一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NLC2019-41(2020-02)

マルチモーダル情報を考慮した議論の取りまとめ役推定

本多　幸希† 　塩田　　宰† 　嶋田　和孝† 　齊藤　剛史†

† 九州工業大学
E-mail: †{k_honda,t_shiota,shimada}@pluto.ai.kyutech.ac.jp, ††saitoh@ces.kyutech.ac.jp

あらまし　議論において参加者が担った役割を推定することは重要である．議論のような対人コミュニケーションでは発言だけでなく動作や声のような非言語的ふるまいも用いられることから，マルチモーダル情報を考慮することで役割推定における精度向上が期待される．先行研究では，参加者の発話から得られる特徴量を用いることで議論のコントロールを行う取りまとめ役を推定した．本研究では映像から動作特徴量および韻律特徴量を抽出し，マルチモーダル化によって取りまとめ役推定における精度向上を図る．先行研究に本手法の特徴量を素性として組み合わせ，決定木分類モデルによる推定精度の比較を行うことでマルチモーダル化の有効性を検証する．
キーワード　複数人議論，マルチモーダル，役割推定，取りまとめ役

Facilitator Identification Using Multimodal Information in Multi-party Conversation

Kouki HONDA†, Tukasa SHIOTA†, Kazutaka SHIMADA†, and Takeshi SAITOH†

† Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
E-mail: †{k_honda,t_shiota,shimada}@pluto.ai.kyutech.ac.jp, ††saitoh@ces.kyutech.ac.jp

Abstract　Predicting roles of participants in a conversation is one of the most important tasks for conversation understanding. In this paper, we propose a prediction model of facilitators in conversations. A previous study has handled verbal features for the facilitator prediction task. However, non-verbal features are also important elements for the prediction. Therefore, we introduce non-verbal features into the prediction model. We utilize two types of non-verbal features: motion features and voice features. The motion features are based on face information and body information. The voice features are based on prosodic information. In the experiment, we compare our model with a verbal feature-based model. The experimental result shows the effectiveness of our method.
Key words　Multi-party Conversation, Multimodal, Role Recognition, Facilitator

1. はじめに

議論とは主義主張の異なる人々が複数人で行う意思決定や意見交換の場である．議論は円滑に進むことが望ましいが，主張された多くの選択肢や判断基準の存在，あるいは発言力や参加態度の違いなどから円滑には進まない状況が考えられる．本研究室では，議論促進を目的として各参加者がどのような意見を持っているかフィードバックする合意形成支援システムの構築を行ってきた [1]．このシステムは，合意候補となる各選択肢を比較し収束させる議論中盤から終盤にかけてを支援対象としている．

また，議論終了後に着目すると，各参加者の議論中のふるまいに対し改善すべき点を提示することで次回以降の議論の円滑化を促すことができる．ふるまいに対する改善案の例として「あまり主張しない受動的な参加者に対し，積極的な主張を促す」のようなものが挙げられる．一方で，「他の参加者の発言を促し議論を盛り上げた参加者に対し，積極的な主張を促す」のような改善案は不適切である．議論中でのふるまいと関わりのない改善案を提示されても改善に繋がらないため，改善案はターゲットとなる参加者の理想的なふるまいを提示しなければならない．加えて，議論において参加者は何らかの役割を担うため，その役割によって理想的なふるまいが異なると考えられる．例えば，複数人による対話において各参加者は議論をコントロールする参加者と従う参加者に分かれる傾向がある．前者においてはどれだけ議論を円滑に進行させ他者の意見をまとめたかが重要であり，後者においてはどれだけ議論に参加し自身

の意見を提示できたかが重要である．

本研究では，議論をコントロールする役割を「取りまとめ役」とする．この取りまとめ役は参加者のアンケートに基づき決定する．先行研究として，塩田ら [2] は参加者の発話から得られる特徴量を用いることで，4 人 1 組による 20 分間の議論を対象に取りまとめ役の推定や分析に取り組んだ．先行研究では議論中の参加者の身振り手振りや声のような非言語的ふるまいは考慮されていなかったが，対人でのコミュニケーションではそのような言葉以外の情報も巧みに用いられており，非言語特徴量を用いることで推定精度向上が見込める．そこで，本論文では先行研究では用いられていない動作特徴量や韻律特徴量を抽出する．また，取りまとめ役分類モデルの素性に組み合わせることで取りまとめ役の推定精度向上に取り組む．

2. 関 連 研 究

古くから，議論における役割には関心が持たれている．Benne ら [3] は，議論に限らず一般的なグループワークにおける参加者を 3 つの大分類と 28 つの具体的な役割に当てはめ，Functional roles (機能的役割) として定義した．また，Zancanaro ら [4] は Functional Role Coding Scheme を提案し，[3] で定義された機能的役割を対面での複数人議論に対応させ，タスク領域と社会的・感情的領域という 2 種類の役割を定義した．タスク領域には議論において取り組むべきタスクの解決を行う 5 つの役割が含まれ，社会的・感情的領域にはグループがグループとして成り立つように参加者同士の関係に焦点を当てる 5 つの役割が含まれる．このような機能的役割は議論中で動的に変化すると考えられ，秒単位でのミクロな観点で推定する研究が行われてきた [4] [5]．また，Huang ら [6] は機能的役割を再定義し，発話の始まりと終わりを単位に推定を行っている．本研究は議論全体を通した参加者のふるまいに着目しており，対話単位というマクロな観点から役割の推定を行っている点が異なる．

特に社会的・感情的領域の役割に着目し，推定を行う研究も行われている [7]〜[9]．Wilson ら [8] は言語活動や韻律，話者の主観性を用いており，その組み合わせで推定精度が向上することを示した．しかしながら，用いられた特徴量は限られたものであり，より多くの特徴量を用いることで更なる精度向上が期待される．

また，具体的な役割ではなくリーダーシップに着目した研究もある．このような研究では，グループリーダーとしての役割が与えられたわけではなく，議論中自然とリーダーとしてふるまった参加者である「Emergent Leader (EL)」を推定する問題に取り組んでいる．Sanchez-Cortes ら [10] は ELEA コーパスの収録と，EL としてふるまった参加者を推定した．Beyan ら [11] は最も EL らしい参加者とそうでない参加者を推定した．本研究で対象としている議論の取りまとめ役は，EL のようにリーダーシップを発揮した参加者とみなすことができる．一方，これらの研究では非言語特徴のみを用いており，リーダーの発話に表れる特徴を考慮していない．

多くの研究が複数人による対話における参加者の役割を明らかにしようとしているが，議論全体を通してふるまわれた特定の役割を発話特徴・非言語特徴の両方を用いることで推定する研究は少ない．

3. データセット

本研究では複数人対話コーパスである Kyutech コーパス(注1) [12] を対象とする．このコーパスは 4 人の話者による意思決定対話を収録している．4 人の参加者は架空のショッピングモールの経営者という役割で対話に参加する．そのショッピングモールのレストラン街にある既存レストランが閉店するため，新規出店するレストランを 3 つの候補店から 1 つ選択するというタスクについて議論する．議論を行うために，3 つの候補店の情報，閉店する既存レストランの情報と閉店理由，他の既存レストランの情報，ショッピングモールの情報，ショッピングモールのある市の人口などの統計情報，隣接する町や市の統計情報の情報などが書かれた資料が参加者には渡される．参加者はこの資料を 10 分間黙読した後 20 分間の議論を行い，出店するレストランを 1 つ決定する．4 つのシナリオが準備されており，同じシナリオを議論しないよう 16 名の学生から 4 名の参加者が選ばれ，計 9 対話収録されている．対話収録後には，参加者を対象に「誰が議論をコントロールしていたと思いますか？」や「最終決定に対して最も重要な発言をしたのは誰だと思いますか?」などのアンケート調査も実施されている．

各発話の書き起こしデータには話者 ID，時間情報，人手でアノテーションされたトピックタグと談話行為タグが付与されている．トピックタグには，候補店に言及している発話に付与される「Cand」や閉店したレストランに言及している発話に付与される「Closed」など，計 28 種類のタグが存在している．談話行為タグには，情報要求を行う発話に付与される「Question (QU)」や情報提供を行う発話に付与される「Inform (Inf)」など，計 22 種類のタグが存在している [13]．また，本研究では書き起こしデータの発話区間を [13] にて用いられた基準である長い発話単位に変換したものを用いる．長い発話単位とは，「話し手と聞き手が行為や情報を交換する際の基本単位に相当し，統語的・談話的・ 相互行為的な一まとまり」である．

対話の収録において 4 名の参加者はテーブルを囲む形で着席して議論を行っており，各参加者の議論の様子を 360 度カメラによって撮影した映像が収録されている．今回は，本映像を用いることで非言語特徴量の抽出を行う．

4. 提 案 手 法

まず，4.1 節にて先行研究で用いられた発話特徴量を説明する．本研究では，この発話特徴量を素性として用いたモデルをベースラインとする．また，4.2 節にて議論の様子を撮影した映像から新たに抽出した動作特徴量および韻律特徴量を説明する．ベースラインに対し動作特徴量および韻律特徴量を組み合わせることによって推定精度向上を目指す．図 1 に提案手法の全体像を示す．

(注1)：http://www.pluto.ai.kyutech.ac.jp/~shimada/resources.html

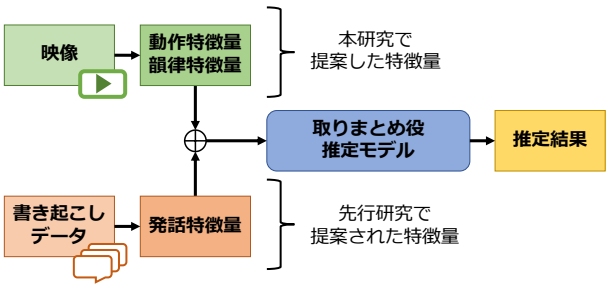

図 1　提案手法の全体像

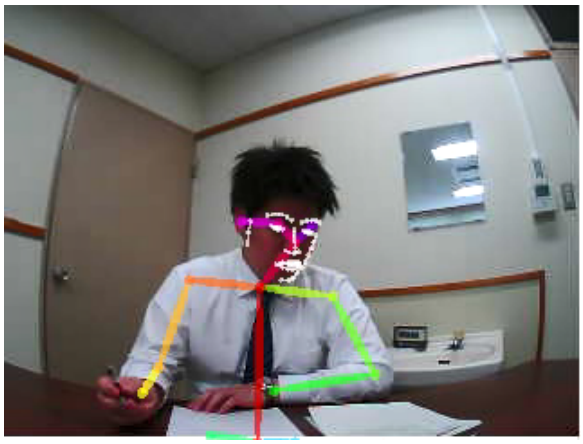

図 2　OpenPose による解析例

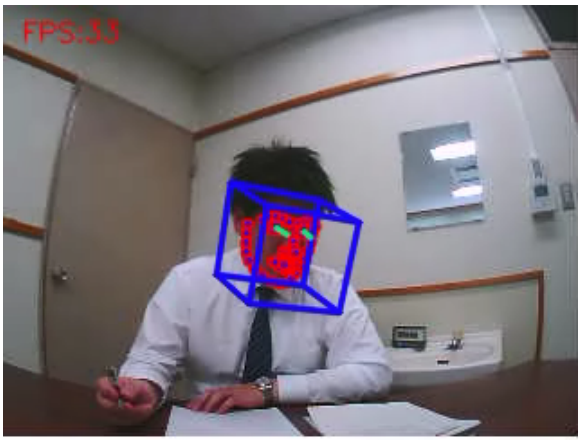

図 3　OpenFacse による解析例

4.1　ベースライン

先行研究 [2] で用いられた，書き起こしデータから得られた発話特徴量を B_1〜B_7 に示す．

B_1:　自身および他者の発話の繰返しの相対的割合

取りまとめ役は他の参加者の発言を繰り返すことで情報を引き出し，非取りまとめ役は自身の発言を繰り返すことで情報をわかりやすく伝えようとすると考えられる．まず対話中に出現する内容語を抽出し，各発話を内容語の頻度ベクトルに変換する．変換した各発話を基準として後ろに続く 10 発話の中で cos 類似度が 0.6 以上となる発話を繰返しとして抽出する．繰返し元の発話が同じ話者によるものなら「自身の発話の繰返し」，異なる話者によるものなら「他者の発話の繰返し」として扱い，それぞれの総数に対する各話者の繰返し数の割合を算出する．

B_2:　対話中のトピックを網羅した割合

取りまとめ役は対話内に存在するそれぞれのトピック内で発言すると考えられる．そこで，対話の全てのトピックを抽出し，各話者が各トピックにおいて発言したか否かを抽出する．そこから全トピック数に対する各話者の発言したトピック数の割合を算出する．

B_3:　Meeting タグの発話の相対的割合

取りまとめ役は議論の進行に関するトピックではより多く発言すると考えられる．そこで，Kyutech コーパスに存在するトピックタグの一つである Meeting タグの発話を抽出し，その総数に対する各話者の発話数の割合を算出する．

B_4:　特定の談話行為の発話の相対的割合

取りまとめ役は他の参加者から意見をくみ取る，他の参加者の意見を傾聴するなどの行為を非取りまとめ役より多く行うと考えられる．そこで，各談話行為の発話数に対する各話者の発話数の割合を算出する．本研究では Kyutech コーパスにおける「QU (情報の要求)」「An (情報要求に対する返答)」「Inf (情報の提供)」「Su (行為の提案)」「PF (肯定)」の 5 つの談話行為タグについて算出する．

B_5:　対話全体および 4 分割した中での発話の相対的割合

取りまとめ役は議論のコントロールのため非取りまとめ役よりも発話量が増えると考えられる．また，議論の状況によって発話するタイミングが決まっていることが考えられる．そこで，対話全体の発話数に対する各話者の発話数の割合，および対話を時間で 4 つに分割したときのそれぞれの発話の総数に対する各話者の発話数の割合を算出する．

B_6:　平均発話文字数・平均発話時間

取りまとめ役は議論の内容をまとめるなど非取りまとめ役よりも発話が長くなると考えられる．そこで，各話者の発話から平均発話文字数および平均発話時間を算出する．

B_7:　沈黙後に行われた発話の相対的割合

取りまとめ役は議論が停滞した場合，再びその議論を活性化させる発言をすると考えられる．そこで，一つ前の発話から 10 秒以上経過してから行われた発話を対話から抽出し，その総数に対する各話者の発話数の割合を算出する．

4.2　動作特徴量および韻律特徴量の抽出

取りまとめ役は身振り手振りのような動作に特徴が現れると考えられることから，議論参加者の動作特徴量を抽出する．動作の抽出には映像から骨格検出を行うことができる OpenPose [14] を用いる．OpenPose を用いた解析例を図 2 に示す．OpenPose によって各参加者の上半身から各フレームにおける鼻・両目・両耳・首・両肩・両肘・両手首といった体の特徴点を推定し，12 部位全てについて P_1，P_2 の動作特徴量を抽出する．

P_1:　各部位の X 座標および Y 座標それぞれの標準偏差

取りまとめ役は議論中長時間資料を見るなど同じ姿勢をとるわけではなく，必要に応じて手を動かす，周りを見渡すなど体の特定部位を大きく動かすと考えられる．すなわち，身体動作

によって特徴点のばらつきが大きくなることが考えられる．各部位の特徴点の座標を用いて X 座標および Y 座標それぞれの 1 対話中の標準偏差を算出する．

P_2: 各部位の移動量の平均値および標準偏差

座標の標準偏差のみではどの程度動いたかという動作のメリハリがわからないため，参加者の動作量に着目する．議論全体を通して各部位の特徴点のフレーム間移動量を計算し，1 対話中の平均値および標準偏差を算出する．

取りまとめ役は発言を促すために周りを見渡す，特定の参加者の方向を向く，といった顔の動作に特徴が現れると考えられるが，OpenPose では部位の座標推定ができるのみであった．そこで，顔情報の推定を行うことができる OpenFace [15] を用いる．OpenFace を用いた解析例を図 3 に示す．OpenFace によって各フレームにおける参加者の顔と目の特徴点，視線方向，頭の位置と向きを推定し，F_1〜F_6 の動作特徴量を抽出する．

F_1: 顔の X 座標および Y 座標それぞれの標準偏差

F_2: 目の X 座標および Y 座標それぞれの標準偏差

F_3: 顔の移動量の平均値および標準偏差

F_4: 目の移動量の平均値および標準偏差

上記と同様の理由から，各フレームごとに顔と目それぞれの特徴点を推定し，全特徴点の平均座標から 1 対話中の X 座標および Y 座標それぞれの標準偏差，フレーム間移動量の平均値および標準偏差を算出する．

F_5: 視線方向の標準偏差

取りまとめ役は議論中長時間同じ方向を見るのではなく，注視したり見渡したりするなど視線方向のばらつきが大きいことが考えられる．左目，右目それぞれの正規化視線方向ベクトルおよび両目で平均された視線角度について，1 対話中の標準偏差を算出する．

F_6: カメラに対する顔の位置と回転角の標準偏差

視線だけでなく顔の向きを考慮するため，カメラに対する顔の位置および XYZ 軸に対する回転角について，1 対話中の標準偏差を算出する．

取りまとめ役は声の大きさや高さに特徴が現れると考えられる．まず，書き起こしデータより得た発話区間にしたがって音声データを分割する．分割した音声データを発話断片とし，各発話断片から各参加者の音声に関する韻律特徴量を抽出する．特徴抽出には音声解析ライブラリである LibROSA [16] を用いる．抽出した V_1〜V_5 韻律特徴量を以下にまとめる．

V_1: 13 次元 MFCC

V_2: RMS

V_3: 基本周波数

V_4: スペクトル重心

V_5: スペクトルコントラスト

今回は各発話断片における最大値，最小値，最大値と最小値の差，標準偏差を算出し，1 対話中の全発話断片について平均したものを韻律特徴量として用いる．

5. 実　　験

Kyutech コーパスを対象に議論参加者を取りまとめ役・非取りまとめ役に分類し，2 値分類タスクを行うことで推定精度を評価する．ベースラインモデルに対し，どの特徴量セットを組み合わせることで推定精度が向上するか，マルチモーダル化が有効であるかを確認する．また，ベースラインを含まない組み合わせについても比較を行う．

表 1　推定実験の評価値

	取りまとめ役			非取りまとめ役		
	精度	再現率	F 値	精度	再現率	F 値
B	0.667	0.611	0.638	0.861	0.926	0.892
P	0.204	0.278	0.235	0.713	0.630	0.669
F	0.055	0.111	0.074	0.574	0.537	0.555
V	0.111	0.111	0.111	0.667	0.741	0.702
B+P	0.537	0.667	0.595	**0.889**	0.889	0.889
B+F	**0.704**	**0.722**	**0.713**	**0.907**	0.926	**0.917**
B+V	0.500	**0.667**	0.574	**0.880**	0.852	0.866
P+F	0.426	0.556	0.482	0.852	0.852	0.852
P+V	0.056	0.111	0.074	0.648	0.667	0.657
F+V	0.083	0.333	0.133	0.435	0.519	0.473
B+P+F	**0.778**	**0.833**	**0.805**	**0.935**	0.926	**0.931**
B+P+V	**0.722**	**0.722**	**0.722**	**0.907**	**0.963**	**0.934**
B+F+V	0.315	0.389	0.348	0.815	0.852	0.833
P+F+V	0.093	0.222	0.131	0.657	0.630	0.643
B+P+F+V	**0.889**	**0.833**	**0.860**	**0.935**	**1.0**	**0.967**

5.1 実 験 設 定

実験データには Kyutech コーパスの 9 対話 (全 36 話者) を用いた．正解データは Kyutech コーパスの被験者に実施されたアンケート「誰が議論をコントロールしていたと思いますか？」の回答で，2 人以上から指名された話者を取りまとめ役とした (取りまとめ役 10 名，非取りまとめ役 26 名)．また，評価値の比較を行うためシード値を固定し，分類学習には先行研究と同様に決定木 (CART) [17] を用いた．

5.2 実 験 結 果

表 1 に 9 対話交差検定によって得られた，決定木分類モデルによる取りまとめ役および非取りまとめ役の推定実験の評価値を示す．表中の「B」は B_1〜B_7 をまとめたものであり，「P」，「F」，「V」についても同様である．なお，各評価値は小数点第 4 位を四捨五入し，ベースラインモデルに対し評価値が向上したものは太字，評価値が最大となったものは下線で示す．

表 1 より，単一モダリティの特徴量セットを用いたモデルについて比較すると，動作特徴量，韻律特徴量をそれぞれ単体で素性として用いた「P」，「F」，「V」モデルの評価値はベースラインモデルよりも低かった．このことから，今回映像から抽出した動作特徴量，韻律特徴量単体では取りまとめ役の推定は難しいことがわかる．

次に複数モダリティの特徴量セットを用いたモデルについて比較する．ベースラインモデルに対し，「B+F」，「B+P+F」，「B+P+V」，「B+P+F+V」モデルの F 値が向上した．また，「B+P+F+V」モデルの評価値が取りまとめ役の推定および非取りまとめ役の推定において最大であった．複数モダリティの特徴量の組み合わせによって推定精度が向上したことから，取

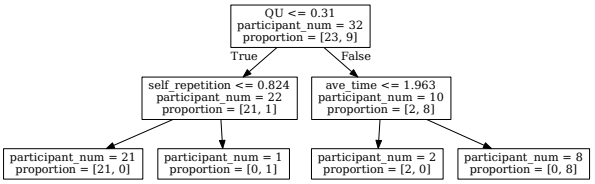

図 4　ベースラインモデルによる木の例

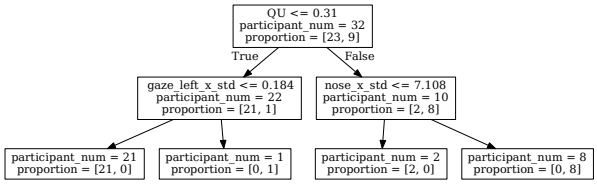

図 5 「B+P+F+V」モデルによる木の例

りまとめ役の推定においてマルチモーダル化の有効性が示唆された．一方，ベースラインモデルで用いた発話特徴量を含まない「P+F」,「P+V」,「F+V」,「P+F+V」モデルの評価値はベースラインモデルには及ばなかった．このことから，発話に現れる特徴を考慮せずに取りまとめ役を推定することは難しいことがわかる．

図 4 に推定実験にて得られたベースラインモデルによる木の例を示す．同様に，図 5 に「B+P+F+V」モデルによる木の例を示す．それぞれのノードには分類に用いられる素性名と閾値，ノードに属する参加者の数 (participant_num)，ノードに属する参加者の取りまとめ役 (括弧内右側) および非取りまとめ役 (括弧内左側) の割合 (proportion) が示されている．

図 4 の木では，

- 情報要求の発話の相対的割合 (QU)
- 自身の発話の繰返しの相対的割合 (self_repetition)
- 平均発話時間 (ave_time)

の 3 つの特徴が取りまとめ役の推定に用いられている．情報要求の発話の相対的割合の大きさは図 5 の木でも分類に用いられており，取りまとめ役推定に大きく貢献する特徴であることがわかる．一方，図 5 の木では，情報要求の発話の相対的割合が大きいことに加えて，

- 左目視線の X 軸方向の標準偏差 (gaze_left_x_std)
- 鼻の X 座標の標準偏差 (nose_x_std)

の 2 つの特徴が新たに取りまとめ役の推定に用いられている．用いられた動作特徴量はどちらも頭部に関するものであるが，視線については右目と左目が別の方向を向くとは考えづらい．加えて鼻だけが動くことは考えづらく，鼻が動いているということは顔全体が動いていると考えるのが妥当である．さらに，どちらも X 軸方向に関するものであったことから，左右方向の顔の動作が重要であることがわかる．以上のことから，議論をコントロールする役割を担う参加者は一方向のみを見ているわけではなく，特定方向を向いたり周りを見渡したりするなど，左右方向の頭部動作にメリハリをつけていることが示唆される．

先行研究では，同じく Kyutech コーパスを対象とした際，「他の話者から情報や行為，提案など言動を引き出す行為を相対的に多く行う」という取りまとめ役の特徴が得られた．本研究での結果と合わせると，そのような他の参加者に働きかける行為をする際に併せて頭部の動作が用いられていることが考えられる．

6. ベースラインの改良

5.2 節の結果より，取りまとめ役の推定において発話特徴量の必要性が示唆された．そこで，発話特徴量を追加することでベースラインモデルを改良し，更なる精度向上に取り組む．

6.1 追加した発話特徴量

先行研究では取りまとめ役のインタラクションに関する特徴が考慮されていなかった．Rienks ら [18] を参考に，新たに追加した B_1^*〜B_6^* の発話特徴量を以下に示す．以降「B」に対し B_1^*〜B_6^* を加えることで改良された発話特徴量をまとめて「B*」と略記する．

B_1^*:　トピックを切り替えた相対的割合

取りまとめ役は議論のトピックを制御しようとすると考えられる．そのため，書き起こしデータに付与されたトピックタグを利用して各参加者がトピックを変更した回数を数え，相対的割合を算出する．

B_2^*:　ターンテーキング回数の相対的割合

「ターン」とは少なくとも 1 単語を含む 1.5 秒以上の無音区間が存在しない同一話者の発話区間であり，「ターンテーキング」とは他者からターンが切り替わることである．取りまとめ役は議論のコントロールのため，他の参加者による発言と発言の間に発言すると考えられる．各参加者について他の参加者の発言の後に存在する発話の数，すなわちターンテーキングの回数を数え，各参加者の相対的割合を算出する．

B_3^*:　他者の発言を遮った相対的割合

B_4^*:　他者から発言を遮られた相対的割合

上記と同様に，取りまとめ役は他の発言を遮ることで議論のコントロールを図ると考えられる．他の参加者の発言を遮った回数および他の参加者から発言を遮られた回数を数え，各参加者の相対的割合を算出する．

B_5^*:　ラージターンテーキング回数の相対的割合

B_6^*:　ラージターンテーキングの平均時間

複数人議論においては各参加者が各発言に反応し，あまりにも多いターンテーキングが発生する．このことから，相槌 (7 文字以下の発話) を考慮しないターンテーキングを「ラージターンテーキング」と定義し，各参加者ついてラージターンテーキングの回数の相対的割合および平均時間を算出する．

6.2 実験結果

5.1 節の実験設定に従って，改良されたベースラインモデルを用いて 9 対話交差検定による分類実験を行った．表 2 に実験による評価値を示す．各評価値は小数点第 4 位を四捨五入し，

表 2　ベースラインを改良した場合の評価値

	取りまとめ役			非取りまとめ役		
	精度	再現率	F 値	精度	再現率	F 値
B	0.667	0.611	0.638	0.861	0.926	0.892
B*	**0.704**	**0.778**	**0.739**	**0.907**	0.889	**0.898**
B*+P	0.611	**0.722**	**0.662**	**0.907**	0.889	**0.898**
B*+F	0.426	0.611	0.502	**0.880**	0.815	0.846
B*+V	0.444	0.500	0.471	0.833	0.852	0.842
B*+P+F	0.500	0.611	0.550	**0.870**	0.852	0.861
B*+P+V	0.204	0.333	0.253	0.778	0.815	0.796
B*+F+V	0.500	**0.667**	0.571	**0.889**	0.889	0.889
B*+P+F+V	0.611	**0.722**	**0.662**	**0.898**	0.852	0.874
B+P+F+V	**0.889**	**0.833**	**0.860**	**0.935**	**1.0**	**0.967**

ベースラインモデルに対し評価値が向上したものは太字，評価値が最大となったものは下線で示す．

「B」モデルと比べて，「B*」モデルは取りまとめ役および非取りまとめ役の推定精度が向上した．このことから，追加した発話特徴量 B_1^*〜B_6^* が精度向上に貢献したことがわかる．一方，「B*」に対し他の特徴量を組み合わせた場合「B」モデルよりも評価値が向上することはあったが「B*」モデルの改善には繋がらず，表 1 の結果にて最高精度であった「B+P+F+V」モデルには及ばなかった．組み合わせによる精度低下の明確な原因分析については今後の課題とする．

7. おわりに

本研究では複数人議論における取りまとめ役の推定精度向上を目的に，特徴量抽出やマルチモーダル化に取り組んだ．まず，Kyutech コーパスを対象に映像から動作特徴量や韻律特徴量を抽出した．次に，決定木を用いて取りまとめ役および非取りまとめ役を推定するモデルを構築した．推定実験の結果，先行研究で用いられた発話特徴量に動作特徴量や韻律特徴量を組み合わせることで取りまとめ役および非取りまとめ役の推定精度が向上したことから，マルチモーダル化の有効性が示唆された．また，実験結果から取りまとめ役の非言語的な行動を分析し，左右方向の頭部動作に特徴があることが示された．

今回は取りまとめと非取りまとめ役の 2 値分類に取り組んだが，今後は非取りまとめ役を役割細分化し推定にも取り組んでいきたい．また，推定したうえでターゲットとなった参加者が相応しいふるまいをできていたか判断するモデルの構築に取り組んでいきたい．

謝辞　本研究は科研費 17H01840 の助成を受けたものです．

文　献

[1] R. Kirikihira and K. Shimada, “Discussion map with an assistant function for decision-making: a tool supporting consensus-building,” Proceedings of the tenth International Conference on Collaboration Technologies (CollabTech 2018), pp.3–18, 2018.

[2] T. Shiota, T. Yamamura, and K. Shimada, “Analysis of facilitators’ behaviors in multi-party conversations for constructing a digital facilitator system,” Proceedings of the tenth International Conference on Collaboration Technologies (CollabTech 2018), pp.145–158, 2018.

[3] K.D. Benne and P. Sheats, “Functional roles of group members,” Journal of Social Issues, vol.4, no.2, pp.41–49, 1948.

[4] M. Zancanaro, B. Lepri, and F. Pianesi, “Automatic detection of group functional roles in face to face interactions,” Proceedings of the 8th International Conference on Multimodal Interfaces, pp.28–34, 2006.

[5] W. Dong, B. Lepri, F. Pianesi, and A. Pentland, “Modeling functional roles dynamics in small group interactions,” Multimedia, IEEE Transactions on, vol.15, pp.83–95, 2013.

[6] H.-H. Huang，Q. Zhang，S. Okada，K. Kuwabara, and T. Nishida, “Adopting functional roles for improving participants’ communication skill in group discussion conversation，” Proceedings of the Group Interaction Frontiers in Technology，pp.1–9，GIFT’18，2018.

[7] F. Valente and A. Vinciarelli, “Language-independent socio-emotional role recognition in the ami meetings corpus,” INTERSPEECH, pp.3077–3080, 2011.

[8] T. Wilson and G. Hofer, “Using linguistic and vocal expressiveness in social role recognition,” Proceedings of the 16th International Conference on Intelligent User Interfaces, pp.419–422, 2011.

[9] A. Sapru and H. Bourlard, “Automatic social role recognition in professional meetings using conditional random fields,” Proceedings of Interspeech, pp.1530–1534, 2013.

[10] D. Sanchez-Cortes, O. Aran, M. Mast, and D. Gatica-Perez, “A nonverbal behavior approach to identify emergent leaders in small groups,” IEEE Transactions on Multimedia, vol.14, pp.816–832, 2012.

[11] C. Beyan, N. Carissimi, F. Capozzi, S. Vascon, M. Bustreo, A. Pierro, C. Becchio, and V. Murino, “Detecting emergent leader in a meeting environment using nonverbal visual features only,” Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp.317–324, 2016.

[12] T. Yamamura, K. Shimada, and S. Kawahara, “The Kyutech corpus and topic segmentation using a combined method,” Proceedings of the 12th Workshop on Asian Language Resources, pp.95–104, 2016.

[13] T. Yamamura, M. Hino, and K. Shimada, “Dialogue act annotation and identification in a japanese multi-party conversation corpus,” Proceedings of the Fourth Asia Pacific Corpus Linguistics Conference, pp.529–536, 2018.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields,” 2018.

[15] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L. Morency, “Openface 2.0: Facial behavior analysis toolkit,” 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), pp.59–66, 2018.

[16] BrianMcFee, ColinRaffel, DawenLiang, DanielP.W. Ellis, MattMcVicar, EricBattenberg, and OriolNieto, “librosa: Audio and Music Signal Analysis in Python,” Proceedings of the 14th Python in Science Conference, pp.18–24, 2015.

[17] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen, Classification and regression trees, CRC press, 1984.

[18] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, “Detection and application of influence rankings in small group meetings,” Proceedings of the 8th International Conference on Multimodal Interfaces, pp.257–264, 2006.