J-036

対話中の頭部運動機能の解釈の曖昧性を推定する深層回帰モデルの検討

A study on deep regression models for estimating the ambiguity in interpretation of head-movement functions in conversations

武田 一輝 *　　大塚 和弘 †
Kazuki Takeda　　Kazuhiro Otsuka

1 はじめに

対話は，社会集団における意思決定や問題解決の場で重要な役割を担う．そのような対話活動を円滑かつ効率的に支援する情報技術を構築するためには，対話活動それ自体についての深い理解が必要とされる．このような課題に対して，頭部運動や顔表情のような非言語行動が重要な手掛かりとなることから注目を集めている．特に頭部運動は対話中で重要な役割を担うことが知られ，多様な機能を持つ．また，頭部運動は見るものにより意味の解釈が異なるという曖昧性を本質的な性質として持つ．

このような頭部運動の機能を自動的に認識するため，Otsuka らは複数の外部観察者による機能の非排他的な評定を実施し，「頭部運動機能コーパス」を構築している．また，このコーパスに基づいて各機能の有無を二値分類する畳み込みニューラルネット (CNN) を構築しているが，二値分類の正解を複数の観察者の評定の論理和として定めており，解釈の曖昧性を十分に捉えることができない．

そこで本稿では，頭部運動の機能について，その解釈の曖昧性を含めて捉えるため，頭部運動機能の認識問題を分類問題ではなく，新たに回帰問題として定式化する．具体的には，複数の観察者による機能の有無の評定値の平均値として，機能の「強度」を定義し，また，評定値の標準偏差として，機能の「曖昧度」を定義し，推定の対象とする．この機能の強度は，ある頭部運動に対して，何割の観察者が機能の存在を認定したかという割合に対応し，曖昧度は，観察者間の評定のばらつきを意味する．これらを推定することで，頭部運動機能を有り無しの二値として認識するよりも，より仔細に捉えることができ，対話者の状態や対話の質をより深く理解するための手掛かりを獲得できると考えられる．

本稿では，機能の強度及び曖昧度を推定するためのモデルとして μ–σ CNN を提案する．μ–σ CNN は従来の機能認識の CNN[1] を起点として，機能の強度及び曖昧度を推定するように拡張を行ったマルチタスク CNN 回帰モデルである．このモデルは，強度と曖昧度を同時に推定することにより，互いの推定に対して有用な特徴を共有し，双方の推定性能の向上を狙う．

* 横浜国立大学 大学院 理工学府 Graduate School of Engineering Science, Yokohama National University
† 横浜国立大学 大学院 工学研究院 Faculty of Engineering, Yokohama National University

2 関連研究

2.1 頭部運動機能の自動認識

従来，頭部運動の自動認識は排他的な運動パターンを識別する問題として扱われていた．しかし，対話中の頭部運動には多様な機能が存在し，一回の動作が複数の意味を持つ「多重性」や見るものによって解釈が異なる「曖昧性」が本質的な性質とされ，運動パターンと機能は一対一に対応しない．

このような頭部運動の機能を認識するため，Otsuka らは 32 種の頭部運動機能を設定し，3 名の外部観察者の解釈を統合した頭部運動機能コーパスを構築している [1]．このコーパスでは，外部観察者が各々，複数の機能の同時出現を許容した機能の評定を行い，かつ，3 名の観察者の評定ラベルの論理和により機能集合を求めることで，機能の多重性や曖昧性を捉えることができるとした．

また，Otsuka らはこのコーパスに基づき，頭部運動機能の有無について自動認識を行う畳み込みニューラルネット (CNN) を提案している [1]．この CNN は，対話者の頭部姿勢角や発話有無の時系列から各機能の有無の二値分類を行うモデルである．この手法では各機能に対して CNN を構築し，認識を行うことで多重性や曖昧性に対処している．しかし，3 名の外部観察者の評定ラベルの論理和を認識対象としており，機能の表出の強さや確からしさという観点において，解釈の曖昧性を十分に扱うことができない．

そこで本稿では，集約前の観察者の評定に立ち返り，その評定値の平均やばらつき度合いを，機能の強さや曖昧さを表す指標として捉え，それらを回帰問題として推定することで，頭部運動機能に対する理解を深めることを目指す．

2.2 CNN による回帰

人間行動認識の分野において，CNN は歩く，座るなどの運動パターンの識別を行うモデルとして広く利用されている [2][3]．また，このようなパターン識別問題以外にも，CNN は身体や頭部の姿勢などを推定する回帰モデルとしても有効であることが知られている [4][5]．このように，回帰モデルとしての CNN の有効性，及び，先行研究 [1] にて，CNN により頭部運動機能認識のための特徴学習が可能であることが示唆されていることから，本研究では，頭部運動機能の強度及び曖昧度の推定のため CNN モデルを構築することとした．

2.3 マルチタスク学習モデル

マルチタスク学習は，複数のタスクを一つのモデルで同時に遂行できるよう学習を行う手法である [6][7]．タスクごと

に別個のモデルを用いる方法に比べ，マルチタスク学習では，タスク間でモデル構造や特徴表現を共有することで学習効率化や汎化性能向上が期待でき，音響信号処理や画像認識などの様々な分野で有効性が確認されている．人物行動認識においても，顔のランドマーク推定を主たるタスクとして，眼鏡の有無や笑顔の有無の推定を同時に行うことで，推定性能が向上することを確認している [8]．

本稿で提案する μ–σ CNN は，頭部運動機能に関する観察者の評定値の平均（強度）及び標準偏差（曖昧度）を同時に推定するマルチタスク学習を特徴としている．これは上述のマルチタスク学習の利点に着目し，双方のタスクの性能向上を狙いとしている．

3　頭部運動機能の強度・曖昧度

3.1　頭部運動機能コーパスの評定プロセス

頭部運動機能コーパス [1] は女性 4 人を一組とした対面対話，2 グループ各 2 セッションの合意形成対話 4 セッションを対象とし，各対話者のバストショット映像，ピンマイク音声，磁気式センサによる頭部位置・姿勢データを含む．この対話映像に対して，まず外部観察者 1 名により頭部運動区間として相槌，首振り，傾げのいずれかが表出される時間区間がラベル付けされた．その後，3 名の外部観察者により，頭部運動区間内の各フレームに対して，32 種の機能の有無が複数機能の同時出現を許容して，評定された．その後，各フレームについて 3 人の評定値の論理和により最終的な機能集合が決定された．また，外部観察者 1 名により各人のピンマイク音声から発話区間が検出された．

3.2　機能の強度と曖昧度の定義

本研究では，各外部観察者が認定した頭部運動機能の評定値（2 値）について，観察者間の平均値を頭部運動機能の「強度」として定義する．この強度の定義は，何割の観察者が機能の存在を認めたかという割合に相当し，より多くの観察者が機能の存在を認めた場合，頭部運動によりその機能がより強く表出され，対話中にて明確にその機能が発揮されていたという仮定に基づくものである．ここで，時刻 t において，ある観察者 $k \in \{1, \cdots, K\}$ による一つの機能 $f \in \{1, \cdots, F\}$ の評定値を

$$l_{k,t,f} = \begin{cases} 1 & (\text{機能 } f \text{ が認定される場合}) \\ 0 & (\text{機能 } f \text{ が認定されない場合}), \end{cases} \tag{1}$$

のように表す．ただし，K は観察者の人数，F は機能数とする．時刻 t の頭部運動機能 f の強度 $\mu_{t,f}$ は，観察者の評定値の加法平均として，

$$\mu_{t,f} = \frac{1}{K} \sum_{k=1}^{K} l_{k,t,f}, \tag{2}$$

のように定義される．

一方，頭部運動機能の曖昧度は，評定値のばらつき度合いを表し，観測者の評定値の標準偏差として定義される．つまり，全ての観察者の評定値が一致した場合には曖昧度は 0 となり，評定が割れた場合には曖昧度は大きくなる．時刻 t の頭部運動機能 f の曖昧度 $\sigma_{t,f}$ は，

$$\sigma_{t,f} = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(l_{k,t,f} - \mu_{t,f}\right)^2}, \tag{3}$$

と定義される．なお，今回，観測者の評定値は 2 値であるため，強度から曖昧度は一意に決定できるが，μ–σ CNN では，これら 2 変数を独立なものとみなして推定対象とする．

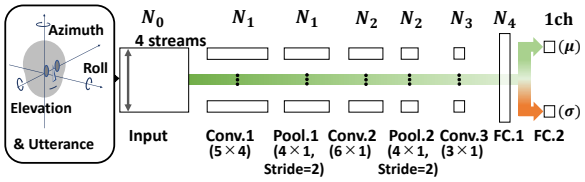

図 1　μ–σ CNN．$N_0 \sim N_4$ は各層のフィルタ数，（μ）は機能の強度を推定する層，（σ）は機能の曖昧度を推定する層をそれぞれ示す

4　μ–σ CNN

頭部運動機能の強度，及び，曖昧度を同時に推定する CNN 回帰モデルとして，μ–σ CNN を提案する．このモデルは，先行研究 [1] で構築された頭部運動機能を検出する CNN を起点として，頭部運動機能の強度及び曖昧度を推定するように拡張したモデルである．本研究では，頭部運動機能の強度を主たる推定対象とするが，μ–σ CNN では，強度と曖昧度のマルチタスク学習により両方の推定タスクへの相乗効果を狙う．

図 1 に μ–σ CNN の構造を示す．μ–σ CNN は，対話中の一時刻における特定機能の強度，及び，曖昧度を出力する回帰モデルであり，推定対象の時間フレームを中心とした窓幅に含まれる部分時系列データを入力とする．その入力は，図 1 左に示すような頭部姿勢角の 3 成分，方位角（Azimuth），仰角（Elevation），ロール角（Roll）各々の角速度，及び，発話の有無（Utterance）である．入力後，初段の畳み込み層にて，頭部姿勢角速度の 3 成分と Utterance のデータを統合し，その後 2 層の畳み込み層を通して，入力したデータから推定に必要な特徴量を抽出する．畳み込み層を通し抽出した特徴量は，全結合層を通して集約される．最終段 FC.2 の 2 つの全結合層の出力に対して，活性化関数として Sigmoid 関数を適用し，頭部運動機能の強度及び曖昧度の実数値を出力する．

μ–σ CNN では，頭部運動機能の強度と曖昧度という 2 つの回帰対象を持つため，どちらに比重をおいてモデルの最適化を行うかによって，最終的なモデルの性能が変化すると考えられる．そこで強度と曖昧度の推定の比重を調整可能とするため，μ–σ CNN の学習時の損失関数を

$$Loss = \beta \left(\alpha(\mu - \hat{\mu})^2 + (1 - \alpha)\left(\sigma - \hat{\sigma}\right)^2 \right), \tag{4}$$

のように定義する．ただし，$\mu, \hat{\mu}$ は，それぞれ強度の正解値，及び，推定値であり，$\sigma, \hat{\sigma}$ は曖昧度の正解値，及び，推定値を表す．$\alpha (0 < \alpha < 1)$ は，強度と曖昧度の損失の比重を調整する「損失重み係数」であり，$\alpha < 0.5$ の場合，強度よりも曖

表 1　機能の強度に関する推定誤差 (wMAE)．太字は各機能で最小の誤差を示す．低減率は CNN−I の推定誤差に対して，μ–σ CNN の推定誤差が減少した割合 [%] を示す．

対象機能	Baseline	CNN−I	μ–σ CNN	低減率
s1 リズム取り	0.438	0.225	**0.222**	1.46
s2 強調	0.470	0.254	**0.221**	12.9
s5 反応確認	0.462	0.257	**0.223**	13.3
s8 思考 [発話時]	0.468	0.321	**0.300**	6.79
r1 相槌	0.333	0.216	**0.206**	4.68
r2 応答	0.470	0.370	**0.316**	14.6
r5 思考 [受け手]	0.464	0.354	**0.328**	7.23
r6 理解	0.483	0.351	**0.325**	7.18
r11 肯定	0.480	0.386	**0.324**	16.2
c1 正の感情表出	0.457	0.346	**0.309**	10.6
Average	0.453	0.308	**0.277**	9.48

昧度に重きをおくことを意味する．また，$\beta(>0)$ は損失関数全体の大きさを調整する「損失全体係数」である．

5　実験

本稿では，頭部運動機能コーパス [1] において対話中で頻出する上位 10 種の機能（リズム取り (s1)，強調 (s2)，反応確認 (s5)，思考 [発話時](s8)，相槌 (r1)，応答 (r2)，思考 [受け手](r5)，理解 (r6)，肯定 (r11)，正の感情表出 (c1)）を対象として頭部運動機能の強度及び曖昧度の推定を行った．コーパスに含まれる，集約前の 3 人の外部観察者の評定値を用いて，各機能の強度及び曖昧度の正解値を求めた．今回，3 人の二値ラベルを対象とするため，正解値の強度 μ の水準は 4 段階 $\{0, 1/3, 2/3, 1\}$ であり，また，曖昧度 σ の水準は 2 段階 $\{0, \sqrt{2}/3\}$ である．

入力データは頭部運動機能コーパスに含まれる 3 自由度の頭部姿勢の角速度と発話の有無の部分時系列（推定対象フレームを中心とした 32 フレーム分）を用いた．頭部姿勢角速度は，頭部姿勢角の時系列からフレーム間差分により求めた．発話の有無の時系列データは 2 値の発話の有無のデータに対して，移動平均フィルタを用いて平滑化を行ったデータを入力した．

μ–σ CNN の有効性を検証するため，ベースラインとして，機能ごと強度，及び，曖昧度の平均値を常に出力するモデルを用いた (以後，Baseline と呼ぶ)．また，μ–σ CNN によるマルチタスク学習の有効性を検証するため，強度あるいは曖昧度のみを推定対象とするシングルタスク CNN 回帰モデルを比較対象として用いた．以後，強度を対象とするモデルを CNN−I，曖昧度を対象とするモデルを CNN−A と記す．この CNN の構造は，図 1 に示す μ–σ CNN の FC.2 を強度または曖昧度のみを推定対象とするよう改変したものである．

正解値のサンプル数の偏りを考慮し，推定性能の評価尺度として，重み付き平均絶対誤差（weighted Mean Absolute Error, 以後 wMAE と略す）を用いた．wMAE は，推定値 $\hat{y}_m$，正解値 $y_m \in \{0, 1/3, 2/3, 1\}$ に対して，

$$\mathrm{wMAE} = \frac{\sum_{m=1}^{M} w_m |\hat{y}_m - y_m|}{\sum_{m=1}^{M} w_m}, \qquad w_m = \frac{M}{M^{(y)}}$$

と定義される（$m = 1, \ldots, M$）．ただし，M はサンプル数，$M^{(y)}$ は正解値が y であるサンプル数をそれぞれ示す．wMAE は，絶対誤差に対して各水準のサンプル数の逆比を掛けて平均をとることにより，サンプルの偏りを是正して，サンプル数の少ない強度水準に対する推定性能を考慮した評価を可能とする．

また学習・評価には交差検証法を用いた．各対話セッションにおける参加者 1 人のデータを 1 データとしたとき，残りの 15 データ（4 人 × 4 対話 −1）を用いて訓練を行い，1 データについてテストを行うというプロセスを全 16 データについて繰り返した．各データの頭部運動区間上の全フレームについて，μ–σ CNN により機能の強度と曖昧度を算出し，最後に 16 データ分まとめて wMAE を計算した．

6　結果と考察

表 1 に強度及び曖昧度に関する推定誤差を示す．表 1 より，CNN−I および μ–σ CNN は，全ての機能にて Baseline よりも誤差が小さかった．また，CNN−I と比較して，μ–σ CNN は，全機能にて誤差の低減が確認でき，r11（肯定）において最大の低減率 16.2% を達成した．

次に損失関数の係数の設定が推定誤差に及ぼす影響を調査するため，図 2 に損失重み係数 α を変化させた場合の wMAE を示す．10 機能中 9 機能において，$\alpha < 0.5$ の範囲で wMAE が最小となることが確認でき，強度よりも曖昧度に重きをおくことで，強度の推定誤差を低減することができることが示唆された．一方，α を 0 に近づけた場合，wMAE が増加する傾向も見られるため，α には適切な範囲があると考えられる．

次に表 2 に曖昧度の推定誤差を示す．この結果より，CNN-A と μ–σ CNN は，ともに Baseline よりも推定誤差が小さくなった．10 機能中 6 機能にて μ–σ CNN の性能が CNN-A を上回り，平均でみると両者の誤差はほぼ拮抗することがわかった．μ–σ CNN の損失関数の係数 α, β は，主たるタスクである強度推定に対して最適化されたものであり，その点を鑑みると，曖昧度の推定にて CNN-A と同程度の推定性能が得られたことは，μ–σ CNN の有効性を裏付ける結果として解釈できる．

7　議論

μ–σ CNN のさらなる性能向上のためには，μ–σ CNN の動作機序について理解することが課題としてあげられる．前節 6 では実験的に μ–σ CNN による強度と曖昧度の同時推定の性能を検証したが，その動作機序と誤差低減の関係性については，まだ検証の余地がある．そこで，動作機序の分析を進め，さらなる性能向上のための手掛かりを探る予定である．

また，頭部運動機能の強度の分解能を向上させることも，頭部運動機能の理解を深めるために必要である．本稿では，3 人の観察者による評定値を用いていたため，強度の正解値は 4 段階に留まっていた．今後，強度の分解能を向上させるため，観察者の人数を増やして機能の評定を行うことが課題である．また，そのような高分解能のデータに対しても，μ–σ

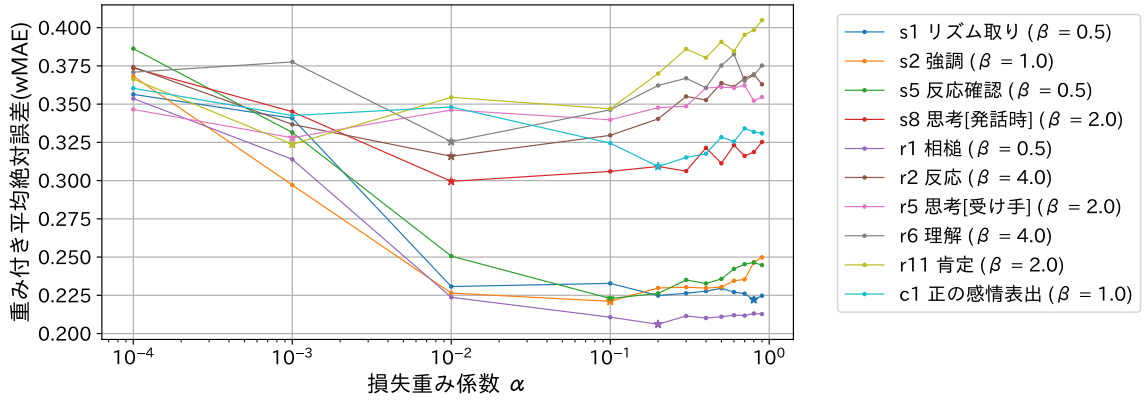

図 2　強度の推定誤差 (wMAE) に対する損失重み係数 α の影響．凡例内に損失全体係数 β を示す．$\star$ は，機能ごとで最小の wMAE を示す．

表 2　曖昧度の推定誤差 (wMAE)．太字は各機能で最小の誤差を示す．

対象機能	Baseline	CNN-A	μ–σ CNN
s1 リズム取り	0.236	**0.132**	0.142
s2 強調	0.236	0.123	**0.120**
s5 反応確認	0.236	**0.147**	0.148
s8 思考 [発話時]	0.236	0.201	**0.197**
r1 相槌	0.236	**0.188**	0.189
r2 応答	0.236	0.203	**0.201**
r5 思考 [受け手]	0.236	0.179	**0.173**
r6 理解	0.236	0.185	**0.181**
r11 肯定	0.236	**0.193**	0.196
c1 正の感情表出	0.236	0.198	**0.195**
Average	0.236	0.175	**0.174**

CNN が有効であるか検証を行うことも必要である．

μ–σ CNN の適用範囲を拡大することも今後の課題である．今回は対話中の頭部運動に焦点を当てたが，他の非言語行動にも同様に機能の解釈には曖昧性が伴う．例えば，顔の表情の場合，同じ笑顔であっても，常に幸福を示すとは限らず，喜び，羞恥，軽蔑など多様な意味を持つ可能性があり，見るものにより多様な解釈が可能である．そのため，今後は顔やその他モダリティの行動についても，複数の観察者による機能の評定を行い，機能の強度や曖昧度の推定を進めたい．

8　結び

対話中の頭部運動が持つ機能の強度，及び，曖昧性を推定する μ–σ CNN を提案した．対話者の頭部運動は様々な機能を担うが，その解釈には曖昧性が伴う．本稿では，複数の観察者による評定の平均値を機能の強度とし，また，評定値のばらつきを曖昧度と定義し，これらを推定する回帰問題を新たに定式化した．その解法として，機能の強度及び曖昧度を同時に推定することで，双方の推定に益することを狙いとしたマルチタスク CNN 回帰モデルを構築した．実験の結果，μ–σ CNN は，強度推定のみを行う CNN 回帰モデルと比較して，全機能で誤差の低減を確認し，提案法の有効性が示唆された．

謝辞

本研究をご支援いただいた日本電信電話株式会社コミュニケーション科学基礎研究所に感謝いたします．また本研究は栢森情報科学振興財団，立石科学技術振興財団，JSPS 科研費 JP21K12011 の助成を受けたものです．

参考文献

[1] K. Otsuka and M. Tsumori. Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access*, Vol. 8, pp. 217169–217195, 2020.

[2] Salwa O. Slim, Ayman Atia, Marwa M.A. Elfattah, and Mostafa-Sami M.Mostafa. Survey on human activity recognition based on acceleration data. *Int. J. Advanced Computer Science and Applications*, Vol. 10, No. 3, pp. 84–98, 2019.

[3] Carlos Avilés-Cruz, Andrés Ferreyra-Ramírez, Arturo Zúñiga-López, and Juan Villegas-Cortéz. Coarse-fine convolutional deep-learning strategy for human activity recognition. *Sensors*, Vol. 19, No. 7, p. 1556, 2019.

[4] Tomas Pfister, Karen Simonyan, James Charles, and Andrew Zisserman. Deep convolutional neural networks for efficient pose estimation in gesture videos. In *Proc. Asian Conference on Computer Vision*, pp. 538–552, 2014.

[5] Massimiliano Patacchiola and Angelo Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, Vol. 71, pp. 132–143, 2017.

[6] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. arXiv:2009.09796, 2020.

[7] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, Vol. 5, No. 1, pp. 30–43, 09 2017.

[8] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Proc. European Conf. Computer Vision (ECCV)*, pp. 94–108, 2014.