発話者と聴取者の役割に基づいた議論への関与姿勢の推定に用いる特徴量の検討

A study of features used for estimating speaker and listener role-based attitudes of involvement in discussions

金岡　翼

Tsubasa Kanaoka

岡山大学 阿部研究室

Abe Laboratory, Okayama University

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を検討する．関与姿勢のラベルを 2 値分類として議論における役割ごとにランダムフォレストを用いて推定を行った．聴取者のみを学習に用いた場合，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 0.73 となった．推定に用いた重要な特徴量を見ると，聴取者の関与姿勢の推定では音響特徴量が有効であることが分かった．一方，発話者の関与姿勢の推定には，音響特徴量だけでなく，映像特徴量も有効であることが分かった．

1　はじめに

議論での議論での円滑な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．非言語情報の分析において，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．

議論を分析するにあたって，議論参加者は，議論を進行するにあたって発話者と聴取者の役割を交互に担うと仮定する．議論中の発話者は，提案としての発言，顔や上半身の動きによるジェスチャ等を伴った発言などの振る舞いが見られ，聴取者は，発話者の方を向く，椅子への座り方や座り直し，メモを取るなどの振る舞いが見られると考えられる．従って，議論を分析するにあたって，発話者と聴取者は別の特徴があると仮定できる．しかし，聴取者は，発話者の話を聞くことが主になるため，聴取者の役割を担う場合，発言を行わないと考えられる．そこで，議論への関与姿勢を推定するためには発話中の音声だけでなく，映像や雑音を用いることが重要であると考えられる．

本稿では，発話の有無や話者の感情が含まれる音響特徴量に，身振り手振りを表す映像特徴量を組み合わせて議論参加者の分析を行う．特に，発話者と聴取者の役割ごとに関与姿勢の推定を行い，それぞれの役割での特徴量の有効性を検討する．

2　提案方式

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．

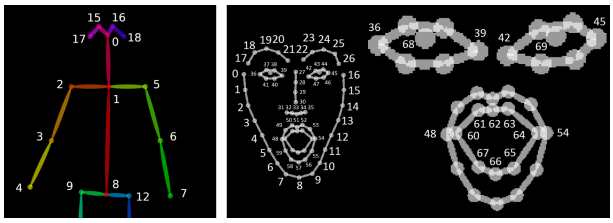


図 1: OpenPose で取得する骨格と特徴点

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

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．

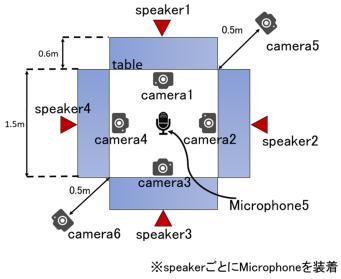
2.2　映像特徴量

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 に示す．図 1(左) は人体の関節をつないだ骨格を表している．図 1(右) は顔のパーツ (目, 鼻, 口, 輪郭等) を表している．

3　実験で使用する会議データ

図 2 に実験で使用する会議データの収録環境を示す．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る．

本実験で使用する議論の関与姿勢の正解ラベルは，1 分ごとに区切られた映像 (シーン) に付与した．議論参加者とは異なる 4 名の評価者により，各議論参加者に



（個人）camera1-4	Zoom Q2n
（全体）camera5,6	SONY HDR-CX180
（個人）Microphone	SHURE MX187
（全体）Microphone5	YAMAHA PJP-20UR
USBオーディオデバイス	steinberg UR-44
収録PC	Macbook Air

図 2: 議論映像の収録環境

表 2: 役割ごとのラベル数

		議題 s1	議題 s2	議題 s3	合計
発話者	悪い	2	2	2	6
	悪くない	30	20	16	76
聴取者	悪い	12	6	9	27
	悪くない	7	8	9	24

対して，関与姿勢を 3 段階で評価した．なお，評価の前に，収録時間の短い議題 s4 を用いて，4 名の評価者の意識合わせをしている．

聴取者と発話者の役割ごとに推定結果を比較するために，1 分間のシーンのうち，発話のアノテーションがされている区間が 14 秒以上である場合を発話者として，それ以外を聴取者とした．表 2 に役割ごとの関与姿勢のラベル数を示す．それぞれの役割ごとに推定器を作成し，重要な特徴量や推定精度を比較する．

4 議論中の関与姿勢の推定

4.1 実験条件

関与姿勢の推定を行うための特徴量を表 3 に示す．音声データから OpenSMILE[5] を用いて音響特徴量を抽出し，関与姿勢の推定実験では 152 次元，映像データから Openpose[8] を用いて抽出した計 83 点の座標の各 x，y 軸の時間差分を 12 種類の基本統計量で計算した値を映像特徴量とした．以下の 3 つの特徴量ごとに推定を行い，比較する．

- 音響特徴量 152 次元
- 映像特徴量 1992 次元
- 結合特徴量 2144 次元

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．

推定器は，ランダムフォレストである．また，Python の機械学習ライブラリ scikitlearn を用いて実装した．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．

表 3: 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 4: 聴取者の議題ごとの交差検証の結果

	議題 s1	議題 s2	議題 s3	平均
音響特徴量	0.67	0.57	0.73	0.66
映像特徴量	0.14	0.60	0.69	0.48
結合特徴量	0.50	0.56	0.67	0.58

表 5: 発話者の議題ごとの交差検証の結果

	議題 s1	議題 s2	議題 s3	平均
音響特徴量	0.13	0.18	0.29	0.20
映像特徴量	0.15	0.29	0.25	0.23
結合特徴量	0.13	0.20	0.31	0.21

4.2 実験結果

表 4 に聴取者の関与姿勢ラベルの推定結果を示す．聴取者の関与姿勢ラベルの推定では，議題ごとの平均 F 尺度を見ると音響特徴量が最も F 尺度が高いことが分かった．映像特徴量を用いて議題 s1 のラベルを推定した場合を確認すると，他の議題に比べて F 尺度が低くなったことから，映像特徴量を用いた際，場合によっては推定精度が落ちることが分かった．

表 5 に発話者の関与姿勢ラベルの推定結果を示す．発話者の関与姿勢ラベルの推定では，すべての特徴量において同等の F 尺度となった．その中でも，映像特

表 6: 聴取者の推定器において重要度の高い特徴量

特徴量	重要度
零交差率_第 1 四分位数	0.0104
音の大きさ (動的特徴量)_最大値と最小値の差	0.0087
音の大きさ_最大値	0.0078
音の大きさ_標準偏差	0.0071
音の大きさ_線形近似での真値との二乗誤差	0.0070

表 7: 発話者の推定器において重要度の高い特徴量

特徴量	重要度
音声確率 (動的特徴量)_最大値	0.0084
零交差率_線形近似での真値との二乗誤差	0.0068
音声確率_最大値	0.0061
右こめかみ (1)_$\Delta y(t)$_平均	0.0054
右眉 (20)_$\Delta x(t)$_線形近似での真値との二乗誤差	0.0052

徴量を推定に用いた際に F 尺度が最も高い結果となった．聴取者の関与姿勢ラベルの推定と比べて，どの特徴量でも推定精度が低い．その理由として，各役割での関与姿勢ラベルの割合が挙げられる．表 2 の役割ごとの関与姿勢ラベルの割合を確認すると，聴取者としたデータのうち関与姿勢が悪いとしたラベルの割合が，53% であるのに対いて，発話者としたデータでの割合は，% であるのに対いて，7% であった．発話者の関与姿勢ラベルを推定する際に，関与姿勢が悪いラベルが少ないため，推定精度が低いのではないかと考えられる．

次に特徴量の重要度について比較を行う．それぞれの推定器で重要度の高い上位 5 つの特徴量を表 6，7 にそれぞれ示す．重要度の高い特徴量を比較すると，聴取者の推定器の場合は，音の大きさや零交差率が有効であるのに対して，発話者の推定器は，音声確率や零交差率，映像特徴量が有効であることが分かった．上位 30 個の特徴量のうち，いくつ映像特徴量があるかを比較すると，聴取者の推定器の場合は 6 個に対して，発話者の推定器の場合では，12 個もの映像特徴量が含まれていた．このことから，聴取者の関与姿勢を推定する際は，音響特徴量が有効であり，発話者の関与姿勢を推定する際は，聴取者に比べて映像特徴量がより有効であることが分かった．

5 まとめ

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，議論への関与姿勢の推定を行った．実験では，議論での役割ごとに各特徴量の有効性の比較を行い，推定結果の分析を行った．結果として，どちらの役割でも音響特徴量が有効であるが，発話者の関与姿勢の推定をする際は，映像特徴量がより有効であることが分かった．

今後の課題として，同期された音声と映像から得られる特徴量の検討が挙げられる．例えば，顔が動いた時の発話のタイミングなどである．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，同期した音声と映像から得られる特徴量を考える必要がある．また，映像特徴量は低レベルの情報しか扱っていないため，議論参加者の視線などの情報を特徴量として取り入れたいと考えている．

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.