5X-07

グループミーティング動画からの発話量抽出手法の検討

徳原 耕亮* Billy Dawton† 荒川 豊† 石田 繁巳† 曽根田 悠介‡ 松田 裕貴‡

*九州大学工学部電気情報工学科 †九州大学大学院システム情報科学研究院
‡奈良先端科学技術大学院大学

1 はじめに

グループミーティングは現代社会において情報交換や問題解決，知識共有の際に必要不可欠である．グループミーティングでは会議本来の目的を見失うことなく議論を行うことが重要であり，参加者全員が意欲的に発言することが求められる．ミーティング参加者の意欲的な発言に向けては，参加者がどれくらいの時間割合で発話していたかを示す「発話量」に基づいた会議支援システムが有効であると考えられる．本稿では，グループミーティング参加者の発話量を動画のみから抽出する手法を検討した結果を報告する．

2 関連研究

発話量推定の既存研究には，ビデオカメラ，マイク，モーションセンサなどの複数のセンサを使用する手法が報告されている [1]．この手法は多数のセンサを用意・装着しなければならず，手間と費用が大きいという問題がある．また，文献 [2] では 2 台のマイクを用いて複数人の会話音声から音源定位により発話者を推定している．しかしながら，音声を用いる手法は発話者が移動すると発話者の特定が困難となる問題がある．また，人数が多い場合の音源分離，すなわち発話者の分離が困難であり，グループミーティング参加者が多い場合に適用することが難しい．

3 提案手法

図 1 に，提案する発話量抽出手法の概要を示す．我々は，360 度カメラ，視線追跡装置，頭部に装着する加速度センサを同期して記録可能なグループミーティング録画システムを構築している [3]．本録画システムを用いて録画した映像データの各フレームから OpenFace を用いて顔及びランドマークを検出し，得られた顔のランドマークから特徴量を抽出する．抽出された特徴量を用いて機械学習により開口状態を推定した上で，

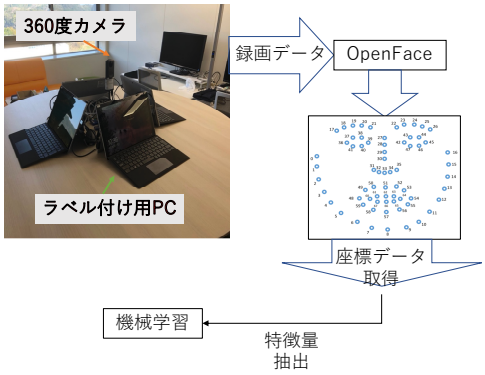

図 1: 発話量抽出手法の概要

口を開いている時間割合を算出することで発話量を推定する．

特徴量として，OpenFace で得たランドマークのうち，開口状態を表す距離として図 2 に示す点 61 と 67，62 と 66，63 と 65 間のユークリッド距離を用いる．この特徴量を用いて機械学習により開口状態，すなわち「口が開いているか」を推定した上で開口時間の割合を「発話量」とする．開口状態の推定において使用する機械学習手法は限定しないが，本稿では Random Forest を用いた．

4 評価実験

4.1 評価環境

文献 [3] のシステムを用い，4 人での 5 分間のグループミーティングを複数回撮影し，身振りや表情に対するラベルを付与したデータセット [4] を構築している．図 3 にグループミーティングの様子を示す．ラベルは，グループミーティング終了後に，各被験者に動画を見てもらいながら，自身の発話状況，うなずき，笑顔など複数種類のラベルを付与してもらった．グループミーティングは 4 人 1 組で各組 2 回ずつ，合計 8 回行って合計 16 人分のデータを収集した．

本稿では，このデータセットの中から 360 度カメラの映像と発話ラベルを用いて映像から発話量をどの程度抽出できるかについて検討する．具体的には，開口状態の推定精度及び発話量を評価した．開口状態推定

Measuring Speech Amount from Group Meeting Data
Kousuke Tokuhara*, Billy Dawton†, Yutaka Arakawa†, Shigemi Ishida†, Yusuke Soneda‡, and Yuki Matsuda‡
*EECS, Kyushu University, Japan
†ISEE, Kyushu University, Japan
‡Nara Institute of Science and Technology, Japan
*tokuhara.kosuke@arakawa-lab.com

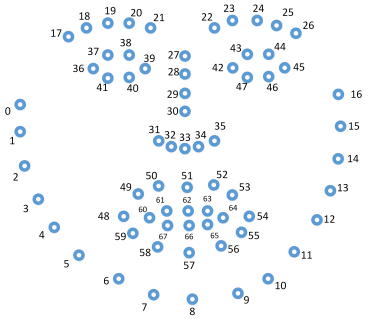

図 2: OpenFace により取得される顔のランドマーク点

図 3: 360 度カメラで撮影したグループミーティングの様子

精度の評価では Leave One Person Out（LOPO）検証を行い，精度として F 値を算出した．口を開いている時間割合は参加者によって異なるため，データをバランスさせた上で評価を行った．学習したモデルを用いて被験者の開口状態を推定して発話量を算出し，実際の発話量と比較した．

4.2 評価結果

表 1 に，開口状態推定結果及び発話量の推定結果を示す．表 1 より，被験者間のばらつきが大きく，推定発話量と実際の発話量の間にも大きな乖離があることが分かる．開口状態推定精度が低いために発話量の推定に大きな誤差を生じたと考えられる．

また，開口状態の推定精度と発話量の推定精度には明確な相関は確認できない．発話量は開口状態の割合のみから算出しているが，実際の発話は開口と閉口が連続して変化する状態であるため，高精度化に向けては開口状態の変化を用いて発話量を推定する必要があると考えられる．

開口状態の推定精度も十分に高いとは言えない．本稿では唇周辺の座標のみを使用して開口状態を推定したが，評価に用いたビデオを確認するとカメラに対する顔の角度の変化によって特徴量として用いている唇間の距離が変化することが確認できる．このため，さらなる高精度化に向けては顔の角度の変化を考慮することが必要であると考えられる．

表 1: 解析結果

被験者	開口状態	発話量 [%]		
	推定精度 [%]	推定	真値	誤差
1	49.9	38.8	57.8	23.9
2	56.5	18.4	44.6	26.2
3	57.2	42.5	55.5	13.0
4	64.6	10.5	37.0	26.5
5	47.1	34.9	60.0	25.1
6	63.2	5.7	39.0	33.3
7	46.3	32.0	49.9	17.9
8	52.7	21.4	46.7	25.3
9	50.6	40.4	58.9	18.5
10	60.0	15.8	40.5	24.7
11	57.4	18.6	46.5	27.9
12	71.5	30.8	22.1	-8.7
13	50.3	33.1	61.3	28.2
14	59.4	11.5	47.9	36.4
15	51.0	34.4	62.2	27.8
16	60.6	11.5	46.7	35.1

5 おわりに

本稿では 360 度カメラを用いて取得した動画から参加者の開口状態を推定することでグループミーティング時の発話量を推定する手法を提案した．実際にグループミーティングを行って収集したデータを用いて発話量を推定した結果，絶対平均誤差 24.9％で発話量を推定できることが確認できた．開口状態と発話の関係性が不明確であること，カメラに対する顔の動きによる座標の変化と発話による座標の変化を区別できていないことから十分な発話量推定精度を得られていないと考えられるため，動画内での参加者の動きを考慮した特徴量抽出を行うなどの拡張を行う予定である．

謝辞

本研究は，大阪大学グランドチャレンジ研究により大阪大学ライフデザイン・イノベーション研究拠点から委託されたものです．

参考文献

[1] Z. Yu, Z. Yu, H. Aoyama, M. Ozeki, and Y. Nakamura: Capture, recognition, and visualization of human semantic interactions in meetings. In *2010 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 107–115 (2010).

[2] 中島大一, 駒谷和範, 佐藤理史: 複数人会話システムにおける複数の音源定位結果の統合による発話者の特定, IPSJ 第 74 回全国大会, Vol. 4, p. 3 (2012).

[3] Y. Soneda, Y. Matsuda, Y. Arakawa, and K. Yasumoto: Multimodal recording system for collecting facial and postural data in a group meeting. In *The 27th International Conference on Computers in Education (ICCE 2019)*, No. 153 (2019).

[4] Y. Soneda, Y. Matsuda, Y. Arakawa, and K. Yasumoto: M3B corpus: Multi-modal meeting behavior corpus for group meeting assessment, In *The ACM 7th International Workshop on Human Activity Sensing Corpus and Applications (HASCA2019)* (2019).