面接試験の観点別自動評価のためのマルチモーダル深層学習自動採点モデルの開発

Development of a Multimodal Deep Neural Automated Scoring Model for Interview Trait Scoring

北嶋太一 *1，宇都雅輝 *1
Taichi Kitajima*1, Masaki Uto*1
*1 電気通信大学
*1The University of Electro-Communications
Email: {kitajima, uto}@ai.lab.uec.ac.jp

あらまし： 近年，人工知能技術を活用した面接試験の自動採点が注目されている．本研究では，面接の記録データから，複数の評価観点で得点予測を行う新たな面接自動採点技術を提案する．具体的には，面接の記録データを構成する映像，音声，テキストのそれぞれのデータから，人手で設計した特徴量と深層学習ベースの特徴量を同時に抽出して処理するとともに，Attention 機構を用いて評価観点間の相関を活用して得点予測を行うことができるマルチモーダルかつマルチタスク型の面接自動採点モデルを提案する．

キーワード： 面接試験，自動採点，深層学習，マルチモーダル，Attention 機構

1 はじめに

面接試験は，コミュニケーション能力や表現力を評価する手段の一つとして様々な評価場面で広く活用されている．しかし，面接試験では，評価結果が評価者の甘さ/辛さなどに依存する可能性があることに加えて，評価に多くの時間的，金銭的なコストを要するという問題を有している．これらの問題を解決する方法の一つとして，人工知能技術を用いた面接試験の自動採点が期待されている．面接試験の自動採点では，面接時の映像や音声，発話内容のテキスト情報などを記録し，それをマルチモーダルな機械学習技術を用いて解析することで，採点を自動化することを目指す．

面接自動採点を実現する従来手法では人手で設計した特徴量を利用する方法が広く採用されてきた．例えば Naim ら [1] は，音声，語彙，顔に関連する特徴量を入力として，SVR (Support Vector Regression) やラッソ回帰モデルを学習し，評価観点別の得点を予測する手法を提案している．しかし，このような従来の特徴量ベースのモデルはその性能が人手で設計した特徴量に依存してしまい，発話内容や音声，映像に内在する複雑な特徴情報を必ずしも十分に考慮できないという問題点がある．

これに対し近年では，深層学習を活用して，自動採点に有効な特徴量を自動的に獲得するアプローチが提案されている [2]．しかし，従来の深層学習ベースの手法は，データのモダリティごとに人手で設計した特徴量か深層学習ベースの特徴量のいずれか一方のみを使用していた．他方で，本来は，いずれのモダリティにおいても，事前に人手設計可能な特徴量と人手では設計が困難な複雑な特徴情報は共存すると考えられる．加えて，既存研究の多くは，表情や抑揚，論理性といった複数の評価観点で得点付けを行うことを目的としているが，その際に，各観点の得点を独立に予測するように設計されている．他方で，小論文自動採点の分野では複数の観点の得点を独立に予測するのではなく，評価観点間の相関関係も考慮して全観点の得点を同時に予測するマルチタスク型の手法が提案され，高精度を達成している [3]．

以上の背景を踏まえ，本研究では，面接の記録データから得られる全てのモダリティデータに対し，深層学習で抽出した特徴量と人手で設計した特徴量を使用し，かつ，Attention 機構を用いて評価観点間の相関を活用できるマルチモーダルかつマルチタスク型の面接自動採点モデルを提案する．

2 提案手法

本研究では採点済み面接データを $\{(\mathcal{A}_i, \mathcal{V}_i, \mathcal{T}_i, \mathcal{S}_i)\}_{i=1}^{I}$ と仮定する．ここで I は面接データの総数，$\mathcal{A}_i$ は i 番目の面接の音声データ，$\mathcal{V}_i$ は映像データ，$\mathcal{T}_i$ は人手で書き起こした対話データを表す．また $\mathcal{S}_i = \{s_{ik}\}_{k=1}^{K}$ は人手でつけられた評価観点別の得点であり，K は評価観点数，s_{ik} は k 番目の評価観点の得点を表す．本研究のタスクは，$\mathcal{A}_i$，$\mathcal{V}_i$，$\mathcal{T}_i$ から，$\hat{\mathcal{S}}_i$ を予測することである．

なお，面接には面接官の質問と受検者の回答が 1 組以上含まれる．$\mathcal{A}_i$，$\mathcal{V}_i$，$\mathcal{T}_i$ から $l \in \{1, 2, \cdots, L\}$ 番目の質問応答に対応する部分を $\mathcal{A}_{il}$，$\mathcal{V}_{il}$，$\mathcal{T}_{il}$ として抽出し，モデルに入力する．ここで，L は質疑応答数を表す．

提案モデルの概念図を図 1 に示す．提案モデルへの入力は，面接データから事前学習済み深層学習モデルを用いて取得した埋め込みベクトルと人手で設計した特徴量ベクトルである．ここで $\mathcal{T}_{il}$ から Sentence-BERT を用いて抽出した対話テキストの埋め込み表現を E_{il}^{T}，$\mathcal{A}_{il}$ と $\mathcal{V}_{il}$ から VideoLLaMA 2 を用いて抽出した音声・映像の埋め込み表現を E_{il}^{V}，$\mathcal{V}_{il}$ から GhostFaceNets を用いて抽出した顔の埋め込み表現を E_{il}^{F}，$\mathcal{A}_{il}$ から ECAPA-TDNN を用いて抽出した声の埋め込み表現を E_{il}^{S} とする．人手で設計した特徴量 $\mathcal{H}_i$ は，中間層に入力される．

各モダリティの埋め込み表現 $\boldsymbol{E}_{il}^{T}$，$\boldsymbol{E}_{il}^{V}$，$\boldsymbol{E}_{il}^{F}$，$\boldsymbol{E}_{il}^{S}$ はまず全観点に共通する共通層に入力される．ここでは，全結合層を通して出力ベクトルを結合し，2 層の全結合層を通して，質問ごとの埋め込み表現を作成する．

表 1　観点別の相関係数

モデル	観点番号																平均
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
従来: SVR	.650	.652	.669	.790	**.409**	.656	.731	.473	.595	.501	.477	.458	.587	.588	.399	.618	.578
従来: ラッソ回帰	.647	.675	.638	.787	.311	**.710**	**.755**	.373	**.739**	.494	.488	.439	.522	.462	.369	.564	.561
提案	**.702**	**.710**	**.739**	**.792**	.399	.658	.705	**.528**	.636	.520	**.513**	.512	**.631**	**.594**	.480	**.636**	**.610**
提案 w/o T	.671	.643	.647	.776	.406	.612	.661	.491	.627	.537	.469	**.575**	.567	.549	.480	.624	.583
提案 w/o V	.680	.653	.636	.780	.324	.658	.681	.409	.612	**.541**	.474	.529	.606	.549	.450	.596	.574
提案 w/o F	.676	.671	.694	.789	.323	.673	.726	.502	.589	.484	.487	.436	.595	.543	.352	.591	.571
提案 w/o S	.665	.674	.655	.788	.342	.659	.693	.489	.597	.527	.488	.555	.573	.536	**.494**	.600	.583
提案 w/o H	.562	.540	.465	.556	.352	.540	.551	.404	.447	.199	.332	.229	.469	.439	.300	.551	.433
提案 w/o TrAtt	.690	.673	.688	.778	.360	.635	.688	.490	.589	.526	.484	.511	**.631**	.576	.381	.618	.582

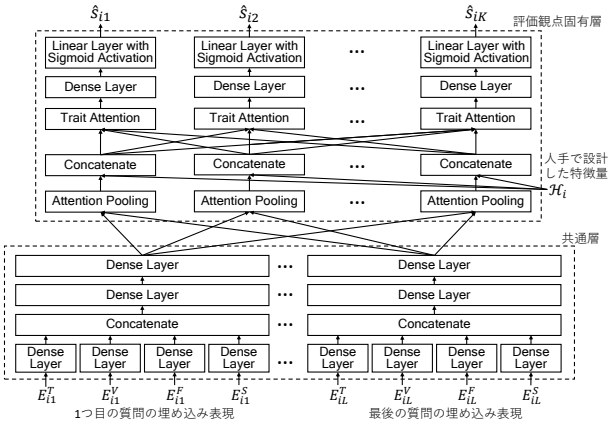

図 1　提案モデルの概念図

続いて，評価観点に固有の層において，l 番目の質問の埋め込み表現 $\boldsymbol{z}_l$ に，次式の Attention Pooling を適用することで，k 番目の評価観点に対するベクトルを得る．

$$\boldsymbol{p}_k = \sum_l \boldsymbol{z}_{kl} \,\mathrm{softmax}(\boldsymbol{v}_k \cdot \tanh\left(\boldsymbol{W}_k \cdot \boldsymbol{z}_l + \boldsymbol{b}_k\right)) \quad (1)$$

ここで $\boldsymbol{W}_k$，$\boldsymbol{v}_k$ は重み，$\boldsymbol{b}_k$ はバイアスベクトルを表す．

次に，$\boldsymbol{p}_k$ に人手で設計された特徴量を結合してベクトル $\boldsymbol{r}_k$ を得る．さらに，観点間の相関を考慮するために，$\boldsymbol{r}_k$ に対して，次式に示す Ridley ら [3] が提案した Trait Attention を適用する．

$$\boldsymbol{x}_k = \left[\sum_{k' \neq k} \frac{\exp\left(\boldsymbol{r}_k \cdot \boldsymbol{r}_{k'}\right)}{\sum_{k'' \neq k} \exp\left(\boldsymbol{r}_k \cdot \boldsymbol{r}_{k''}\right)} \boldsymbol{r}_{k'},\ \boldsymbol{r}_k \right]. \quad (2)$$

最後に，$\boldsymbol{x}_k$ に評価観点ごとに 2 層の全結合層を適用し，シグモイド関数で正規化された得点を予測する．

提案モデルの出力は，K 個の評価観点別の得点であり，最小二乗誤差を損失関数として学習される．

3　評価実験

実験には Naim ら [1] の研究で収集された MIT Interview Dataset を用いる．このデータセットには，138 本の模擬面接動画が含まれており，各面接に対して，人手による観点別の評価得点が付与されている．評価観点は 16 項目あり，1-7 点のリッカート尺度で評価されている．

本実験では，10 分割交差検証法を用いて性能評価を行った．各分割ではデータを訓練・検証・テストデータを 8:1:1 に分割し，訓練データでモデルを訓練した．検証データによる精度評価を通してハイパーパラメータ最適化を行った．性能評価の指標には相関係数を用いた．

また，ベースライン手法として，人手で設計された特徴量を入力として SVR やラッソ回帰を用いて各評価観点を予測する Naim ら [1] のモデルを作成した．また，提案モデルで導入した各特徴量の影響を評価するために，入力から $\boldsymbol{E}_{il}^T$，$\boldsymbol{E}_{il}^V$，$\boldsymbol{E}_{il}^F$，$\boldsymbol{E}_{il}^S$，$\mathcal{H}_i$ の 1 つを除いたモデルをそれぞれ作成した．さらに，提案モデルにおいて Trait Attention を用いないモデルも作成した．

実験結果を表 1 に示す．表の提案 w/o T，V，F，S，H，TrAtt は，提案手法から，T（テキスト），V（音声 + 映像），F（顔），S（声）の深層学習ベースの特徴量，および，H（人手設計の特徴量）と TrAtt（Trait Attention）をそれぞれ除いたモデルを意味する．

表 1 から，提案手法はベースライン手法と比べて優れた平均性能を達成したことがわかる．また，提案手法からいずれかの特徴量や Trait Attention 層を除くと性能が低下することもわかり，全ての提案機構の有効性が確認できた．

4　まとめ

本研究では，面接動画の各モダリティから深層学習で抽出した特徴量と，人手で設計した特徴量を組み合わせ，評価観点別の得点の相関を活用するモデルを提案し，実験によりその有効性を示した．今後は，特徴量やモデルの改良を行うことで性能向上を目指す．

参考文献

(1) Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and et al. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *Proceedings of 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–6, 2015.

(2) Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and et al. Multimodal hierarchical attention neural network: Looking for candidates behaviour which impact recruiter's decision. *IEEE Transactions on Affective Computing*, Vol. 14, No. 2, pp. 969–985, 2023.

(3) Robert Ridley, Liang He, Xin-yu Dai, and et al. Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13745–13753, 2021.