0.68
Recall	0.14	0.73	0.27	0.38
F 尺度	0.21	0.75	0.40	0.45

表 11　重要度の高い音響特徴量の説明変数

説明変数	重要度
音の大きさ_線形回帰係数の切片	0.0149
音の大きさ_線形二乗誤差	0.0142
音の大きさ (動的特徴量)_線形二乗誤差	0.0139
音の大きさ (動的特徴量)_線形回帰係数の切片	0.0135
零交差率_歪度	0.0118

表 12　重要度の高い映像特徴量の説明変数

説明変数	重要度
左頬 (14)_$\Delta y(t)$_線形二乗誤差	0.0045
右目の下瞼 (40)_$\Delta x(t)$_最大値	0.0029
右目の下瞼 (40)_$\Delta x(t)$_線形二乗誤差	0.0029
右眉右端 (17)_$\Delta y(t)$_標準偏差	0.0028
右顎 (6)_$\Delta x(t)$_勾配	0.0026

重要度が高い順番に音響特徴量と映像特徴量の重要度の一覧を表 11，表 12 に示す．なお，表 12 にある括弧内の数字は図 2 の座標と対応している．

説明変数の重要度を確認すると音の大きさ (Loudness) と零交差率 (Zero-clossing rate) が重要であることが分かった．また，音響特徴量と比べて映像特徴量は重要度が低い結果となった．

4.1.3　考察

重要度の高い特徴量である音の大きさの線形二乗誤差について分析を行った．線形二乗誤差とは，音の大きさの時間変化を二乗誤差が最小になるように直線で近似した時の誤差である．図 7 は，この誤差を関与姿勢が悪いと，悪くないとの判断に分けて表示している．ここで示されるように，関与姿勢が悪くない場合には誤差は大きな値をとることが分かる．誤差が大きいことは，次の理由により，発話時間が長いことに関係していると考えられる．4 名の議論であることから個々の参加者についてみれば，1 分の時間区間では聞いている時間が長く，発話している時間は短いと予想される．そのため，線形近似すると聞いている時間（音の時間が小さい）に偏った直線式となる．発話があった場合は，この近似式より外れるため，誤差が大きくなる．以上のことから，発話時間が長いほど関与姿勢が高く評価されたと考えられる．

4.2　発話者の発話態度の推定実験

3.3 で作成したラベルを用いて，音響特徴量と映像特徴

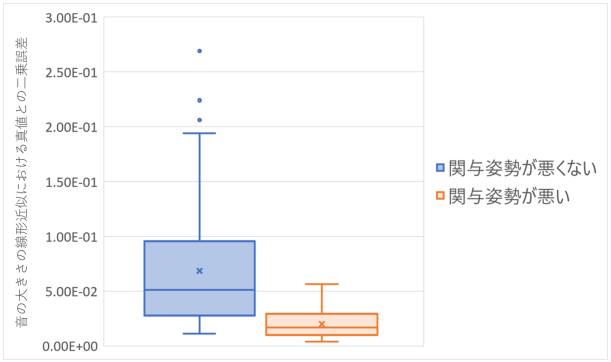

図 7　関与姿勢ごとの線形二乗誤差

表 13　発話推定で用いた特徴量

音響特徴量 (384 次元)	
LLD	音声確率 (Probability)
	音の大きさ (Loudness)
	信号エネルギーの二乗平均 (RMS energy)
	メル周波数ケプストラム (MFCC[1]-[12])
	零交差率 (Zero-Crossing Rate)
	各 LLD の動的特徴量
基本統計量	最大値/最小値とその範囲
	最大値/最小値のあるフレーム位置算術平均
	標準偏差
	線形近似における勾配と切片
	線形近似における真値との二乗誤差
	尖度
	歪度
	各四分位数
	各四分位数ごとの範囲

映像特徴量 (1992 次元)	
LLD	座標の時間差分 ($\Delta x(t)$, $\Delta y(t)$)
基本統計量	最大値/最小値とその範囲
	最大値/最小値のあるフレーム位置
	算術平均
	標準偏差
	線形近似における勾配と切片
	線形近似における真値との二乗誤差
	尖度
	歪度

量から SVM 法により，発言の肯定 と否定を推定する実験を行った．

4.2.1　実験条件

本実験では，4 名の参加者の PA/NA ラベルを用いた 4 交差検証により肯定否定の認識精度を比較する．なお，未知話者条件を模擬した 4 交差検証と，未知セッション条件を模擬した 4 交差検証を，それぞれ行う．

発話態度の推定を行うための特徴量を表 13 に示す．このとき，以下の 3 つの特徴量セットごとに評価実験を行い特徴量の有効性を比較する．

- 音響特徴量 384 次元

参加者	音響特徴量			映像特徴量			結合特徴量		
	Rec	Pre	F	Rec	Pre	F	Rec	Pre	F
m1	.610	**.872**	.718	.623	.817	.707	**.667**	.854	**.749**
m2	.603	.872	.713	.603	.911	.726	**.618**	**.913**	**.737**
m3	.604	**.762**	.674	.642	.708	.673	**.660**	.761	**.707**
m4	**.639**	**.885**	**.742**	.528	.826	.644	.556	.769	.645

(a) PA の検出

参加者	音響特徴量			映像特徴量			結合特徴量		
	Rec	Pre	F	Rec	Pre	F	Rec	Pre	F
m1	**.771**	.360	**.478**	.553	.313	.400	.632	**.369**	.466
m2	.750	.400	.522	.833	.426	.563	**.862**	**.490**	**.625**
m3	**.697**	.523	.597	.576	.500	.535	.677	**.550**	**.603**
m4	**.800**	**.480**	**.600**	.733	.393	.512	.600	.360	.450

(b) NA の検出

表 14　未知話者条件における実験結果

議題	音響特徴量			映像特徴量			結合特徴量		
	Rec	Pre	F	Rec	Pre	F	Rec	Pre	F
s1	.638	**.892**	.744	.723	.825	.770	**.746**	.882	**.808**
s2	.622	**.944**	.750	.610	.833	.704	**.756**	.886	**.816**
s3	**.756**	**.816**	**.785**	.732	.618	.667	.732	.789	.759
a4	.667	.818	.735	.667	.818	.735	**.778**	**.913**	**.840**

(a) PA の検出

議題	音響特徴量			映像特徴量			結合特徴量		
	Rec	Pre	F	Rec	Pre	F	Rec	Pre	F
s1	**.773**	.420	.544	.546	.400	.462	.705	**.484**	**.574**
s2	**.870**	.392	**.541**	.565	.289	.382	.652	**.429**	.517
s3	**.774**	**.706**	**.738**	.387	.522	.444	.741	.676	.708
s4	**.667**	.471	.552	.667	.471	.551	.833	**.625**	**.714**

(b) NA の検出

表 15　未知セッション条件における実験結果

- 映像特徴量 1992 次元
- 結合特徴量 2376 次元

4.2.2　未知話者条件における実験結果

肯定的態度 (PA) の評価値を表 14(a) に示す．F 尺度に関しては, m1, m2, m3 の評価データに結合特徴量を用いた認識で最大となっている．Recall に関しても F 尺度と同様に m1, m2, m3 の評価データに結合特徴量を用いた場合に最大となっている．Precision に関しては, m1, m3, m4 の評価データに音響特徴量を用いた場合に最大となっている．

否定的態度 (NA) の評価値を表 14(b) に示す．F 尺度に関して, m1 と m4 では音響特徴量を用いた認識で最大となり, m2 と m3 では結合特徴量を用いた認識で最大となっている．Recall に関しては, m1, m3, m4 で音響特徴量を用いた認識で最大となっている．Precision に関しては, m1, m2, m3 の評価データに結合特徴量を用いた場合に最大となっている．

4.2.3　未知セッション条件における実験結果

肯定的態度 (PA) の評価値をを表 15(a) に示す．F 尺度に関しては, s1, s2, s4 の評価データに結合特徴量を用いた認識で最大となっている．Recall に関しても F 尺度と同様に，s1, s2, s4 の評価データに結合特徴量を用いた認識で最大となっている．Precision に関しては s1, s2, s3 の評価データに音響特徴量を用いた認識で最大となっている．

否定的態度 (NA) の評価値を表 15 に示す．F 尺度に関

しては, s1, s2, s4 の評価データに結合特徴量を用いた認識で最大となっている．Recall に関しても F 尺度と同様に s1, s2, s4 の評価データに結合特徴量を用いた認識で最大となっている．Precision に関しては s1, s2, s3 の評価データに音響特徴量を用いた認識で最大となっている．

4.2.4 考察

表 14(a) と表 15(a) に示した，各実験条件下における PA の評価値を比較すると, F 尺度に関しては結合特徴量を用いた場合に最大となる評価データの方が多かった．しかし，一部の例外がみられた．例えば, 話者方向の評価値では, 参加者 m4 に関してのみ, 結合特徴量の F 尺度が最大値ではない．このとき, 再現率を表す Recall に着目すると, 結合特徴量では 0.556, 映像特徴量では 0.528 となっており, 各参加者と比較すると, どちらの特徴量の Recall も最小となっている．これより，参加者 m4 に関して, 結合特徴量の F 尺度が最大値ではないのは, 他の参加者と比べて再現率が低いことが原因であると考える. そして, 参加者 m4 の映像の特徴量は他の参加者とは異なることが考えられる．また, 表 15(a) のセッション方向では, セッション 3 に関してのみ, 結合特徴量の F 尺度が最大値ではない．Precision に着目すると, 映像特徴量で 0.612 となっており，他の特徴量と比べて低い．これより, セッション 3 においてのみ, 結合特徴量のうち映像特徴量が PA の検出における信頼性に悪影響を与えて, 結合特徴量の F 尺度が最大値とならなかったと考える．

表 14(b) と表 15(b) に示した，各実験条件下における NA の評価値を比較すると, F 尺度に関しては, 音響特徴量と結合特徴量を使った場合で最大値は分かれている．表 14(b) の話者方向の F 尺度では, 参加者 m1 と m4 で音響特徴量, 参加者 m2 と m3 で結合特徴量でそれぞれ最大値となっている．Recall に関しては, 表 14(b) の話者方向において, m2 の評価データ以外の場合で音響特徴量を用いたときに最大になっている．これより, NA の検出においては, 音響特徴量を用いた方が再現率が高くなると考える．一方, Precision に関しては, 結合特徴量を用いた方が他の特徴量と比べて, 表 14(b) の話者方向では m4, 表 15(b) のセッション方向では s3 以外の評価データで最大となった．このことから, NA の検出の信頼性が高いのは結合特徴量を用いた場合であると考える．

4.3 両実験を踏まえた考察

関与姿勢の推定を議題ごとの交差検証で行った結果，分類性能を表す F 尺度は最大 0.75 となった．また，議題 1 または議題 3 をテストデータとした際の F 尺度はそれぞれ，0.21，0.75，0.40 となった．F 尺度が低い原因として，Recall が低いことが挙げられる．重要な特徴量の分析の結果，音響特徴量と比べて，映像特徴量の重要度が低くなっていた．このことから，関与姿勢の推定では音響特徴量の寄与が大きく，あくまで映像特徴量は補助的な精度向上に貢献しているものと考えられる．

肯定・否定的態度の推定では，4 人中 3 人の話者において結合特徴量を用いたときに PA の Recall が最大値をとった．また，precision では 4 人中 3 人の話者において音響特徴量のみを用いたときに最大値をとった．否定的態度 (NA) の Recall においては音響特徴量がすべての話者において最大値をとった．Precision においては結合特徴量を用いた際に，比較的に最大値を示した．音響特徴量のみでもある程度の精度で分類できていることから，映像特徴量は音響特徴量に対して補助的な特徴量になっていると考えられる．

5. まとめ

本報告では議論の映像データと音声データから抽出した特徴量を用いて，議論への関与姿勢と肯定･否定的態度を推定する方式について検討した．関与姿勢が悪く見える議論参加者には発話回数が少なかったり，話を聞く姿勢が悪いなど特徴があることに着目した．映像特徴量と音響特徴量を用いることで，議論への関与姿勢や発言の肯定・否定的態度を推定する方式を考案した．音響特徴量として，各議論参加者に取り付けたラベリアマイクロフォンの録音データから，4 種類の音響特徴量を抽出した．映像特徴量として，各議論参加者の前に設置したカメラから，骨格や顔の特徴点の動的特徴量を抽出した．議論参加者の分析を行うという観点から，4 人の議論評価者に議論の関与姿勢の評価を行ってもらい，議論映像に対してラベル付与を行った．また，発言者の分析を行うという観点から，議論参加者本人による肯定・否定的態度のラベル付与を行った．評価実験として，映像特徴量と音響特徴量を用いて，関与姿勢と肯定・否定的態度の推定と重要な特徴量の分析を行った．

両方の実験では，音響特徴量及び映像特徴量を用いて検証を行ったが，顔が動いた時の発話のタイミングなど完全に同期した特徴量は使われていない．議論への関与姿勢の評価は，ボディランゲージと発話を合わせてのコミュニケーションを見て評価をするため，今後，音声と映像の同期のとれた特徴量を考える必要がある．

6. 謝辞

本研究は JSPS 科研費 18K02862 の助成を受けて実施したものである．

参考文献

[1] 坂原誠, 岡田将吾, 新田克己. ”マルチモーダル情報を用いた情緒的な発話検出と議論分析.” 人工知能学会全国大会論文集 第 27 回全国大会 (2013). 一般社団法人 人工知能学会, 2013.

[2] 市野順子, 田野俊一. ”発言の時系列的パターンを用いた会議における発散/収束の判別の可能性.” 人工知能学会論文

誌 25.3 (2010): 504-513.
[3] Y. Kobayashi, M. Nakamura, H. Nambo, and H. Kimura,"Discrimination of positive/negative attitude using optical flow and prosody information,"IEEJ Transactions on Electronics, Information and Systems, vol.136, no.3, pp.401-408, 2016.
[4] S. Fujie, Y. Ejiri, H. Kikuchi, and T. Kobayashi, "Recognition of positive/negative attitude and its application to a spoken dialogue system,"Systems and Computers in Japan, vol.37, pp.45-55, Nov. 2006.
[5] Eyben, Florian, Martin Wöllmer, and Björn Schuller." Opensmile: the munich versatile and fast open-sourceaudio feature extractor." Proceedings of the 18th ACM international conference on Multimedia. 2010.
[6] Yu-Chang Ho, et al. "Automatic Opinion Leader Recognition In Group Discussions," 2016 Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, 2016.
[7] L. Nichola, E. Walker, and H. Pon-Barry. "Relating entrainment, grounding, and topic of discussion in collaborative learning dialogues." Proceedings of Computer Supported Collaborative Learning. 2015.
[8] Z. Cao, G. Hidalgo, T. Simon, S.E. Wei, and Y. Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).
[9] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. "Hand keypoint detection in single images using multiview bootstrapping." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.