グラフニューラルネットワークを用いた多人数会話における発話バランス推定の試み

A Study on Speech Balance Estimation in Multi-Party Conversations Using Graph Neural Networks

吉井　一総 [1*]　吉田　直人 [2]
Kazusa Yoshii[1]　Naoto Yoshida[2]

[1] 工学院大学大学院
[1] Graduated School of Informatics, Kogakuin University
[2] 工学院大学
[2] Kougakuin University

Abstract: 本研究では，多人数対話における話者関与の度合いは定性的で教師信号を得にくい点に着目し，参加者の正面映像から顔向き・身体方向・発話状態から疑似ラベルを生成して，GNN の注意機構を自己教師ありで学習させた．検証に際して対話映像が収録されたデータセットを用い，正解値として俯瞰映像から関与の度合いを 5 段階で主観的に評価させた．比較にはこの主観評定値と重み付き隣接行列の行和を代表値に用い，それぞれを順位化して一致率を比較した．結果として最下位の一致率は 42.3%であり，消極的発話者支援への活用可能性が示唆された．

1　はじめに

グループディスカッションや，アクティブラーニングといった多人数会話において,参加者は主体的な参加を求められることが多い．しかしながら対話の中で対話者の積極性を判断するとき,「空気を読む」というような直感的な判断に頼る場合が少なくない．この場合においては対話を聞いていたにも関わらず，エンゲージメントの少なさから消極的だと判断されてしまう可能性さえ考えられる．対話をはじめとした社会的相互作用を計算論的なアプローチで分析する領域に社会的信号処理が挙げられ [1]，マルチモーダル信号を入力に人間の感情や関係性などを推定する研究が行われている．酒井ら [1] はこの分野の推定対象について，アノテーションや回答には主観的な判断が含まれるため正解が一意に定まらないことを指摘している．また，感情推定について熊野 [2] は主観評定値を用いた教師データは時間的，または認知的な負荷がかかることを指摘している．つまり，計算的なアプローチを用いて話者間の積極性を推論しようとするとき，それは定性的であるために入力特徴に適さないという課題がある．

対話における古典的な議論には F 陣形があり，頭部方向や身体方向などが人がどのような活動に参加を試みているかの手がかりになるとされている [3]．加えて，グループダイナミスの研究として発話のオーバーラップが発話の盛り上がりに寄与することが指摘されており [4]，多人数対話における参加者の動向を探るモダリティには身体方向と発話の共起性などが重要であると考えられる．以上より，本研究ではセンサから得られる身体特徴や音声特徴を利用して，多人数会話内の参加者における積極性をはじめとした発話バランスの推定を試みる．発話バランスとは参加者間の関係性の強さを指し，多人数会話に現れる定性的特徴を計算的に捉えることを目的とする．

2　関連研究

2.1　多人数会話に着目した支援システム

関係性の把握に当たり，話者の属性に焦点を当てた多人数会話を分析対象とするシステムは数多く提案されている．石井ら [5] は発話タイミングの予測を注視の遷移パターンに基づく遷移確率を算出するナイーブベイズモデルでの予測を行っている．さらに，森ら [6] は韻律・言語・視線情報を用いた Feed Forward Network(FFN) を用いた予測システムを提案している．これらのシステムは多人数会話の中で発話交代の中で生じる話し手・聞き手に着目した研究である．また，会話集団の検出

*連絡先：工学院大学大学院工学研究科情報学専攻
〒 192-0015 東京都八王子市中野町 2665-1
E-mail: em25038@ns.kogakuin.ac.jp

に着目した研究 [7] では，GNN を用いて F 陣形のクラスタリングを行い，既存手法よりも個人の特徴を捉えやすくなっていることを報告している．対話における役割は相互関係によって定まるものの，話し手や聞き手などを含めた包摂的な関係性を定量的に推測する研究は少ない．したがって本研究では対話における構造を直接的に表現できるグラフ理論に着目し，推論を行うにあたり GNN を活用した．

2.2 社会低関係の推定

計算的に社会的関係を把握する研究も行われている．中田ら [8] はモーションキャプチャによる身体動作と頭部運動や，マイクによる発話の有無の取得から会話状況に参加する積極性を推定し，重回帰分析でのモデル化を試みている．また，岡田ら [9] はコミュニケーション能力の高低についてサポートベクターマシンを用いた分類器を提案している．近年では Transformer を用いた対話における心象推定も取り組まれているが [10]，個人の状態を直接的に表現できる研究はいまだ少ない．本研究は会話参加者を各グラフの頂点としてみなすことで，多人数会話における個人間の関係の表現を試みるものである．

3 提案方法

3.1 システム概要

本システムはグラフニューラルネットワーク（GNN）の注意機構を用いて，多人数会話における各参加者の発話バランスを推定することを試みた．入力特徴として，MediaPipe(0.10.32)[1]による身体特徴と，FFMpeg(8.0)[2]を利用した音声特徴を組み合わせた疑似ラベルで疑似隣接行列を作成した．学習はこの隣接行列をソフトターゲットとみなし，GNN の注意重みに対するカルバックライブラー距離 (KLD) が最小化するように自己教師あり学習を行わせた．学習で得られた重み付き隣接行列を各時刻ごとに計算することによって，多人数会話における話者間の積極性を連続的に表現することを試みた．なお，積極的であるほど注意を向けられやすくなると考えたため，得られた隣接行列における行方向の総和を推定結果の代表値として用いた．

3.2 特徴量の抽出と作成

MediaPipe を用いて，着席して対面している会話参加者の正面映像から頭部方向と身体方向の 3 次元ベクトルを取得した．頭部方向は 15 点の顔メッシュ画像によって得られたオイラー角をカメラ中心 (0, 0, 1) を基準として 3 次元の方向ベクトルに変換させ，身体方向は両肩の二点の中心を直交するベクトルとして算出した．さらに，音声は FFMpeg を用いて映像から音声を抽出して発話の有無を二値化した．以上の特徴量を各参加者ごとに取得したうえ，姿勢に関しては座席配置をもとにして実世界での姿勢と一致するように角度の補正を行った．これらをもとに，入力特徴は頭部方向，身体方向が向き合う一致度の指標 $H_{ij}(t)$, $B_{ij(t)}$，発話の共起性に S_{ij} を定め，これらの指標の線形和を疑似ラベル隣接行列 $W_{ij}(t)$ の要素として設定した．また会話を連続的に捉えるにあたり窓幅とストライドを時間窓として導入し，各指標は時間窓ののスライドごとにさせた．

$$H_{ij}(t) = 1 - H_i(t) \cdot H_j(t) \tag{1}$$

$$B_{ij}(t) = 1 - B_i(t) \cdot B_j(t) \tag{2}$$

$$S_{ij}(t) = S_i(t) \cdot S_j(t) \tag{3}$$

$$W_{ij}(t) = \alpha \cdot H_{ij}(t) + \beta \cdot B_{ij}(t) + \gamma \cdot S_{ij}(t) \tag{4}$$

3.3 学習方法

頂点間の注意重みをもつ隣接行列 A_ij は，疑似ラベル隣接行列 W_{ij} との KLD によって学習させた (式 (5))．式 (6), 式 (6) は GNN の注意重みを更新する式であり，h_i は頂点の特徴ベクトル，$e_{ij}^{(h)}$ が正規化前の信号ベクトルである．$\alpha_{ij}^{(h)}$ は正規化された学習可能な注意ベクトルである．なお，$e_{ij}^{(h)}$ が正規化前の近傍ノードへの信号ベクトルである．これらの式を用いた理由には，身体特徴や音声特徴などのモダリティは発話のバランスを推定する上での主要な特徴であるものの，これらは主観評定値を十分に代替するものであるとは考えにくい．そこで KLD を用いた確率分布の主要特徴の傾向のみを学習させることで，主観評定値を代理させた学習が可能になるのではないかと考えられる．また，疑似ラベル生成で用いた各特徴を，頭部・身体・音声の 3 次元の埋め込みエッジ特徴としても利用した．したがって主観的評定値の代理としての疑似ラベルと，実際の特徴としてのエッジ埋め込みを併用した．なお，学習に際して初期状態のグラフで関係性が推測できないため，完全無向グラフのグラフの状態から疑似ラベルによる学習を行い，注意重みの大きさの変動が多人数会話の関係性を表現するものであると仮定した．

$$\mathcal{L}_{\mathrm{KLD}} = KL\Big(\mathrm{softmax}\Big(\frac{\mathbf{A_{ij}}}{\tau}\Big)\Big\|\mathrm{softmax}\Big(\frac{\mathbf{W_{ij}}}{\tau}\Big)\Big) \tag{5}$$

$$\alpha_{ij}^{(h)} = \mathrm{softmax}_j\big(e_{ij}^{(h)}\big) \tag{6}$$

[1]https://pypi.org/project/mediapipe/
[2]https://www.ffmpeg.org/

$$\mathbf{h}_i^{(l+1)} = \underset{h=1}{\overset{K}{\mathrm{concat}}}\,\sigma\Big(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(h)}\,\mathbf{W}^{(h)}\mathbf{h}_j^{(l)}\Big) \quad (7)$$

4 検証および結果

4.1 検証手続き

本検証では，GNN の注意重みが多人数会話の参加者が持つ積極性（以下，参与度）を代替するとして，主観評価における参与度を正解データとした比較を行った．提案手法の検証にあたり，NII が提供するグループコミュニケーションコーパス (TDU-NEDO) を利用した [11]. このコーパスには 6 名によるグループディスカッションの様子とそれに伴う発話内容，手の動き，視線の動きなどのアノテーションデータが記録されている．本検証では映像の中から連続していない 5 分間の映像を 2 つ抽出して，それぞれ学習用と検証用に用いた．このコーパスに記録されている参加者 A E のうち，B を除く 5 名の正面のカメラ映像から，30Hz の標本周波数で頭部方向と身体方向の 3 次元ベクトルを取得した．また，音声はカメラ映像内の音声を用い発話の有無に関して二値化を行った．さらに窓幅，ストライドはそれぞれ 90, 15 に設定し，映像の全域を 600 区間のうち時間窓の境界効果を除くため冒頭のデータを除去した 595 区間について疑似隣接行列を作成した．ハイパーパラメータの値は探索的に設定し，$\alpha = 0.1, \beta = 0.3, \gamma = 0.6$ に，KLD における変数 τ は 2 とした．また，学習のエポック数は 20 で実施した．

主観評価の方法については，システムに用いる抽出した 5 分の動画と同区間の俯瞰映像を重複しない 5 秒ごとの区間に分割して，各区間の参加者ごとに参与度をリッカート尺度で評価させた．事前の教示として「参与度とは，発話や体の動きなどを手がかりに，議論の様子を外から観察して，どの程度議論に貢献しようとしているかを示す総合的な指標」という説明を行い,「1: まったくそうは思わない 5: とてもそう思う」とした 5 段階での評価を行った．評定者は人事採用経験のない大学生 3 名であり，集計後に平均した値を主観評定値に定めた．評定者間の回答の相関係数は 0.81 であったため，回答者間でのブレは少ないと考えられる．なお，主観評価においては正面映像が提供されていない参加者についても行ったものの，提案システムの評価に際して利用できないため除外した．

検証には上記の提案システムによる評定値と，主観評価による回答をそれぞれ順位化した 60 区間のうち，冒頭を除いた 59 区間について，top-k 一致率による評価を行った．

4.2 結果

結果は表 1 のようになった．行の要素が主観評価における評定値を 1 位から 5 位まで順位化したもの，列の要素は推定値を順位化したものである．ベースラインをランダムとする 20%と定めたとき，1 位から 3 位までの上位の結果はベースラインの値に近いものや，下回るものが見られた一方で，4 位，5 位の下位においては上回る結果となった．また，全セグメントを結合して主観評価と推定値の相関係数は $\rho = 0.33 (n = 295)$ となった．

表 1: 順位化した主観評定値と推定値の一致率

	1 位	2 位	3 位	4 位	5 位
1 位	14.0%	31.3%	29.7%	15.6%	9.4%
2 位	35.0%	20.0%	17.5%	15.0%	12.5%
3 位	24.0%	28.0%	22.0%	16.0%	10.0%
4 位	11.9%	11.9%	10.2%	33.9%	32.2%
5 位	7.1%	4.8%	21.4%	23.8%	42.3%

5 考察

上位の一致率が低いことを受け，推定値が注意重みとしてどのような特徴を持つのかを探索するにあたり，コーパスに収録されていた各参加者の視線の動きの回数を数えた．表 2 は参加者自身が視線を向けた注意の回数と，視線を向けられた被注意の回数を割合にしたものである．これより，参加者 C が注意の回数と被注意の回数の両指標において最小であることが明らかになった．また．注意に関しては参加者 D や F の割合が高く，被注意に関しては参加者 E や A の割合が高いことが伺えた．

表 2: 検証区間における各参加者の注意と被注意の回数の割合

	A	C	D	E	F
注意 (%)	19.7%	10.3%	35.0%	12.1%	22.9%
被注意 (%)	22.9%	7.1%	10.0%	38.8%	21.2%

表 3: 注意・被注意と主観評定値・推測値における相関係数

	注意	被注意
主観評定値	0.28	0.41
推測値	0.25	0.03

さらに，これらの両指標について，4.1 節と同様に視線の動きを 60 区間に分割した中で各参加者の回数をカ

ウントした．主観評定値と推測値のそれぞれに対して，冒頭を除く 59 区間を結合 (n=295) させた相関係数を計算した結果を表 3 に示す．視線の被注意と主観評価に正の相関がみられる一方で，推定値の間では無相関であった．なお，注意と被注意の間の相関は $\rho = -0.0041$ であり相関はなかった．これは人事経験などを持たない評定者が議論における積極性を直感的に評価する際には，注目の集まりやすいその場における話し手を高く評価する傾向があったためであると考えられる．これは，平野ら [12] が行ったファシリテーションにおける初心者と熟達者の間の視線配布対象を分析した研究において．初心者は発言している参加者に向けた注意を向けていた報告と整合するものである．一方で，注意については推測値と主観評定値とともに弱い正の相関がみられた．推定値と被注意の間に相関がみられなかったことから，参加者の注意を優先的に表現しうる可能性が示唆される．表 1 において上位の一致率は低く，下位の一致率が高いのは，表 2 において高い割合の持つ話者は異なっていたものの，参加者 C が低い割合であったことが共通していたことが要因として考えられる．これらより，主観評定値と推定値は異なる特徴を指していた可能性があるが，0.33 もの相関を持つことは直感の延長線上に立って，評価対象や支援対象の認識を拡張する足掛かりとして有用になりえる．応用例として，本システムは能動的な働きかけが少ない消極的発話者を検知し，コミュニケーション支援を行う活用できる可能性が考えられる．

6　おわりに

本研究では，GNN における注意重みを用いて，他人数会話における発話バランスの推定を試みた．データセットを用いた検証の結果，議論での貢献を指し示す参与度を主観評定値とした top-k 一致率は 4 位で 33.2%，5 位で 42.3%とランダムベースラインである 20%を上回った．一方で 1 3 位の上位においてはランダムベースライン付近あるいは，それを下回る結果になった．したがって推定値は主観評定値を代替することは困難であるものの，消極的発話者に関しては検出できる可能性が示唆された．

本研究の限界として，単独の会話データセットから検証を行ったため，汎用的にこのような傾向が発見されるかは定かではない．したがって複数の会話データセットによる検証が必要である．また本研究の展望として，提案手法を組み込んだアプリケーションが実際に消極的発話者支援として機能しうることを確かめる必要がある．

謝辞

本研究は JSPS 科研費 23K11202，23K11278，22K19792 の助成を受けたものです．また，国立情報学研究所の IDR データセット提供サービスにより東京電機大学から提供を受けた「グループコミュニケーションコーパス（TDU-NEDO）」を利用しました．

参考文献

[1] 酒井元気, 岡田将吾, 近藤一晃, 湯浅将英, and 酒造正樹, “これからのコミュニケーション研究とは?,” in **人工知能学会全国大会論文集 第** *38* **回** *(2024)*, pp. 2R6OS13a01–2R6OS13a01, 一般社団法人 人工知能学会, 2024.

[2] 熊野史朗, “主観感情推定の研究動向,” **人工知能**, vol. 36, no. 1, pp. 13–20, 2021.

[3] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*, vol. 7. CUP Archive, 1990.

[4] 森田大樹 and 瀬島吉裕, “マルチスケール集団コミュニケーションにおける盛り上がり推定モデルの開発,” in **人工知能学会全国大会論文集 第** *39* **回** *(2025)*, pp. 2D4GS904–2D4GS904, 一般社団法人 人工知能学会, 2025.

[5] 石井亮, 大塚和弘, 熊野史朗, 松田昌史, and 大和淳司, “複数人対話における注視遷移パターンに基づく次話者と発話開始タイミングの予測,” **電子情報通信学会論文誌** *A*, vol. 97, no. 6, pp. 453–468, 2014.

[6] 森大河 and 伝康晴, “多人数会話におけるマルチモーダル聞き手反応予測,” in **人工知能学会研究会資料 言語・音声理解と対話処理研究会** *96* **回** *(2022/12)*, p. 02, 一般社団法人 人工知能学会, 2022.

[7] S. Thompson, A. Gupta, A. W. Gupta, A. Chen, and M. Vázquez, “Conversational group detection with graph neural networks,” in *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 248–252, 2021.

[8] 中田篤志, 角康之, 西田豊明, and 來嶋宏幸, “移動・動作に関するセンサデータからの多人数会話状況の解釈,” **人工知能学会全国大会論文集**, no. 0, pp. 91–91, 2008.

[9] 岡田将吾, 松儀良広, 中野有紀子, 林佑樹, 黄宏軒, 高瀬裕, and 新田克己, “マルチモーダル情報に基づくグループ会話におけるコミュニケーション能力の推定,” **人工知能学会論文誌**, vol. 31, no. 6, pp. AI30–E_1, 2016.

[10] 堅田俊, 岡田将吾, and 駒谷和範, “注意機構を用いた生体信号時系列と言語系列の 統合に基づく本人心象推定,” **人工知能学会論文誌**, vol. 40, no. 2, pp. B–O72_1, 2025.

[11] 東京電機大学 (2019), “グループコミュニケーションコーパス (TDU-NEDO).” 国立情報学研究所 情報学研究データリポジトリ. https://doi.org/10.32130/rdata.1.1.

[12] 平野智紀, 山内祐平, *et al.*, “ワークショップのファシリテーションにおける熟達者と初心者の視線配布傾向の比較 応用演劇ワークショップを例に,” **日本教育工学会論文誌**, vol. 47, no. Suppl., pp. 117–120, 2023.