K-018

座圧分布の時系列変化に基づく協調学習における話者推定
Speaker Recognition in Cooperative Learning Based on Time-Series Seat Pressure Distribution

江角 翼[1)]　武村 紀子[1)]
Tsubasa Esumi　Noriko Takemura

1　はじめに*

近年の学習改革の影響によりアクティブラーニングや PBL などの知識活用型の学習が注目されており，教育の場で協調学習を行う機会が増えている．しかし，協調学習において教員が全学生の状態を把握するのは困難であり，適切な声掛けや公正な成績評価が難しいといった問題がある．この問題を解決するために，誰がいつ，どの程度発話したかという情報は有益であると考える．話者や発話量が分かると，ディスカッションにおいて発話回数が多い学生を高く評価したり，発話回数が少ない学生をリアルタイムでフォローすることができる．話者推定に関する従来研究では，音声情報が用いられることが一般的であるが [1]，話者とマイクとの位置関係や環境ノイズなどにより安定した話者推定が困難なケースが見られる．さらに，音声を録音されることに対する抵抗感や会話内容の秘匿性の担保が難しい点など，プライバシーの問題が実システムへ応用する際の障壁となりうる．そこで，本研究ではプライバシー情報が少なく，安定した計測が可能な座圧センサに着目し，座圧情報を用いた話者推定を行う．

本研究で使用する座圧センサは図 1 のように座布団の形をしており，椅子に敷いて座るだけで着座面の圧力（計測点 16 × 16 = 256）が計測できる．座圧の値は着座姿勢をよく反映していると言われており，森田ら [2] は座圧情報を用いて 32 パターンの姿勢を F1-score=0.67 の精度で推定している．また，姿勢情報から協調学習の評価を行っている研究もあり，福井ら [3] はカメラ映像から抽出した骨格モデル情報を用いて，協調学習の活性度（Active/Passive）を F1-score=0.70 で推定している．カメラ映像から姿勢を推定する場合は，人物同士あるいは障害物による人物領域の遮蔽の問題や撮影する向きによる見えの違いの問題が存在するが，座圧センサについては遮蔽や撮影方向の問題がなく，安定した計測が可能であるという利点がある．以上より，座圧データを用いた協調学習時における話者推定は十分可能であると考える．

本研究では，協調学習時（ディスカッションタスク時）の座圧データを用いて，モデルの学習および評価を行う．まず，グループ学習時の音声データから話者ラベルを作成し，話者推定用の座圧データベースを構築する．その後，1 枚の座圧画像を入力とする VGG-16 モデル，および座圧画像列を入力とする 3D-CNN モデルを用いて話者推定を行い，話者推定性能の比較を行う．

2　データベースの作成

教師あり機械学習により話者推定モデルを構築するため，収集したグループ学習時の音声データから話者ラベルを作成し，座圧データに話者ラベルを対応付けることで，学習・評価用のデータベースを作成する．

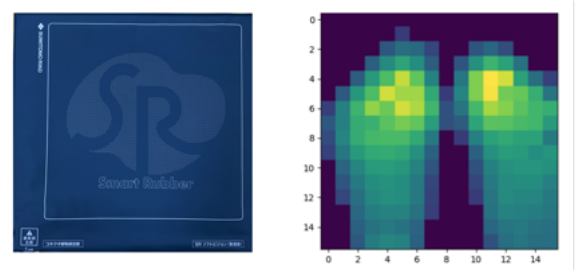

図 1　座圧センサと座圧画像．住友理工の SR ソフトビジョンを使用．

図 2　グループディスカッションの様子

2.1　収集データ

図 2 のように 4 名でグループディスカッションを行っている様子を web カメラ，座圧計およびピンマイクにより計測した．被験者は 20 ～ 40 代の男女 32 名（8 グループ）で，25 分間のタスクを 2 回ずつ，計 50 分行った．ディスカッションは付箋を用いてアイデアを整理する手法である KJ 法 [4] を用いて行われ，各タスクの最初 10 分はブレーンストーミング，後の 15 分はアイデアを整理する作業を行うものとした．ただし，ブレーンストーミングの最初の 5 分間は各被験者が付箋にアイデアを書き出す作業を行い，発話はしないため，本研究では使用しないものとする．

座圧計は 350mm×350mm の範囲内に 256 箇所 (16×16) の計測点があり，各計測点においてサンプリングレート約 5Hz で 20 ～ 200mmHg の範囲で圧力が計測可能である．また，被験者の襟元にピンマイクを付け，被験者ごとの音声データを収集した．Web カメラで撮影した映像については，話者推定には使用せず状況確認にのみ用いるものとする．

1）　九州工業大学

＊　本論文は画像の認識・理解シンポジウム MIRU2023 においてコンセプト論文として発表した内容に基づく

00:04:34 5
野菜の皮も食べるとかちょっと。
00:04:36 5
似てるんじゃないよ？
00:04:37 3
見てる気がするなあ、頑張って食べる。
00:04:38 1
ああ。

図 3　トランスクリプト機能による文字起こし．

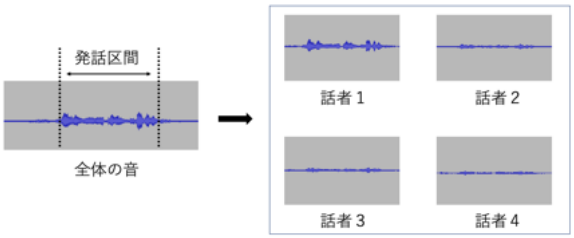

図 4　発話区間および話者の決定方法．図の例では，最も音量が大きい話者 1 を発話者とする．

2.2　話者ラベルの作成

ピンマイクから得られた音声データ（WAV ファイル）を元に，誰が発話しているかラベル付けを行う．収集した音声データには紙をめくる音や咳払いの音など様々なノイズが含まれる．そこで，まず最初に発話区間を切り出し，切り出した発話区間に対して 1 秒ごとに，各被験者のピンマイクの音量の大きさに基づいて話者を求める．

発話区間の切り出しには，Microsoft Word for the web のトランスクリプト機能を用いる．この機能では，音声データを入力すると，図 3 のようなタイムスタンプ付きの文字起こしが出力される．ここでは，4 人分の音声データを結合したものを入力とし，文字起こしが行われた区間を発話区間とみなす．なお，本機能では話者情報についても出力されるが，話者推定の精度が十分ではないため本研究では使用しない．

トランスクリプト機能で出力されるタイムスタンプは発話開始時刻のみであるため，図 4（左）のように全体の音量の大きさから閾値処理により発話区間がどこまで続いているかを判断する．1 秒単位で発話区間であるかどうかを判断し，発話区間である場合は図 4（右）のように区間内で最も音量が大きい人を発話者とする．ただし，マイクの取り付け位置や個人によってマイク音量は異なるため，各被験者について 25 分間のタスクごとにマイク音量の平均が 0，分散が 1 になるように標準化を行うものとする．

2.3　座圧データの補正

座圧計のサンプリングレートは一定ではないため，前後のデータを用いて線形補間を行うことで，サンプリングレートが 5Hz になるようにデータを補正する．図 5 にデータ補正の概略図を示す．補正したい時刻の前後で最も近い座圧データを $\hat{S}_j, \hat{S}_{j+1}$ とし，実際のサンプリング時刻と補正時刻との間隔の比が m ： n のとき，補正

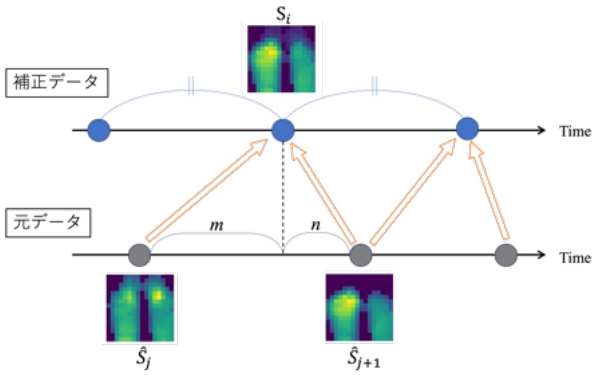

図 5　線形補間による座圧データの補正

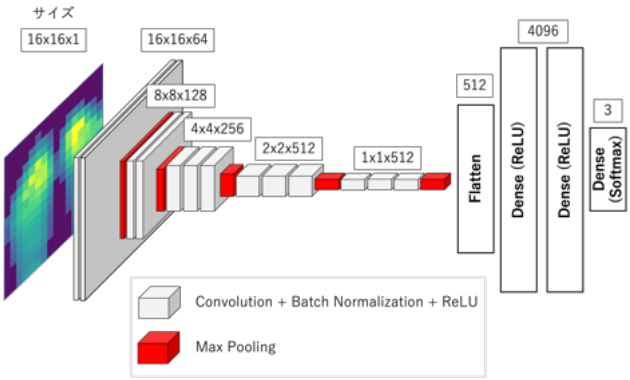

図 6　VGG-16 モデルのネットワーク構造

データ S_i は各画素に対して線形補間を行い，以下の式で求める．

$$S_i = \frac{n\hat{S}_j + m\hat{S}_{j+1}}{m+n} \tag{1}$$

3　提案手法

前章で作成した座圧データベースを用いて教師あり学習により話者推定を行う．ここでは，2 パターンの入力データおよび推定モデルについて検討する．

3.1　話者推定モデル

発話時は姿勢が少し前傾になったり，身振り手振りなどの動作が行われるケースが多い．一方発話していないときは，姿勢の変化が少なく，頷きなどの小さな動作のみが行われるケースが多い．以上より，話者推定には姿勢変化やジェスチャといった動きの情報が重要になると考えられる．そこで，時系列情報の有用性を検証するため，以下の 2 つのモデルについて検討する．

VGG-16 モデル [5]　1 枚の画像を入力とする深層学習モデルであり，一般画像認識タスクにおいては高い性能を示している．VGG-16 モデルのネットワーク構造を図 6 に示す．

3D-CNN モデル [6]　画像列（時系列データ）を入力とする深層学習モデルであり，動画像を用いた行動認識タスク等において高い性能を示している．3D-CNN モデルのネットワーク構造を図 7 に示す．

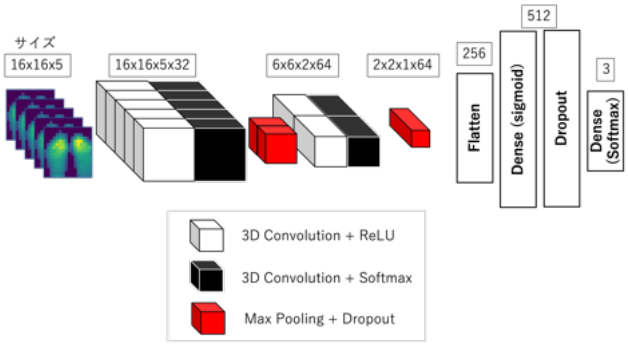

図 7　3D-CNN モデルのネットワーク構造

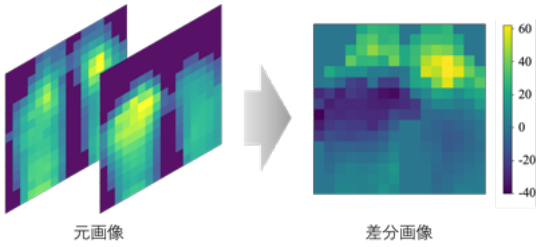

図 8　差分画像の生成

3.2　入力データ

推定モデルだけではなく入力データについても，動きの情報を考慮した場合について比較を行う．ここでは時系列情報を考慮するため，フレーム間の差分画像を推定モデルへの入力データとする．差分画像は現時刻の座圧画像から 1 フレーム前の座圧画像を各画素について減算することで生成する．生成した差分画像の例を図 8 に示す．

ただし，3D-CNN モデルにおいては，各フレームにおいて 1 フレーム前との差分画像を生成し，差分画像列を入力とする．3D-CNN モデルは時系列情報を考慮したネットワーク構造であるが，差分画像列を入力することで，より明示的に動きの情報を活用することが可能であると考える．

また，座圧の大きさは座り方や体重により異なるため，各被験者について 25 分間のタスクごとに各画素値の平均が 0，分散が 1 となるように標準化を行う．

4　評価実験

グループディスカッションにおける座圧データベースを用いて，推定モデルとして VGG-16 モデル，3D-CNN モデルを用いる場合，入力データとして元画像（列），差分画像（列）を用いる場合についてそれぞれ評価実験を行った．

4.1　評価方法

ディスカッションタスクを行ったグループごとに Leave one-group out 交差検証を行う．8 グループのうち，1 グループをテストデータ，他の 1 グループを検証データ，残りの 6 グループを訓練データとして，全てのグループが 1 回ずつテストデータになるように計 8 回検証を行い，8 回分の評価指標の平均により性能評価を行う．評価指標には多クラス分類においてクラスごとに計算した F1 score の平均を取る macro-F1 score を用いる．なお，検証データの macro-F1 score が最も高かった Epoch における学習モデルを用いて，テストデータで評価する．

表 1　評価実験で使用する 3 値分類ラベルと 2 値分類ラベル

3 クラス分類	誰も発話していない 自分が発話している 他人が発話している
2 クラス分類	自分は発話していない 自分が発話している

表 2　各種関数とハイパーパラメータ

Loss function	Categorical cross-entropy
Optimizer	Adam
Batch Size	1024
Epoch	50 or under
Learning Rate	Initial value:10^{-2}, epoch ≥ 40 : 10^{-3}
Sampling	oversampling (SMOTE)

本評価実験では，2.2 章で作成した話者ラベルを用いて，表 1 に示す 2 つの実験を行う．

4.2　実験設定

話者推定モデルの学習に用いた各種関数およびハイパーパラメータを表 2 に示す．本データベースのグループディスカッションは 4 人 1 組で行っているため，「自分が発話」に比べて「他者が発話」「発話なし」のデータが多い傾向にある．学習データの不均衡の問題に対処するため，Synthetic Minority Over-sampling Technique (SMOTE)[7] という手法を用いてオーバーサンプリングを行う．

3D-CNN モデルにおいては，1 秒間の座圧画像列（5 フレーム分）を入力とする．ただし，座圧データのサンプリングレートは 5Hz，話者ラベルは 1 秒ごとに付けられているため，5 フレームのデータを切り出す際は，重複がないように話者ラベルの切れ目で 1 秒分ずつデータを切り出す．つまり，3D-CNN モデルに使用するデータ数は VGG-16 モデルに使用するデータ数の 5 分の 1 になる．

4.3　実験結果

3 クラス分類の結果を表 3 に，2 クラス分類の結果を表 4 に示す．表 3，4 より，3 クラス分類，2 クラス分類どちらにおいても，差分画像列を入力とする 3D-CNN モデルが最も良い結果となった．実験結果をまとめると以下のことが言える．

- VGG-16 モデルより 3D-CNN モデルの方が精度が高い
- 元画像より差分画像を入力する方が精度が高い

以上より，座圧情報を用いて話者推定を行う際は，モデルの構造として時系列情報を考慮するのが有効であることが示せた．さらに差分画像を用いることにより，入力データにおいて明示的に時系列情報を考慮することも有効であることが示せた．

表 3　実験結果 (3 クラス分類)

モデル	入力	F1-score
VGG-16	元画像 (1 枚)	0.306
VGG-16	差分画像 (1 枚)	0.395
3D-CNN	元画像列 (5 枚)	0.452
3D-CNN	差分画像列 (5 枚)	**0.456**

表 4　実験結果 (2 クラス分類)

モデル	入力	F1-score
VGG-16	元画像 (1 枚)	0.486
VGG-16	差分画像 (1 枚)	0.623
3D-CNN	元画像列 (5 枚)	0.717
3D-CNN	差分画像列 (5 枚)	**0.728**

5　おわりに

本研究では座圧情報を用いて深層学習モデルにより話者推定を行う手法を提案した．推定モデル，入力データそれぞれについて時系列情報を考慮する場合について検討し，グループディスカッション時の座圧データベースを用いて評価実験を行った．実験の結果，推定モデル，入力データともに時系列情報を用いた場合，つまり，差分画像列を入力とした 3D-CNN モデルを用いた場合の精度が最も高くなり，macro-F1 score が 3 クラス分類で 0.456，2 クラス分類で 0.728 となった．

現在は 1 秒間のデータを入力としているが，入力するデータの期間を変えた場合について今後検討を行う．また，画像データから推定した姿勢情報（骨格モデル）を用いて話者推定を行った場合との比較も行う予定である．

参考文献

[1] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6301–6305. IEEE, 2019.

[2] 森田陽介, 飯島安恵, 今野将ほか. 時間変化を考慮した機械学習による着座姿勢推定手法. 第 79 回全国大会講演論文集, Vol. 2017, No. 1, pp. 249–250, 2017.

[3] 福井嵐士, 武村紀子, 白井詩沙香, Mehrasa Alizadeh, 長原一. グラフ畳み込みネットワークを用いたグループ学習時の活性度推定. 画像の認識・理解シンポジウム, 2021.

[4] 川喜田二郎. 発想法 – 創造性開発のために. 中央公論社, 1967.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representation (ICLR)*, 2015.

[6] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 35, No. 1, pp. 221–231, 2012.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, Vol. 16, pp. 321–357, 2002.