一般発表 | Track 4: メディア処理・HCI・人間中心情報マネジメント

2026年2月28日(土) 15:30 ~ 17:40 | H会場

[3H] 学習支援

座長:幸島 匡宏(NTT株式会社)　コメンテータ:白石 優旗(筑波技術大学)　ジュニアコメンテータ:井上 沙紀(関西学院大学)

15:30 ~ 15:55

[3H-01] アウトプット型学習のための主張抽出とRAGに基づく訂正・補足情報の生成

*米村 琉衣[1]、中井 香那子[1]、山本 岳洋[1] (1. 兵庫県立大学)

15:55 ~ 16:20

[3H-02] 基本感情の複合分析に基づく表情画像からの困惑深度推定：Transformerを用いた特徴抽出と評価

*石川 昂樹[1]、一色 夢香[1]、寺田 憲司[1]、遠藤 雅樹[1]、大野 成義[1] (1. 職業能力開発総合大学校)

16:20 ~ 16:45

[3H-03] スマートリングによる指先生体情報を用いたプログラミング学習者の困惑状態検知

*一色 夢香[1]、石川 昂樹[1]、寺田 憲司[1]、遠藤 雅樹[1]、大野 成義[1] (1. 職業能力開発総合大学校)

16:45 ~ 17:10

[3H-04] 動画に基づく教本参照型コーチングエージェントの構築

*川田 拓朗[1]、藤若 雅也[2]、JI XIAOTONG[2]、劉 健全[2] (1. 法政大学、2. 日本電気株式会社)

アウトプット型学習のための主張抽出とRAGに基づく訂正・補足情報の生成

米村　琉衣† 　中井香那子†† 　山本　岳洋††

† 兵庫県立大学 社会情報科学部　〒651–2197 兵庫県神戸市西区学園西町 8–2–1
†† 兵庫県立大学 大学院情報科学研究科　〒651–2197 兵庫県神戸市西区学園西町 8–2–1
E-mail: †{fa22p101,ad24c041}@guh.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本研究では，議論の発話内容から話者の主張の抽出と，その主張の中に誤った情報が含まれている場合は訂正文を，不足情報がある場合は補足文の生成に取り組む．学生が行うアウトプット型学習を目的としたプレゼンテーションにおける発話には，誤った理解や不足した背景情報を含む主張が見られることがあり，これらは学生の学習の妨げの要因となる．提案システムでは，発表音声から大規模言語モデル（LLM）を用いて発表者および質問者の主張を抽出する．また，その後発表の根拠となる文書を外部データとした検索拡張生成（RAG）を用いて訂正文と補足文を生成する．実験のため，大学生または大学院生 11 名を対象に実施した論文紹介のプレゼンテーションを用いてデータを収集した．収集したデータに対して音声認識モデルの違いによる主張抽出精度の差と訂正文および補足文の生成精度を，適合率，再現率，F_1 値を用いて評価した．その結果，主張抽出の精度は音声認識の精度に影響されることを確認した．また，提案システムは主張抽出を行わない場合と比較して，訂正文および補足文の生成において，適合率，再現率，F_1 値のいずれも向上することを示した．

キーワード　アウトプット型学習，RAG，主張抽出

1　はじめに

近年の中等教育および高等教育においては，学習者が主体的に学習活動に参加することで，知識の理解を深め，思考力や表現力を育成することを目的とした学習方法であるアクティブ・ラーニングが盛んに行われている [1] [2]．具体的には，授業で習った内容を生徒が各自でまとめて他の生徒に発表するプレゼンテーションなどのアウトプット型学習がある．このような学習方法は授業を聞き知識を身に着けるインプット型学習よりも学習者の知識の獲得に役立つと考えられる．

しかし，アウトプット型学習には課題も存在する．特に，指導教員などの学習を支援する立場にある者が存在しない場合，あるいは指導教員が生徒の発表分野に関する知識を十分に有していない場合，学習者が発表した知識が間違っていたり，議論の最中に突発的に発言をしてしまうことがあると考えられる．このような学習者が突発的に発言する場面では，学習者が自律的に調べた情報の正確性を即座に判断することは困難であり，誤った情報や十分ではない情報を共有してしまうリスクがある．さらに，このような誤情報は発表を聞いている他の学習者にも共有されてしまう可能性があり，学習内容全体の理解に悪影響を及ぼす恐れがある．

そこで本研究では，アウトプット型学習を支援することを目的として，発表および質疑応答の音声から外部文書を用いた検索拡張生成（RAG）に基づいて訂正文および補足文を生成するシステムを提案する．

本システムの実現にあたり，実際の発話には曖昧な表現や言い直しが多く含まれるため，発話全体を対象として正確な訂正または補足を生成することは困難であると考えられる．そこで発話の中から話者が伝えようとしている主張を抽出することにより，訂正文および補足文生成の精度向上を目指す．本システムにより，学習者が共有した情報の誤りの訂正や不足情報を補足することで，より正確で理解の深い学習環境の実現を目指す．

例えば，大学の研究室にて情報検索における nDCG という評価指標についてプレゼンテーション形式で発表を行い，学生のみで学習した情報を共有したとする．その際，図 1 のように発表者が誤った情報を発信した際に，「nDCG の算出式における対数の底を変更するとスコアが変化する.」という主張を抽出した後，誤りの認識と訂正を行い，「発表で用いている nDCG の算出式の対数の底を変化させると DCG のスコアは変わりますが nDCG のスコアは変化しません.」という訂正文や，「nDCG の定義は 2 種類ある」という補足文を画面に返すシステムを実現する．

提案システムでは，まず発表および質疑応答の音声に対して音声認識を行い，得られた文字起こしテキストから大規模言語モデル（LLM）を用いて話者の主張を抽出する．次に，抽出された主張を入力として，発表の根拠となる文書を外部データとした RAG を用い，主張に誤りが含まれる場合には訂正文を，不足情報がある場合には補足文を生成する．

本研究では，研究室内で実施された論文紹介プレゼンテーションを対象として実験を行い，発話内容からの主張抽出精度と主張抽出の有無が訂正文および補足文生成の精度に与える影響を評価した．実験の結果，主張抽出の精度は音声認識の精度

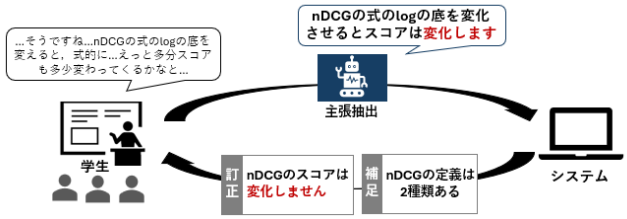


図 1　提案システムの概要.

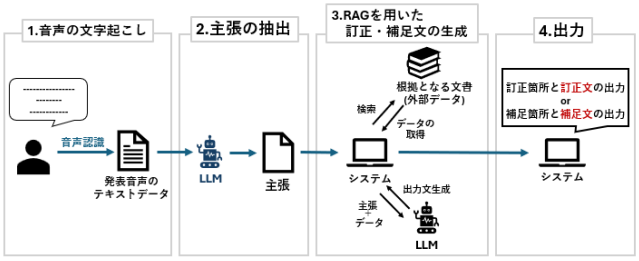


図 2　提案システムの流れ.

に影響されることを確認した．また，主張抽出を導入することで，訂正文および補足文の生成精度が向上することを確認した．

2　関 連 研 究

2.1　LLM を用いた議論支援

LLM を用いた議論支援システムは，多く提案されている．Imamura らは，ブレインストーミング中の議論の書き起こしから埋め込みベクトルを生成し，関連論文を提示する Serendipity Wall を提案している [3]．木下らは，市民参加型 Web 議論を対象に，情報提示の必要性を判定したうえで，LLM による関連情報の検索・要約とファクトチェックを行う情報推薦システムを提案している [4]．

教育分野では，Steinert らが，教育学的理論に基づくプロンプトを用いて学習者の回答に形成的フィードバックを生成する学習支援システム LEAP を提案している [5]．また，Kirstein らは，LLM と RAG を組み合わせ，議事録と補足資料を統合する要約パイプラインを提案し，要約の質向上を示している [6]．加えて，Yun らは，教育理論に基づく専門家ルールベースと LLM を統合したハイブリッド AI エージェントを開発し，教室内の複雑な対話シーケンスを自動分析する手法を提案している [7]．

さらに，創造的協調作業を支援する研究として，Shi らの IdeaWall [8]，Andolina らの InspirationWall [9]，中らの AIR-VAS [10]，Fede らの The Idea Machine [11]，および IdeaExpander [12] が提案されており，議論内容に関連する情報や刺激を提示することで参加者を支援している．

しかしながら，教育分野において，議論の発話内容に基づいて訂正情報や補足情報を提示することを目的とした議論支援システムの提案は少ない．

2.2　LLM を用いた主張抽出やファクトチェック

発話や文章から主張を抽出する研究や，ファクトチェックに関する研究が行われている．

Panchendrarajan らは，自動ファクトチェックにおける主張検出を対象に，単言語・多言語・クロスリンガル設定における既存手法を，検証可能性や優先度などの観点から体系的に整理している [13]．Ullrich らは，文脈付きテキストから事実確認対象となる主張を生成的に抽出する手法を整理・比較し，主張抽出の品質を評価する自動評価指標 Ffact を提案している [14]．また，Tran らは教室内ディスカッションを対象に，LLM を用いて主張構造のエンドツーエンド抽出を行い，少量のアノテーションでも高い性能を示している [15]．

ファクトチェック支援に関して，Gupta らは日常会話中の誤情報をリアルタイムに検知し，ウェアラブル端末による非言語的フィードバックを行うシステム Factually を提案している [16]．さらに，Venktes らはライブ音声ストリームを対象に，発言の書き起こしと話者同定を行い，主張の真偽を数秒以内に検証するシステム LIVEFC を提案している [17]．

一方で，Rashkin らが示すように，既存研究の多くは書き言葉を中心に，真偽判断に有効な言語的特徴の分析に焦点を当てている [18]．そのため，人間の発言，特に発表や議論といった口語的な発話から主張を抽出し，その内容に対して体系的にファクトチェックを行う研究は著者らの知る限り少ない．また，発話内容は冗長表現や言い直しを含むことが多く，発話全体から正確な訂正情報や補足情報を生成することは困難であるため，正確な訂正情報や補足情報を提示するには，発話から話者の主張を正しく抽出する必要がある．本研究は，この点に着目し，議論に対する訂正および補足を行う新たな支援手法を提案するものである．

3　主張抽出に基づく訂正文および補足文の生成

3.1　システムの流れ

本研究ではアウトプット型学習を支援するために，発言に対する訂正と補足を行うシステムを提案する．提案システムの流れを図 2 に示す．まず，発表や質疑の音声の文字起こしを行い，文字起こしテキストから LLM を用いて主張の抽出を行う．その後，得られた主張を入力として，RAG より発表内容における補足が必要な箇所や訂正が必要な箇所を自動的に検出し，対応する訂正文および補足文を生成して画面上に出力する．

3.2　音声認識による発話内容のテキスト化

提案システムにおける主張抽出と訂正文および補足文の生成の入力として用いるため，発話音声に対して音声認識を行い，発話内容をテキスト化する．音声は，話者が一つの発話を話し終えたタイミングで音声ファイルを区切るために，無音区間に基づいて分割した．具体的には，音量が 45dB 以下の状態が 2 秒以上継続した区間を無音区間と定義し，音声ファイルを無音区間ごとに wav ファイルとして分割した．その後，分割された各音声ファイルに対して音声認識モデルを適用し文字起こしを行う．

3.3 発話内容からの主張抽出

発表内容に含まれる誤りや不足情報に対して適切な訂正文および補足文を生成するためには，発話全体から検証対象となる情報を特定する必要があると考えられる．そのため，発話内容から話者が提示している重要な説明や断定的な内容を抽出し，訂正文および補足文生成の対象となる単位を明確にする．

無音区間ごとに分割された発表音声から得た文字起こしデータに含まれる話者の主張を自動で取り出すために，LLM を使用する．本研究における主張とは話者が発表の中で「伝えたい中心的な内容」や「説明したいポイント」をまとめた文のことであり，訂正文および補足文生成における基本的な単位となる．

例えば，nDCG には複数の定義が存在するが，そのうち一般的に用いられている Burges らによる nDCG の定義 [19] について発表が行われたとする．その際，「nDCG の算出式における対数の底を変更するとスコアが変化する」といった性質に関する主張や，「nDCG では対数の底は定義によると後ほどキャンセルされるのでなんでもよい」といった主張などを，発話内容から取り出すことを想定している．

以下は, 実際に LLM に与えたプロンプトであり，以下のプロンプトにおいて文字起こしデータは変数 transcription に格納されている．

主張抽出のプロンプト

```
あなたは発表会の記録を整理する専門アシスタントです．
以下は，発表者と質問者の途中の会話を文字起こししたテキストです．
このテキストから，実際に述べた要点だけを抽出してください．
推測や補完は禁止です．話されていないことを想像して書かないでください．

【出力形式】
- 発話者が説明・主張・条件・前提・結果・評価・意見・補足として述べた内容を箇条書きにして下さい．
- 箇条書きの結果のみを出力してください．

【制約条件】
- 実際の発話に存在しない情報を補完・推測しないこと．
- 言い換えはしてもよいが，意味を変えない．
- 発表者や質問者など役割ごとにまとめることはせず一括で出力すること．
-可能な限り多く，重複しない主張を抽出すること．

=== 文字起こし ===
transcription
```

主張抽出プロンプトは，発話内容から実際に述べられた要点のみを安定して抽出するため以下の特徴を持つように設計した．

- **推測と補完の禁止**
  「推測や補完は禁止」と明示することによって，入力テキストに存在しない情報が LLM の憶測によって出力されることを防ぐ．
- **意味の保持を重視した言い換えの許可**
  意味を変えない範囲での言い換えを許可することによって，発話の冗長性を減らし，要点を簡潔に表現できるようにする．
- **役割ごとに分けない一括出力**
  発表者や質問者といった役割で分けずに一括で出力させることで，質問文のみを主張として抽出することを防ぎ，発話全体の文脈に基づかない不必要な補足の生成を減らす．

例えば，「提示する応答っていうのがそのまま表示するものが 1 つで事実性が低い情報にハイライトしたものが 1 つで事実性が高い情報にハイライトしたものが 1 つでその提案手法である事実性が低い情報を隠したものっていうのが 1 つで最後事実性が低い情報っていうのを曖昧な表現に変更したものっていうのを 1 つで用意しています」という発話から「提示する応答は、そのまま表示するもの、事実性が低いものまたは高いものにハイライトしたもの，隠したもの，曖昧な表現に変えたものである」という主張を抽出する．

3.4 主張に対する関連文章の取得

主張に対する関連文章を取得するために根拠となる文書を埋め込みベクトルに変換し，RAG を用いる際に検索できるようにする．RAG に用いる外部データとして今回の研究では論文の PDF ファイルを対象とした．まず，PDF を読み込みテキストチャンクに分割する．チャンク分割を行うことで，主張内容と対応する局所的な文脈を含む記述を効率的に検索でき，発表内容を根拠となる文書の記述に基づいて検証したうえで，訂正文および補足文を生成することが可能となる．次に，得られたテキストチャンクに対して，埋め込みモデルを用いて埋め込みを行う．各チャンクは独立してベクトル化され，論文内の内容を表現する検索用ベクトルとして保存される．検索時には，入力となる主張文の埋め込みベクトルとのコサイン類似度を用いて各チャンクとの適合度を算出し，適合度の高い上位 k 件のチャンクを取得する．

3.5 訂正文および補足文の生成

主張抽出によって得られた主張文に対して，発表内容の誤りや情報不足を明らかにするため RAG を用いて訂正文および補足文の生成を行う．例えば，「nDCG の算出式における対数の底を変更するとスコアが変化する.」という主張に対し，nDCG の算出式の対数の底を変化させても nDCG のスコアは変化しないという点を示した訂正文を提示する．また，nDCG の定義は 2 種類存在することなどを補足文として示すことを想定している．

本研究では，訂正と補足を同一の処理として扱わず，訂正用および補足用に別々のプロンプトを設計し，個別に処理を行う．これは，誤りの修正と情報の追加では判断基準や出力内容が異なるため，個別のプロンプトで生成した方が訂正文および補足文生成の精度が高くなると考えたからである．

以下は訂正文および補足文の生成のために与えたプロンプトである．以下のプロンプトにおいて生成された主張は変数 claim に，埋め込みベクトルに変換された外部データは変数 context に格納されている．

訂正文生成のプロンプト

対象テキスト:
claim
参照可能な外部情報:
context
このテキストは論文紹介のプレゼンテーションの内容を音声から書き起こし，主張をまとめたものです．
発表者の主張が外部データである論文の情報と明らかに異なっている場合や，一般的な知識や常識と異なる場合訂正が必要な箇所とみなし，以下を簡潔に示してください：

・誤っている箇所の抜粋
・誤りの理由
・正しい情報

【制約条件】
-文章が異なっていても意味が同じ場合は訂正箇所とはみなさないでください．
-誤字がある場合でも意味が通じる場合は訂正箇所とはみなさないでください．
-訂正が不要な場合は「(何も出力しない)」で返してください．
-可能な限り多く，重複しない訂正を抽出すること．

補足文生成のプロンプト

対象テキスト:
claim
参照可能な外部情報:
context
このテキストは論文紹介のプレゼンテーションの内容を音声から書き起こし，主張をまとめたものです．
外部データの論文を紹介するにあたって重要な前提知識・定義・根拠・背景説明が欠けている場合，以下を含む補足情報を簡潔に示してください：

・補足が必要な箇所の抜粋
・補足情報

【制約条件】
-補足を行う際、誤った情報の訂正は行わないでください．
-補足が不要な場合は「(何も出力しない)」で返してください．
-可能な限り多く，重複しない訂正を抽出すること．

訂正文の生成では，抽出された主張が外部データである論文の内容，あるいは一般的な知識や常識と明らかに異なる場合に，訂正が必要な箇所として検出する．プロンプトでは，誤っている箇所の抜粋，誤りの理由，および正しい情報を簡潔に出力するよう指示する．一方で表現が異なっていても意味が同じ場合は訂正対象としない．もし訂正が不要な場合には何も出力しないように指示している．

補足文の生成では，主張自体に誤りはないものの論文を理解する上で重要な前提知識，定義，根拠，背景説明が不足している場合に補足文を生成する．プロンプトでは，補足が必要な箇所の抜粋と対応する補足文を簡潔に示すよう指示する．また補足生成では誤った情報の訂正は行わず，訂正文の生成の際と同様に補足が不要な場合には何も出力しないものとする．

4 実　　験

本研究では，提案手法の有用性を検証するため，主張抽出の精度評価と，主張抽出を行うことによって訂正文および補足文の生成精度が向上するかどうかを評価した．

4.1 実験に用いたデータ

研究室内で行われた発表を対象にデータを収集し，実験に用いた．具体的には，著者らが所属する兵庫県立大学山本研究室に所属している大学生または大学院生 11 名を対象とし，各自が選んだ論文に関する発表を実施した．実験の際，すべての参加者から研究目的およびデータの利用に関する説明を事前に行い，同意書への署名を得た．データの収集は，2025 年 11 月 7 日に実施した．

発表者には，質疑応答を含めて約 10 分間で発表を行うよう求め，その際に作成したスライドを基に内容を説明してもらった．この発表において，発表および質疑応答における音声データ，スライド映像，ならびに発表資料（スライド）と各自が選んだ対象論文を取得した．実験では，収集した発表および質疑応答の音声データから音声を文字起こししたテキストを入力として，提案手法による訂正文および補足文の生成を行った．

4.2 実 験 設 定

本実験では，3 節で述べた提案システムを用いて，主張抽出と訂正文および補足文生成の性能を評価した．

音声認識モデル: 音声認識モデルには，OpenAI が提供する Whisper（API）[1]およびローカル環境で動作する Whisper（Local）[2]（Multilingual model，medium）の 2 種類を使用した．Whisper（API）は，著者らが試した限り高い文字起こし精度を示した一方で，API を利用するためコストが発生する．一方，Whisper（Local）はローカル環境で動作するため利用コストはかからないが，文字起こし精度は Whisper（API）と比較して低い傾向がある．このような精度とコストのトレードオフを考慮し，本研究では両者を

1：https://openai.com/ja-JP/index/introducing-chatgpt-and-whisper-apis/
2：https://github.com/openai/whisper

比較対象として採用した．

主張抽出のモデル: 主張抽出には，Google が開発したローカルで動作する LLM である gemma3:12b [20] を使用した．

エンベディングモデル: RAG における外部データの埋め込みの際，PDF からのテキスト抽出には Python ライブラリ fitz を用いた．抽出したテキストは，文脈のまとまりを考慮しつつ 200 字ごとに分割し，各テキストを 1 チャンクとして扱った．各チャンクは Google による埋め込みモデルであり多言語に特化している embeddinggemma:300m[3] を用いて埋め込みベクトルに変換し，検索用データベースとして保存した．検索時には，入力となる主張文を同様に埋め込みベクトルに変換し，各チャンクとのコサイン類似度に基づいて適合度を算出し，上位 5 件のチャンクを取得した．

訂正文および補足文生成のモデル: 訂正文および補足文の生成には，gemma3:12b と OpenAI の GPT-4o-2024-11-20（GPT-4o）[4]の 2 つのモデルを使用した．GPT-4o は，文脈を安定して理解でき，訂正文および補足文生成においてプロンプトにより忠実に出力を生成できると考えたため，本研究では GPT-4o を採用した．一方で，gemma3:12b はローカル環境で動作可能なモデルであり，外部 API を利用せずに運用できるという利点を有する．本研究では，提案手法が特定のモデルに依存せず有効に機能するかを検証するため，性能特性の異なるこれら 2 種類のモデルを用いて比較を行った．

4.3 評価指標

4.3.1 主張抽出の評価指標

主張抽出の性能を評価するため，人手により作成した理想的な主張の集合と，システムによって抽出された主張の集合を用いて評価を行った．まず，発話内容から抽出されるべき m 個の理想的な主張の集合を $A = \{a_1, a_2, \ldots, a_m\}$ と定義する．ここで，各要素 a_i（$1 \leqq i \leqq m$）は，人手により作成された主張の 1 単位を表す．なお，理想的な主張の集合は，著者が作成したものである．次に，実際に抽出された n 個の主張の集合を $B = \{b_1, b_2, \ldots, b_n\}$ と定義する．ここで，各要素 b_j（$1 \leqq j \leqq n$）は，システムが出力した主張の 1 単位を表す．集合 A と B の要素間で，意味的に同一であると人手で判断された対応関係の集合を $A \cap B$ とする．なお，文面が完全に一致しない場合であっても，意味が同一であると判断できる場合は一致とみなす．この一致判定は，理想的な主張の集合とシステム出力の集合が対応しているかどうかを著者が人手で判断することにより行った．

このとき，主張抽出における適合率（Precision），再現率（Recall），および F_1 値をそれぞれ以下の式で求める．

$$\text{Precision} = \frac{|A \cap B|}{|B|}$$

$$\text{Recall} = \frac{|A \cap B|}{|A|}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.3.2 訂正および補足文生成の評価指標

訂正文および補足文生成の性能を評価するため，主張抽出と同様に，人手で作成した理想的な出力例と，システムによる生成結果を比較する．訂正文と補足文は異なるタスクであるが，同一の評価方法を用いた．まず，理想的な o 個の訂正文または補足文の集合を $C = \{c_1, c_2, \ldots, c_o\}$ と定義する．ここで，各要素 c_k（$1 \leqq k \leqq o$）は，人手により作成された訂正文または補足文の 1 単位を表す．なお，理想的な訂正文および補足文の集合は，著者が作成したものである．次に，提案手法または比較手法によって生成された p 個の訂正文または補足文の集合を $D = \{d_1, d_2, \ldots, d_p\}$ と定義する．ここで，各要素 d_l（$1 \leqq l \leqq p$）は，システムが生成した訂正文または補足文の 1 単位を表す．集合 C と D の要素間で，意味的に同一であると人手で判断された対応関係の集合を $C \cap D$ とする．文面が完全に一致しない場合であっても，訂正内容や補足内容の意味が同一であると判断できる場合は一致とみなす．この一致判定は，理想的な訂正文および補足文の集合とシステム出力の集合が対応しているかどうかを著者と指導教員が人手で判断することにより行った．このとき，訂正文および補足文生成における適合率，再現率，および F_1 値を 4.3.1 節と同様に求めた．

4.4 比較手法

実験では，提案手法の有用性を検証するため，主張抽出を行うことで訂正文および補足文生成の精度が向上するかという点と，使用するモデルの違いが生成精度に与える影響という 2 点に着目し，主張抽出と訂正文および補足文生成の各段階において条件を整理した比較を行った．

4.4.1 主張抽出の比較手法

主張抽出の比較手法として，音声認識手法の違いによる影響を検証するため，Whisper（API）と Whisper（Local）の 2 種類の音声認識結果を用いて主張抽出を行い，それぞれの抽出精度を比較した．この際，主張抽出に用いる LLM やプロンプトは同一の条件とした．

4.4.2 訂正および補足文生成の比較手法

訂正文および補足文生成の比較では，以下の手法を対象とした．

はじめに，主張抽出の有無による比較を行った．一つ目は，Whisper（API）および Whisper（Local）によって文字起こしされたテキストから主張抽出を行い，得られた主張を入力として訂正文および補足文を生成する手法である．二つ目は，主張抽出を行わず，音声を文字起こししたテキストを直接入力として訂正文および補足文を生成する手法である．主張抽出を行わない手法は，発話内容全体をそのまま入力として生成を行う一般的な方法に相当するため，提案手法の有効性を検証する比較対象として適切であると考えられる．この比較により，発話内容から主張を抽出することが，訂正文および補足文生成の精度向上に寄与するかを明らかにする．

3：https://huggingface.co/google/embeddinggemma-300m

4：https://platform.openai.com/docs/models/gpt-4o?snapshot=gpt-4o-2024-11-20

表 1　主張抽出の評価結果.

音声認識のモデル	適合率	再現率	F_1 値
Whisper（Local）	0.55	0.60	0.56
Whisper（API）	0.65	0.71	0.67

次に，使用するモデルの違いが生成精度に与える影響を検証する．訂正文および補足文生成に用いる LLM として GPT-4o および gemma3:12b を使用し，同一条件下で比較を行った．これにより，提案手法の性能が特定のモデルに依存するか，あるいは複数のモデルにおいて一貫した傾向が得られるかを分析する．

また，これらの比較においては，使用する外部データや RAG の設定など，主張抽出およびモデル以外の条件を可能な限り統一することで，各要因が生成精度に与える影響を個別に評価できるようにした．

4.5 実験結果

提案手法による実験結果について述べる．主張抽出の精度評価には，研究室内で実施した論文紹介プレゼンテーションに参加した 11 名分のデータを用いた．一方で，訂正文および補足文の生成精度の評価については，評価データの作成に詳細な評価が必要なことから，2 名分のデータを対象として評価を行った．

4.5.1　主張抽出の精度の結果

表 1 に，主張抽出の評価結果を示す．表中の値は 11 名分データそれぞれについて主張抽出の評価を行いマクロ平均をとったものである．Whisper（API）を用いた場合，適合率 0.65，再現率 0.71，F_1 値 0.67 となり，Whisper（Local）を用いた場合の適合率 0.55，再現率 0.60，F_1 値 0.56 と比較して，すべての指標において高い値を示した．

4.5.2　主張抽出の有無による訂正および補足生成の結果

表 2 および表 3 に，訂正文および補足文生成の評価結果を示す．表中の値は 2 名分データそれぞれについて訂正文および補足文生成の評価を行いマクロ平均をとったものである．

まず訂正文生成について見ると，主張抽出を行わずに文字起こしテキストを直接入力した場合はモデルを問わず適合率，再現率，F_1 値はいずれも 0.00 となり，有効な訂正文を生成できなかった．一方で，主張抽出を行った場合，GPT-4o を用いた条件において，適合率 0.17，再現率 0.25，F_1 値 0.20 となり，訂正文生成が可能であることが確認された．

次に補足文生成については，表 3 に示すように，主張抽出を行った条件で高い精度が得られた．特に，GPT-4o を用いた場合には，適合率 0.65，再現率 0.58，F_1 値 0.59 と最も高い値を示した．これに対し，主張抽出を行わない場合や，補足文生成に gemma3:12b を用いた場合には，F_1 値が低下する傾向が見られた．

表 2　訂正文生成の評価結果.

主張抽出	音声認識モデル	訂正文生成のモデル	適合率	再現率	F_1 値
なし	Whisper（Local）	gemma3:12b	0.00	0.00	0.00
あり	Whisper（Local）	gemma3:12b	0.00	0.00	0.00
なし	Whisper（API）	GPT-4o	0.00	0.00	0.00
あり	Whisper（API）	GPT-4o	0.17	0.25	0.20

表 3　補足文生成の評価結果.

主張抽出	音声認識モデル	補足文生成のモデル	適合率	再現率	F_1 値
なし	Whisper（Local）	gemma3:12b	0.07	0.06	0.07
あり	Whisper（Local）	gemma3:12b	0.33	0.29	0.28
なし	Whisper（API）	GPT-4o	0.34	0.29	0.32
あり	Whisper（API）	GPT-4o	0.65	0.58	0.59

5 議　　論

5.1 主張抽出の議論

表 1 の結果より，主張抽出の精度は音声認識結果の品質に大きく影響されることが分かる．Whisper（API）を用いた場合は Whisper（Local）と比較して，適合率，再現率，F_1 値のすべてが高く，特に再現率が高い傾向が見られた．これは，音声認識の誤りが少ないほど，発話中の重要な説明や断定的な内容を LLM が正確に把握しやすくなるためであると考えられる．一方で，Whisper（API），Whisper（Local）のいずれにおいても，発話内容に論文由来の専門用語や略語が多く含まれる場合には，単語の認識誤りが増加する傾向が確認された．このような音声認識の誤りは，主張抽出において主張の欠落や意味の誤解釈を引き起こす要因となっていた．例えば，発表において「低リソース言語に対応した指示追従モデルを構築したい」という内容の発話が行われた場合，音声認識モデルによる文字起こしでは「この論文何をしているかと言いますとペリソース言語に対応した市立移住モデルを構築したいという論文になります」といった誤認識が生じることがあった．このような誤った文字起こし結果に基づいて主張抽出を行うと，本来の意味とは異なる，「ペリソース言語に対応した市立移住モデルを構築したい」といった意味の通らない主張が抽出されてしまった．

また，発表者が事前に作成した資料を基に，論文の背景，提案手法，実験設定，結論などを説明している発話においては，主張抽出が質疑応答の発話よりも高い精度で行われていることが確認された．これらの発話は，スライド構成に沿って論理的に説明されることが多く，定義や結果に関する断定的な表現が明確である一方で，言い直しや曖昧な言い回しが少ないという特徴を持つ．そのため，音声認識結果が一定の品質を満たしている場合，LLM が発話の要点を把握しやすく，主張として抽出すべき内容を安定して取り出すことが可能であったと考えられる．

一方で，Whisper（API）を用いた場合でも F_1 値は 0.67 にとどまっており，すべての主張を完全に抽出できているわけではない．特に，質疑応答における発話では，主張が正しく抽出されないケースが確認された．例えば，「論文の実験は倫理審査を通していたか」という質問に対し，発表者は複数の発話を経た後に「分からない」と回答していた．しかし，発話の途中に含まれる言い直しや曖昧な表現に加え，音声認識の誤りの影響により，会話の文脈が十分に反映されず，抽出された主張で

は Whisper（API）および Whisper（Local）のいずれにおいても「この論文の実験は倫理審査に通っている」という誤った内容が出力された．このように，質疑応答のような対話的で断片的な発話に対しては，音声認識の誤りと文脈理解の困難さが重なり，誤抽出が生じる可能性があることが示唆される．事前に内容を整理した発話と突発的な発話とでは，主張抽出の難易度に差があると考えられる．

5.2 主張抽出の有無による訂正文および補足文生成の議論

表 2 および表 3 の結果より，主張抽出を行うことで，訂正文および補足文の生成精度が向上することが確認された．特に補足文生成においては，主張抽出を行った条件で適合率，再現率，F_1 値が大きく改善しており，発話内容から主張を抽出することが，補足対象の明確化に有効であることが示唆される．これは，発話内容全体をそのまま入力とする場合と比較して，重要な説明や断定的な内容のみが入力として与えられることで，RAG による情報検索および生成が適切に機能したためであると考えられる．

訂正文生成においても，全体として精度は高くないものの，主張抽出を行うことで有効な訂正文が生成された事例が確認された．具体的には，発表中に「評価実験 1 と評価実験 2 の違い」について説明する場面において，発表者が不明確な情報を発言していた事例が挙げられる．この場面での文字起こしされた具体的な発話は「1 と 2 差がトピックと人数以外にあるのかどうか論文を見た感じでは，えーっと人数とテーマと，あと，参加している人が，評価実験 1 は結構大学とかその 身内で，2 はそれ以外の一般の人に参加を募ってやっているっていう差があって，他は多分ほとんど同じ内容で実験していると思います.」となっている．この発話において発話中に言い直しや補足的な説明が含まれていたため，主張抽出を行わずに文字起こしテキストを直接入力した場合には，訂正すべき主張が明確にならず，訂正文は出力されなかった．一方で，主張抽出を行った場合には，「評価実験 1 と 2 は，人数とテーマ以外に、評価実験 1 は身内，2 は一般の人に参加を募っている点が異なる.」という主張が抽出されており，この主張を入力として RAG による照合を行うことで，「評価実験 1 の参加者はファシリテーション協会の会員とその知り合いであり，評価実験 2 の参加者は外部委託業者によって募集された一般の人々である.」という適切な訂正文が生成された．この結果は，主張抽出が訂正文生成においても重要な前処理として機能することを示唆している．一方で，訂正文生成の全体的な評価値が低くなった要因として，訂正文生成のプロンプトにおいて「誤字の訂正は行わない」と指示していたにもかかわらず，音声認識の過程で生じた誤字を訂正対象として検出してしまうケースが多く見られた点が挙げられる．例えば，「正例」という語が音声認識により「精霊」と誤認識されていた場合，この誤字が訂正対象として出力される事例が確認された．このような出力は人手評価において不正解と判定されるため，適合率および再現率の低下につながったと考えられる．

また，訂正文生成の評価値が低くなったもう一つの要因として，一回の発表において訂正が必要な理想的な訂正文の数自体が少ない点が挙げられる．訂正すべき誤りがほとんど含まれない発表に対しても，システムが訂正文を生成しようとすることで，相対的に誤検出の影響が大きくなり，評価指標が低下した可能性がある．このことから，訂正文生成においては，訂正が本当に必要かどうかを事前に判定する仕組みの導入も重要であると考えられる．

モデル間の比較では，gemma3:12b を用いた場合に，GPT-4o と比較して訂正文および補足文生成の精度が低くなる傾向が見られた．この原因として，gemma3:12b は GPT-4o と比べてモデルサイズが小さく，複雑なプロンプトの意図を十分に理解できていなかった可能性が考えられる．特に，訂正と補足を厳密に区別する必要がある本タスクでは，プロンプト理解能力の差が出力精度に影響したと考えられる．

さらに，主張抽出を行わなかった場合には，訂正や補足を行うべき範囲が不明確となり，不適切な訂正文や補足文が生成されるケースが確認された．例えば，発表中で「例えば『1 日に何 ml の水を飲むべきですか』というクエリに対して，大人は 3000ml の水を飲むべきといったコーパスを生成したい」という説明がなされていた時に，主張抽出を行わない条件では，この例示部分に対して「男性は 3000ml，女性は 2100ml の水を飲むべきです」といった訂正文が生成された．これは，例として提示された仮定の内容が主張として誤って扱われたことによるものである．一方で，主張抽出を行った場合には，このような例示は主張として抽出されないため，不適切な訂正文および補足文の生成を防ぐことができた．この結果は，主張抽出が訂正文および補足文生成の前処理として重要な役割を果たしていることを示していると考えられる．

一方で，本研究では RAG における検索性能やエンベディング手法そのものの評価は行えておらず，訂正文および補足文の生成精度が，どの程度検索結果の品質に依存しているかについては明らかにできていない．また，訂正文および補足文生成に用いたプロンプト設計や，外部データの検索方法，エンベディングの粒度や単位についても，さらなる検討の余地がある．これらの要素を体系的に評価することが，訂正文および補足文生成の精度向上につながると考えられる．

5.3 学習支援への適用に関する課題

本研究では，発話内容から主張を抽出し，訂正文および補足文を生成する手法を提案し，主張抽出と訂正文および補足文生成の精度に関する基礎的な有用性を示した．しかしながら，本手法を実際の学習支援に適用するためには，生成された情報をどのような形で学習者に提示するかという点について，さらなる検討が必要である．

フィードバックのタイミングは学習成果形成に決定的な役割を果たす [21]．このことから訂正文や補足文を提示するタイミングも学習効果に大きな影響を与えると考えられる．例えば，発表中にリアルタイムで提示する場合には，発表者や聴講者の注意を分散させ，内容理解を妨げる可能性がある一方で，誤りや不足に即座に気づかせるという利点もある．一方で，発表後

にまとめて提示する場合には，誤った理解がその場で修正されないまま進行してしまう可能性がある．このように，提示タイミングの設計は，学習者の理解度や発表の目的に応じて慎重に検討する必要がある．

また，提示する情報量や表現方法も重要な課題である．訂正や補足が過剰に提示されると，学習者にとっては情報量が多くなりすぎ，どの点が重要なのか分かりにくくなる恐れがある．特にアウトプット型学習の場面では，学習者自身が考えながら説明を行うことが重要であるため，訂正や補足が一方的に与えられすぎると，主体的な学びを阻害する可能性も考えられる．そのため，訂正と補足の重要度に応じて提示量を調整したり，簡潔な要約として提示したりするなど，情報の取捨選択が求められる．

さらに，本研究では生成された訂正文および補足文の正確性を評価しているが，それらが実際に学習者の理解や学習成果にどのような影響を与えるかについては検証できていない．例えば，訂正文を提示された学習者が，自身の誤りをどの程度理解し，その後の学習に活かすことができているか，あるいは補足文が理解の深化につながっているかといった教育的効果については，別途評価が必要である．

このように，本手法を学習支援システムとして実運用するためには，主張抽出や訂正文および補足文生成の精度向上に加えて，提示タイミング，提示量および学習者への影響といった観点から総合的な検討が必要である．今後は，実際の教育現場での利用を想定したインタフェース設計や運用方法の検討を行い，学習者の理解促進や主体的な学びにどのように寄与するかを実証的に評価していくことが課題である．

6 まとめと今後の課題

本研究では，アウトプット型学習における発表内容の理解支援を目的として，発話音声から話者の主張を抽出し，RAG に基づいて訂正文および補足文を生成するシステムを提案した．提案手法では，音声認識によって得られた文字起こしデータから LLM を用いて主張を抽出し，抽出された主張を単位として根拠となる文書との照合を行うことで，発表における発話に含まれる訂正文および補足文を生成する．

研究室内で行われた論文紹介のプレゼンテーションを対象とした実験の結果，主張抽出を導入することで，訂正文および補足文の生成精度が向上することを確認した．特に，主張抽出を行わずに文字起こしテキストを直接入力した場合と比較して，主張を抽出した場合には，訂正および補足の対象が明確化され，RAG による情報検索および生成が効果的に機能することが示された．また，発表資料に基づいて論理的に説明される発話においては，主張抽出が質疑応答の際の発話よりも高い精度で行われることも確認された．

一方で，本研究にはいくつかの課題が残されている．主張抽出と訂正文および補足文生成の精度は，音声認識結果の品質に大きく依存しており，特に専門用語や略語を含む発話や，質疑応答における断片的で言い直しの多い発話に対しては，誤った抽出が生じる場合があった．また，訂正文生成においては，音声認識の誤りによる誤字が訂正対象として検出されてしまうなど，訂正が本当に必要な主張を見極めることの難しさも明らかになった．

さらに，本研究では生成された訂正文および補足文の正確性を中心に評価を行ったが，それらの情報が学習者の理解や学習成果にどのような影響を与えるかについては検証できていない．訂正や補足を提示するタイミングや提示量，提示方法によっては，学習者の注意を分散させたり，主体的な学びを阻害したりする可能性も考えられる．

今後の課題としては，質疑応答のような対話的な発話に対して複数の発話を統合して主張を抽出する手法の検討や，訂正が本当に必要かどうかを事前に判定する仕組みの導入が挙げられる．また，実際の教育現場での利用を想定し，訂正文および補足文の提示タイミングやインタフェース設計を含めた運用方法についても検討する必要がある．これらの課題に取り組むことで，アウトプット型学習をより効果的に支援する実用的な学習支援システムの構築を行っていきたい．

謝　　辞

本研究は，JSPS 科研費 JP24K03228 の助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] 文部科学省高等教育局. 令和 3 年度の大学における教育内容等の改革状況について（概要）. https://www.mext.go.jp/content/20230908-mxt_daigakuc01-000031526_1.pdf. 2025 年 1 月 15 日閲覧.

[2] 中央教育審議会. 新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け、主体的に考える力を育成する大学へ～(答申). https://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2012/10/04/1325048_1.pdf. 2025 年 1 月 15 日閲覧.

[3] Shota Imamura, Hirotaka Hiraki, and Jun Rekimoto. Serendipity Wall: A discussion support system using real–time speech recognition and large language model. In *Proceedings of the Augmented Humans International Conference 2024*, pp. 237–247, 2024.

[4] 木下良輔, 櫻井崇貴, 白松俊. LLM を用いたファクトチェック機能の試作と Web 議論における関連情報推薦システムへの応用. 第 12 回市民共創知研究会 2023 (CCI-012), pp. 41–44, 3 2024.

[5] Steffen Steinert, Karina E. Avila, Stefan Ruzika, Jochen Kuhn, and Stefan Küchemann. Harnessing large language models to develop research–based learning assistants for formative feedback. *Smart Learning Environments*, Vol. 11, No. 62, 2024.

[6] Frederic Kirstein, Terry Ruas, Robert Kratel, and Bela Gipp. Tell me what I need to know: Exploring LLM–based (personalized) abstractive multi–source meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 920–939, 2024.

[7] Yun Long and Yu Zhang. Enhanced classroom dialogue sequences analysis with a hybrid AI agent: Merging expert rule-base with large language models. *arXiv preprint arXiv:2411.08418*, 2024.

[8] Yang Shi, Yang Wang, Ye Qi, John Chen, Xiaoyao Xu, and Kwan-Liu Ma. IdeaWall: Improving creative collaboration through combinatorial visual stimuli. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative*

Work and Social Computing, pp. 594–603, 2017.

[9] Salvatore Andolina, Khalil Klouche, Diogo Cabral, Tuukka Ruotsalo, and Giulio Jacucci. InspirationWall: Supporting idea generation through automatic information exploration. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pp. 103–106, 2015.

[10] 中明理沙, 吉添衛, 服部宏充. 議論支援システム AIR-VAS への LLM に基づく議論エージェントの導入とその効果. 人工知能学会全国大会論文集, 2F6GS505, 2021.

[11] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. The Idea Machine: LLM–based expansion, rewriting, combination, and suggestion of ideas. In *Proceedings of the 14th Conference on Creativity and Cognition*, pp. 623–627, 2022.

[12] Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. Idea Expander: Supporting group brainstorming with conversationally triggered visual thinking stimuli. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pp. 103–106, 2010.

[13] Rrubaa Panchendrarajan and Arkaitz Zubiaga. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *arXiv preprint arXiv:2401.11969*, 2024.

[14] Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. Claim extraction for fact–checking: Data, models, and automated metrics. *arXiv preprint arXiv:2502.04955*, 2025.

[15] Nhat Tran, Diane Litman, and Amanda Godley. Using large language models to analyze students' collaborative argumentation in classroom discussions. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference*, pp. 111–125, Pittsburgh, PA, USA, 2025.

[16] Chitralekha Gupta, Hanjun Wu, Praveen Sasikumar, Shreyas Sridhar, Priambudi Bagaskara, and Suranga Nanayakkara. Factually: Exploring wearable fact–checking for augmented truth discernment. In *Proceedings of the 2025 ACM Workshop on Human–AI Interaction for Augmented Reasoning*, 2025.

[17] Venktesh V and Vinay Setty. LiveFC: A system for live fact–checking of audio streams. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 1060–1063, 2024.

[18] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of Varying Shades: Analyzing language in fake news and political fact–checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2931–2937, 2017.

[19] Burges Chris, Shaked Tal, Renshaw Erin, Lazier Ari, Deeds Matt, Hamilton Nicole, and Hullender Greg. Learning to rank using gradient descent. In *ACM ICML 2005*, pp. 89–96, 2005.

[20] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[21] Hongyu Mai. The comparative effect of immediate and delayed feedback on EFL learners' engagement and willingness to collaborate. *PsyCh Journal*, pp. 1008–1017, 2025.

基本感情の複合分析に基づく表情画像からの困惑深度推定 : Transformer を用いた特徴抽出と評価

石川 昂樹† 一色 夢香† 寺田 憲司† 遠藤 雅樹† 大野 成義†

†職業能力開発総合大学校 電子情報専攻 情報通信ネットワークユニット 〒187-0035 東京都小平市小川西町 2-32-1

E-mail: †{b22302, b22305, k-terada, endou, ohno}@uitec.ac.jp

あらまし 急速に常態化した非対面の学習・作業環境での困惑状態の把握は作業停滞の早期発見や心理的安全性の確保を実現しうるデジタル社会に求められる技術である．そこで本研究では，困惑が「怒り」,「悲しみ」等の複数の基本感情の要素を含んでいる可能性に着目する．具体的には，作業時に取得した表情画像から学習済み深層学習を用いて複数の基本感情を抽出し，教師付き機械学習によって困惑を推定する手法を提案する．実験では 7 名の被験者に対しプログラミング課題を課し，Web カメラで表情を撮影した．表情画像から Transformer を用いて 7 つの基本感情スコアのデータセットを構築した．さらに，データセットには特徴量増加を施し，説明変数を生成した．プログラミング課題中に 4 段階の困惑深度のアンケートを実施し，目的変数とした．機械学習による困惑推定モデルを構築した結果，各困惑深度の平均 F 値は 0.8882 となった．研究から困惑は基本感情を組み合わせた複合感情により推定でき，被験者毎に困惑推定が可能であるという知見を得た．

キーワード パターン認識，深層学習，機械学習，感情分類，生体センシング

1. はじめに

近年，社会全体においてオンライン環境での作業が普及している.テレワークの導入率は 2018 年から 2024 年にかけて 20%以上増加している[1]．また，教育現場においても大学生のオンライン授業の受講率は約 80%に達している[2]．しかし，このようなオンライン環境での作業の普及に伴い課題も顕在化している．オンライン授業中に教員が受講生の反応が見えないという回答が約半数を占めており[3]，教員の約 60%が学生の困惑が把握できないことが報告されている[4]．

特にオンライン環境におけるプログラミング初学者を対象とした授業実践では，学習者がつまずきの原因を言語化できず，質問機能を用いても適切な支援を求められないケースがあることが指摘されており[5]，オンライン環境下における潜在的な困惑の検知は困難を極める．そのため，非対面の学習作業環境での困惑状態の把握は，作業停滞の早期発見や心理的安全性の確保を実現しうるデジタル社会に求められる技術である．

しかし，困惑の表情の検出には課題がある．第一に，困惑の表情は他の思考中の表情と類似していることが示唆されており，その識別の複雑さが浮き彫りになっている[6]．第二に，学習者が困難に直面した際に生じる「困惑」が重要であるが，表情のみでは中立的な表情や他の感情と誤認識されやすい[7]．第三に，学習中の困惑は頻繁に発生するが，表情は微妙で検出しにくいことも報告されている[8]．

そこで本研究では，プログラミング作業時の表情から学習済み深層学習を用いて複数の基本感情を抽出し，教師付き機械学習によって困惑深度を推定することを目的とする．

2. 関連研究

2.1. 困惑の表情特徴と既存検出手法の課題

Yasser ら[9]の研究によれば，眉の引き下げと瞼の緊張の組み合わせは，困惑状態において有意に頻出し，口元の動きは，困惑が生じている際の「何か言いたげだが言えない」状態や，自己嘲笑的な反応を反映していると分類している．表情から困惑を推定できるとしており，これらの特徴はポール・エクマンが提唱する「怒り（angry)」,「嫌悪（disgust)」,「恐怖（fear)」,「驚き（surprise)」,「幸福（happy)」,「悲しみ（sad)」，「中立（neutral)」の 7 つの基本感情の特徴にみられるものと類似している．

一方で，既存手法の侵襲性とコストの問題，従来の混乱検出技術の多くは，脳波（EEG）や筋電図（EMG）といった生理学的センサを身体に装着する必要がある．これらは侵襲的であり，高価で，かつユーザーの動きを制限するため，日常的な E ラーニング環境などでの利用には適していないと課題に挙げている．

2.2. 困惑検出における時間的変化の重要性

Pachman ら[10]は，困惑を単なる一瞬の表情としてではなく，時間的な広がりを持つ動的なプロセスとして捉えている．学習のきっかけとなる「建設的な混乱（Productive Confusion)」と，解決できずにフラストレーションへと移行する「非建設的な混乱（Unproductive Confusion)」に分類するとしている．困惑の検出には

Hesitation と呼ばれる意思決定や発話の前の「間」として現れる行動的シグナルを検出するためには，個別の特定の顔面微細表情の検出が必要であるとされている．

この関連研究から得られることは，一瞬の画像から得られる単なる分類モデルでは困惑推定が難しいということである．

2.3. 困惑状態の細分化の必要性

Hussain ら[11]の研究によれば，感情的干渉がある状況下では生理信号よりも顔表情の方が認知的負荷の困惑検出においてロバストであることが示されている．これは，生理反応が情動に強く反応してしまう一方で特定の顔面筋は認知的努力をより純粋に反映することを明らかにしている．但し，顔，生理信号などのマルチモーダル情報が増えると複数の情報源を統合することで一般的には検出精度が向上するが，本研究では感情的覚醒の強度が高まると統合してもなお検出精度が低下することが明らかになった．つまり，強い感情的干渉下では，単純なマルチモーダル統合だけではロバスト性を維持できない点が課題として浮き彫りになった．

この研究から，困惑の深度が高まった場合，どのような感情の変化がみられるかを明らかにすることが困惑推定で必要となることを特定した．

3. 提案手法

3.1. 概要

本研究では，困惑が基本感情の複合感情である可能性に着目し，プログラミング作業中の作業者の表情画像から困惑を推定する手法を提案する．

従来の研究では，表情画像から困惑推定モデルを構築する際，「困惑」状態をラベル付けした学習画像データが必要である．しかし，十分な困惑画像データの収集は大きな労力が必要であるという課題がある．ここで，7 つの基本感情分類には学習済み深層学習モデルが存在することに注目する．本研究では，画像から直接「困惑」を学習させるのではなく，既に確立された学習済み深層学習モデルから得られるスコア結果から困惑を分類する手法を提案する．

データ取得から評価までの提案手法を示す（図 1）．Web カメラを用い，被験者の表情を撮影し，表情画像を取得する．表情画像から深層学習を用い，7 つの基本感情スコアを取得する．このスコアからスライドウィンドウを用い，統計的特徴量増加を行った．これを説明変数とした．プログラミング作業中に 4 段階の困惑深度アンケートを実施し，これを目的変数とした．説明変数と目的変数をタイムスタンプで結合し，特徴量を生成した．教師付き機械学習を用い，困惑推定モデルを構築する．モデルは評価および複合感情分析を行う．

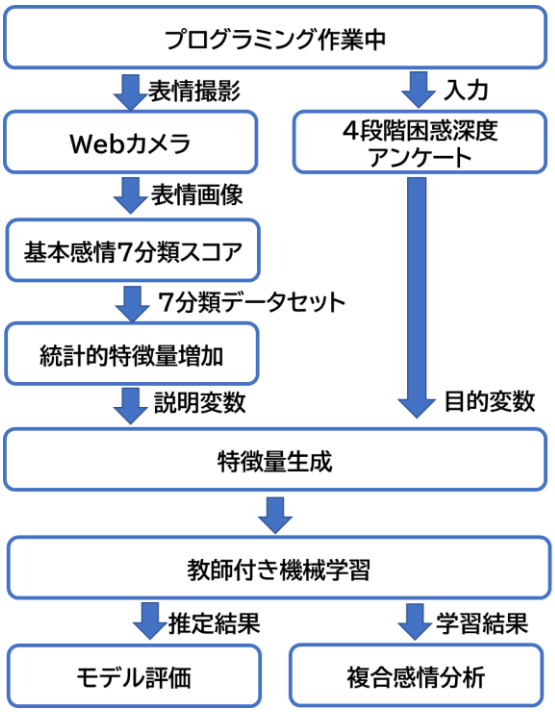


図 1　提案手法のブロック図

3.2. 特徴量

「2.2. 困惑検出における時間的変化の重要性」より，困惑は一瞬の表情としてではなく，時間的なプロセスとして捉える必要がある．そこで，機械学習を適用する際は統計的特徴量増加によって時間的変化を特徴量の中に持たせる．

基本感情の複合感情から困惑推定を行うために，表情画像から 7 つの基本感情スコアを取得する（式 1）．

$$E_{x,t} = D_t(I)x \quad (式\ 1)$$

$E_{x,t}$：基本感情スコア（0.0～1.0）
t：時間
D_t：深層学習
I：表情画像
x：an(angry)，di(disgust)，fe(fear)，su(surprise)，ha(happy)，sa(sad)，nu(neutral)

ある時間における 7 つの基本感情スコアを示す（式 2）．

$$E_{B,t} = \{E_{an,t}, E_{di,t}, E_{fe,t}, E_{su,t}, E_{ha,t}, E_{sa,t}, E_{ne,t}\} (式\ 2)$$

$E_{B,t}$：ある時刻における 7 つの基本感情スコア

スライドウィンドウを用いた（式 3）．

$$E_S = \{E_{B,t},\ E_{B,t-1},\ \cdots E_{B,t-w}\} \quad (式\ 3)$$

E_S：ローデータ

w ：ウィンドウサイズ
取得したE_Sに対し，最小値，最大値，中央値，平均値，分散，標準偏差，尖度，歪度の統計的特徴量増加を行った（式 4）.

$$F_g = \{F_{min}, F_{max}, F_{med}, F_{mea}, F_{var}, F_{std}, F_{kur}, F_{ske}\} \quad (式\ 4)$$

F_g：特徴量
F_{min}：E_Sの最小値
F_{max}：E_Sの最大値
F_{med}：E_Sの中央値
F_{mea}：E_Sの平均値
F_{var}：E_Sの分散
F_{std}：E_Sの標準偏差
F_{kur}：E_Sの尖度
F_{ske}：E_Sの歪度
ローデータと統計的特徴量増加によって説明変数に適用する特徴量数は 63 となった．機械学習を用い，予測結果を評価する(式 5).

$$P_e = M_{RF}(E_{B,t},\ F_g) \quad (式\ 5)$$

P_e：困惑深度
M_{RF}：機械学習

3.3. 評価方法

3.3.1. 機械学習モデルの評価

機械学習モデルの評価指標は F 値（F-score）を適用する．F 値は混同行列から求めることができる「適合率(Precision)」と「再現率(Recall)」のバランスを評価するための重要な指標である．特にデータセットのクラスに偏りがある場合や，誤検出と見逃しの両方を考慮したい場合に利用する．混同行列とは機械学習の分類問題において，モデルの予測結果と実際の正解値を対照させた表のことである．モデルが「どのクラスをどの程度正しく予測できたか」，あるいは「どのクラスと間違えやすいのか」を評価するための指標である．例えば二値分類（Positive / Negative）の場合，行に実際の正解クラス，列に予測クラスを配置する(表 1).

表 1　混同行列

		予測	
		Positive	Negative
実際	Positive	TP	FN
	Negative	FP	TN

混同行列を構成する 4 つの要素は以下の通りである.

- TP (True Positive)：実際も正で，予測も正
- TN (True Negative)：実際も負で，予測も負
- FP (False Positive)：実際は負だが，予測は正
- FN (False Negative)：実際は正だが，予測は負

これら 4 つの要素を利用した適合率はモデルが「正（Positive)」と予測したもののうち，実際に正であったものの割合である．すなわち，「予測の正確さ」に焦点を当てた指標である(式 6).

$$Precision = \frac{TP}{TP + FP} \quad (式\ 6)$$

再現率は，実際に「正」であるもののうち．モデルが正しく正と予測できたものの割合である．すなわち，「見逃しの少なさ」に焦点を当てた指標である(式 7).

$$Recall = \frac{TP}{TP + FN} \quad (式\ 7)$$

F 値とは適合率と再現率により求めることができる．適合率と再現率はトレードオフの関係にある．これら 2 つの指標を統合し，一つの値でモデルの性能を評価する指標が F 値である(式 8).

$$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (式\ 8)$$

3.3.2. 複合感情分析の評価

複合感情分析の評価には，Explainable AI(XAI)手法の 1 つである SHAP[12]を適用する．SHAP はゲーム理論のシャープレイ値に基づき，各特徴量が予測値の偏差に与える寄与度を算出する手法である．モデル全体において，「どのデータが分類に重要であるか」また，「互いにどう影響しているか」を分析するものである．これにより，ブラックボックス化しやすい機械学習モデルにおいて，個々の予測に対する特徴量の影響度を定量的に評価することが可能になる．本研究では Violin Summary Plot（図 2）を適用する．左側の縦軸は特徴量を示し，寄与度が大きい順に上から整列している．横軸は SHAP 値を示し，0 より右側は分類値に対する正の影響を示し，左側は負の影響を示す．各点は個々のサンプルを表し，色は特徴量の値（赤：高い，青：低い）を表している．つまり，赤い点が右側に多いほど予測に有効な特徴量となる．図 2 では「嫌悪」，「怒り」，「恐怖」が予測に有効な特徴量となる．

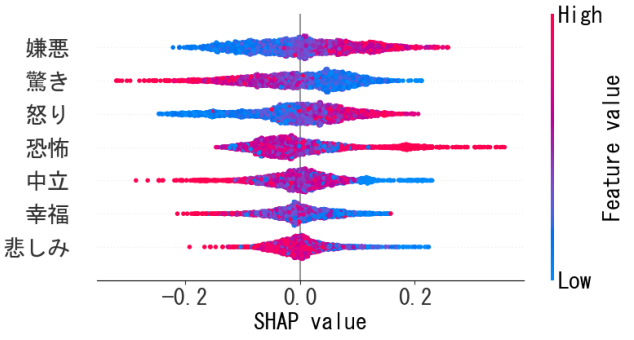


図 2　Violin Summary Plot

4. 実験

4.1. データ収集方法

プログラミング作業中の作業者の表情画像を取得することを目的とし，プログラミング課題を課した．18 歳から 22 歳の大学生 7 名が参加した．プログラミング課題に対し，内容は講義として履修したことがあるものに限定した．

本研究は，職業能力開発総合大学校倫理審査委員会の承認を受けた．

4.2. 実験環境

本実験環境を図 3 に示す．実験は，PC と 2 台のディスプレイ（サイズ：24 インチ，解像度：1920×1080）を使用した．2 台のディスプレイは横に並べ，各ディスプレイの画面を左右 2 分割し，計 4 つの領域でウィンドウを常時表示させた．ウィンドウはそれぞれ左から課題内容を表示する課題提示画面，解答コードを入力するコードエディタ，教科書などの参照資料，困惑深度アンケート画面である．コードエディタ画面の上部に記録解像度 1080P の Web カメラ（Angetube 社 967）を設置した．

図 3　実験環境

4.3. 実験手順

プログラミング課題は計 2 題実施し，設置した Web カメラを用い，表情画像を取得した．データセットの総数は 126,125 サンプルとなった．表情画像から深層学習である Transformer[13]を用い，基本感情 7 分類スコアのデータセットを取得した．このデータセットに「3.2 特徴量」にある特徴量生成を施し，説明変数とした．このとき式 3 ウィンドウサイズ w は 5 とした．プログラミング課題と並行して Web アプリケーションによる 4 段階の困惑深度アンケートを実施した．本研究における「困惑」は教育心理学における「発達の最近接領域(ZPD)」の状態[14]と近いと考えた．ZPD では学習支援の最適化において，学習者が独力では解決に至らないものの，教育的介入によって達成可能となる心理領域を指す．ZPD を通過する学習過程は「自動化」,「自己による支援」,「有能な他者による支援」の段階があるとされている．そこで，本研究ではこの理論に基づき困惑深度の評価基準を設定した（表 2）．アンケートは被験者が困惑を生じた瞬間にディスプレイ上のボタンを速やかに押下するよう指示した．ボタンが押下された時刻は，0.1 秒の精度で自動的に記録する．押下され，次に押下されるまでの困惑状態は最初に押下された状態と定義した．以上の手順で取得した困惑深度アンケート結果を機械学習の目的変数とした．取得した説明変数と目的変数をタイムスタンプで結合し，困惑推定モデルを構築した．

表 2　困惑深度評価基準

困惑深度	評価基準
0	非困惑(手が動いている状態)
1	少し困惑(手が止まり，想起する状態)
2	困惑(配布資料や教材で調べる状態)
3	とても困惑(他者に質問する状態)

4.4. 最適な分類器の調査

最適なモデルを選定するために全被験者モデルのデータセットを結合し，各機械学習モデル及び回帰分析における性能評価を実施した．本研究の考察に用いるモデルとして，最も F 値が高い機械学習モデルを採用することとした．比較した学習モデルは Random Forest(RF)[15]，Light Gradient Boosting Decision Tree(LightGBM)[16], Gradient Boosting[17]，Support Vector Machine(SVM)[18], AdaBoost[19]及び回帰分析(Logistic Regression)である．なお，各指標の算出にはクラス間のサンプル数の偏りを考慮し，マクロ平均を採用した結果を表 3 に示す．

表 3　各モデルの評価結果

モデル	再現率	適合率	F 値
RF	0.8897	0.8917	0.8882
LightGBM	0.7771	0.7787	0.7727
Gradient Boosting	0.6935	0.6929	0.6852
SVM	0.5948	0.5850	0.5726
AdaBoost	0.5689	0.5705	0.5632
Logistic Regression	0.6119	0.6032	0.6022

性能評価の結果，本研究では F 値が最も高い RF を最適な困惑推定モデルとして選定した．

4.5. 結果

教師付き機械学習 RF を用い困惑推定モデルを構築した．モデルの評価は F 値を使用し，基本感情複合分析は SHAP を使用した．

全被験者のデータセットを統合したときの混同行列を図 4 に各困惑深度における Violin Summary Plot を図 5 に示す．

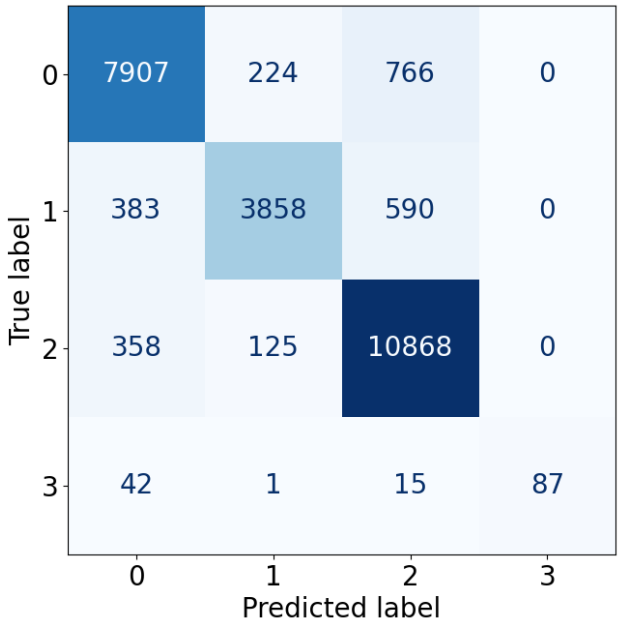
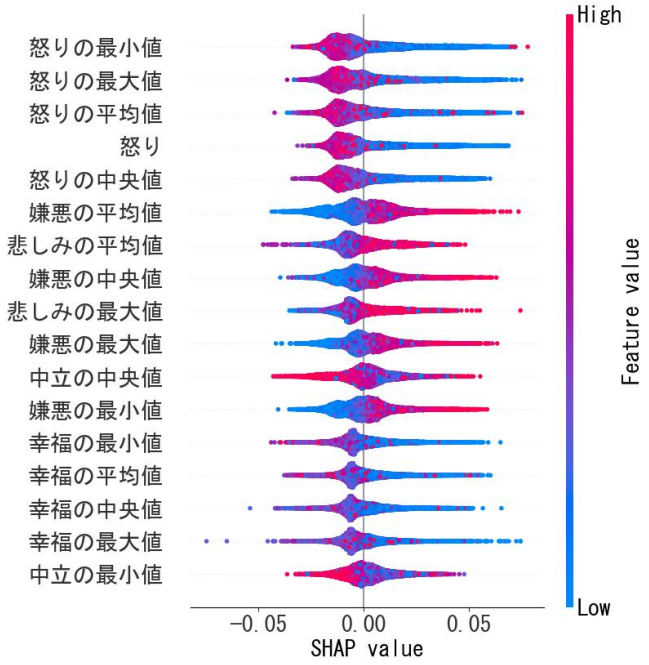
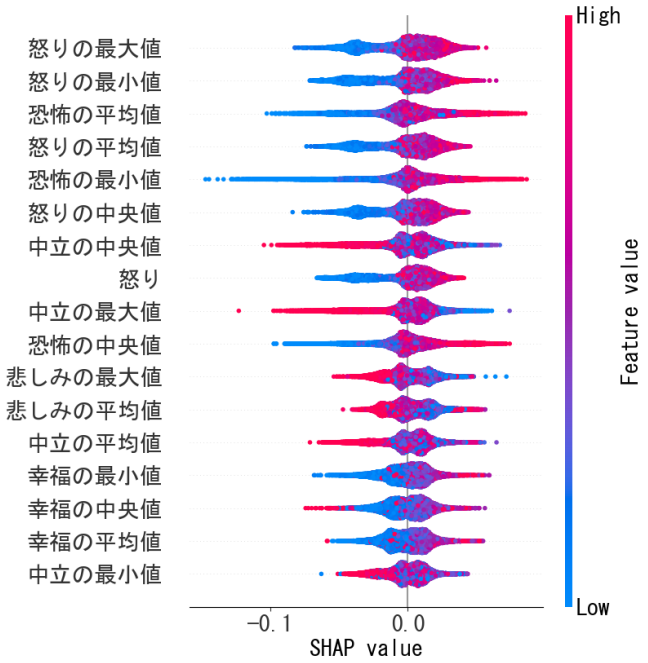


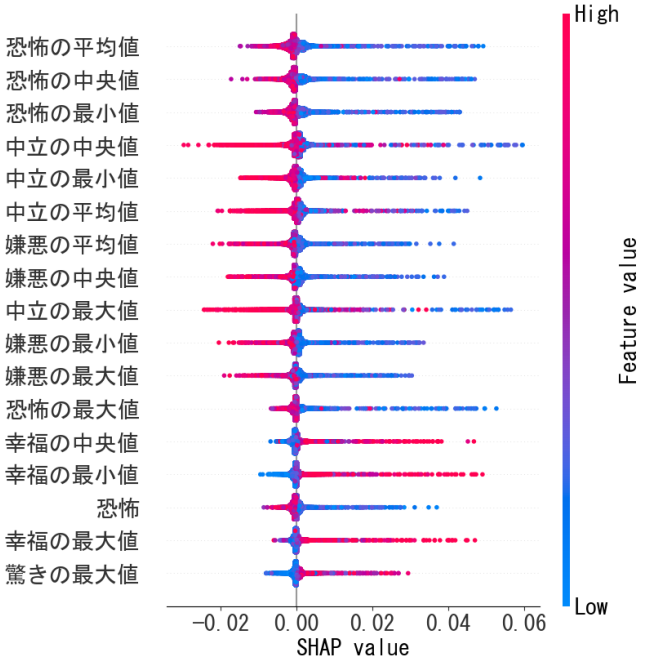
図 4　全被験者統合の混同行列



(a)　困惑深度 1



(b)　困惑深度 2



(c)　困惑深度 3

図 5　全被験者統合の Violin Summary Plot

図 5(a)から困惑深度 1 では「嫌悪」，「悲しみ」に関係する特徴量が正の方向に影響を与えることがわかった．このため，「嫌悪」，「悲しみ」の複合感情が少し困惑している状態を示唆すると考えられる．図 5(b)から困惑深度 2 では「怒り」，「恐怖」に関係する特徴量が正の方向に影響を与えることがわかった．このため，「怒り」，「恐怖」の複合感情が困惑状態を示唆すると

考えられる．図 5(c)から困惑深度 3 では「幸福」,「驚き」が正の方向に影響を与えることがわかった．このため,「幸福」,「驚き」の複合感情がとても困惑状態を示唆すると考えられる．

また，全被験者統合の SHAP 値と顕著に違いが見られた 2 名の被験者について示す．被験者 A の混同行列を図 6 に SHAP を図 7 に示し，被験者 B の混同行列を図 8 に SHAP を図 9 に示す．

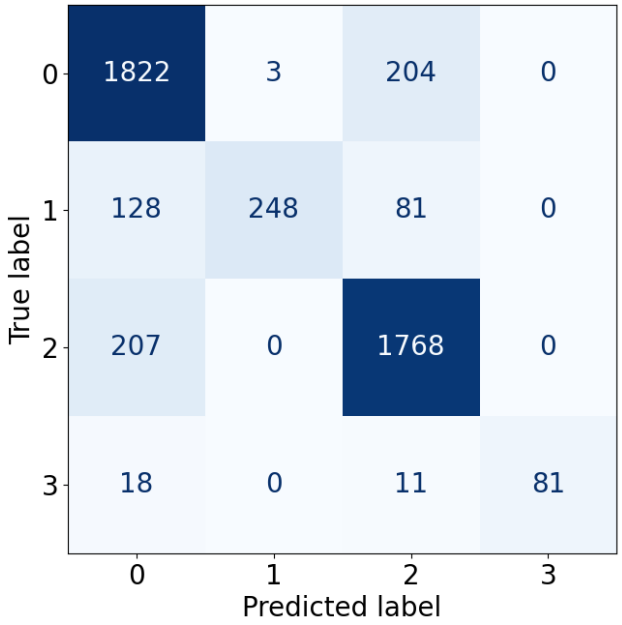
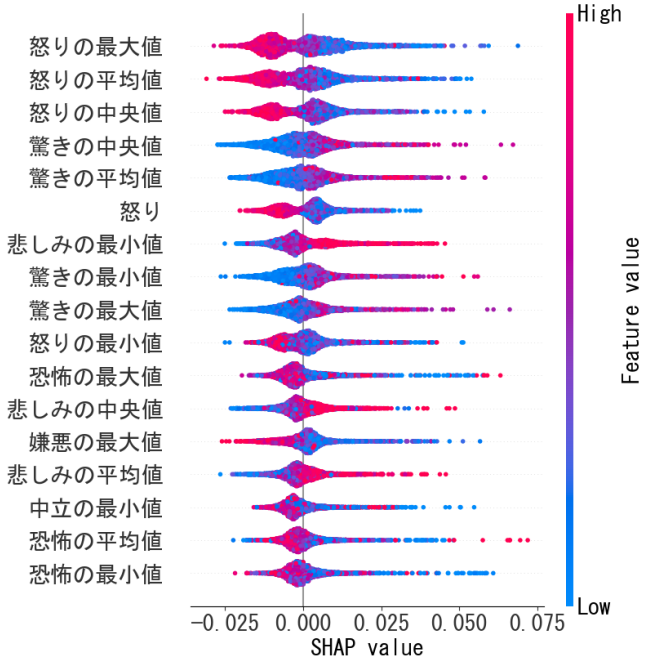
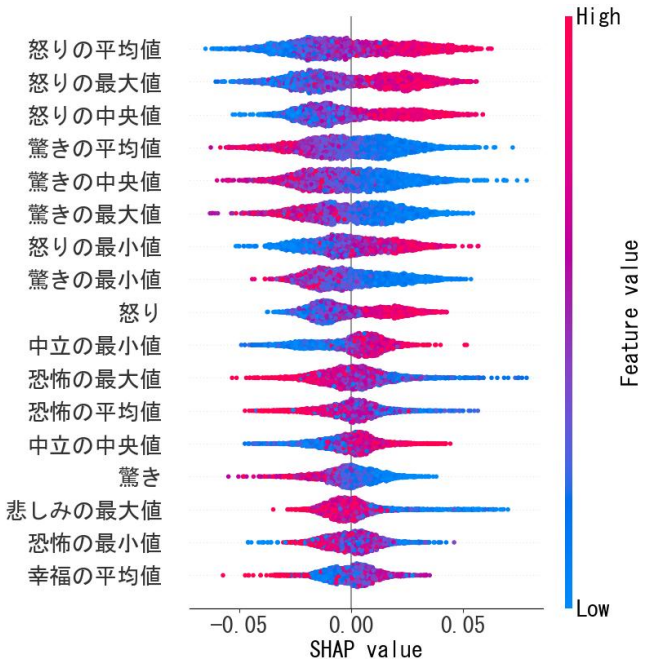


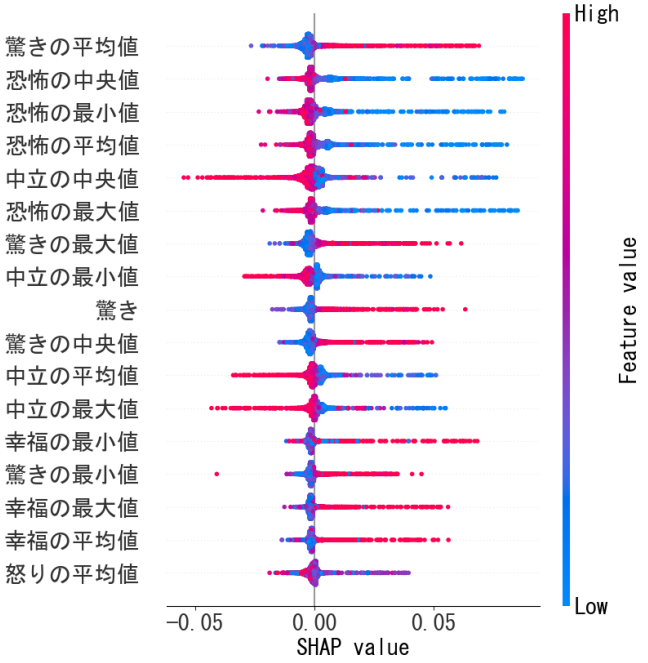
図 6　被験者 A の混同行列



(a)　困惑深度 1



(b)　困惑深度 2



(c)　困惑深度 3

図 7　被験者 A の Violin Summary Plot

図 7(a)から困惑深度 1 では「驚き」,「悲しみ」に関係する特徴量が正の方向に影響を与えることがわかった．このため,「驚き」,「悲しみ」の複合感情が少し困惑している状態を示唆すると考えられる．図 7(b)から困惑深度 2 では「怒り」,「中立」に関係する特徴量が正の方向に影響を与えることがわかった．このため,「怒り」,「中立」の複合感情が困惑状態を示唆すると

考えられる．図 7(c)から困惑深度 3 では「驚き」,「幸福」が正の方向に影響を与えることがわかった．このため,「驚き」,「幸福」の複合感情がとても困惑状態を示唆すると考えられる．

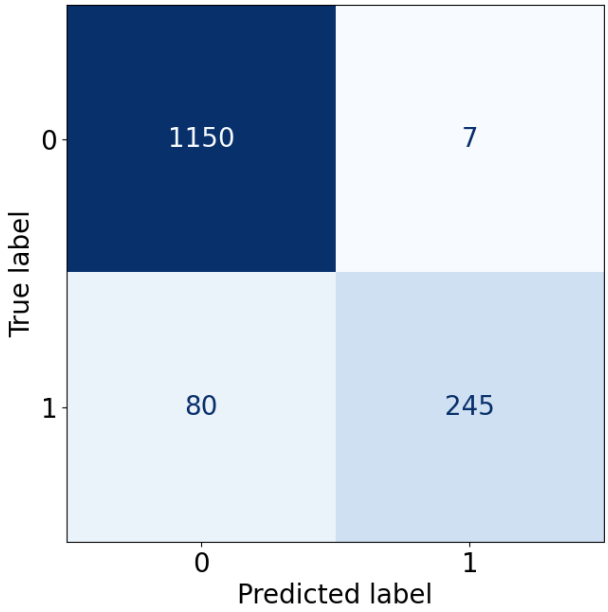


図 8　被験者 B の混同行列

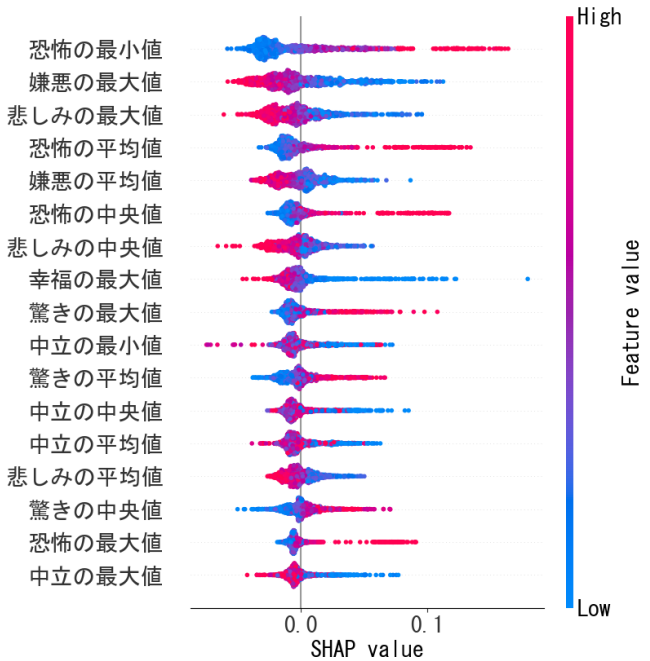


図 9　被験者 B の困惑深度 1 の Violin Summary Plot

図 9 から困惑深度 1 では「恐怖」,「驚き」に関係する特徴量が正の方向に影響を与えることがわかった．このため,「恐怖」,「驚き」の複合感情が少し困惑している状態を示唆すると考えられる．また，他の困惑深度は実験中に検出できなかった．

混同行列から求められる各困惑深度における被験者ごとの F 値を表 4 示す．

表 4　各困惑深度における被験者ごとの F 値

困惑深度	全被験者	被験者 A	被験者 B
0	0.899	0.867	0.964
1	0.854	0.701	0.849
2	0.921	0.876	
3	0.750	0.848	

5. 考察

5.1. 全体モデルと個人モデルの精度比較

表 4 より，全体モデルでは平均 0.856 の F 値となった．これは，困惑推定モデルにおいて高い精度となった．各困惑深度において，困惑深度 3 データ数が他の困惑深度と比べて少ないため学習が十分ではなく，F 値の低下がみられたと考えられる．

5.2. 自己による支援

図 5，7，9 から困惑深度 1 に着目すると，困惑に影響を与える特徴量に差異があることがわかった．全体モデルでは「嫌悪」と「悲しみ」の寄与度が高い．一方，個人モデルにおいて被験者 A では「驚き」と「悲しみ」，被験者 B では「恐怖」と「驚き」の寄与度が高い．被験者ごとに感情が異なるため，個人モデルの SHAP 値を確認することで困惑の把握をすることが可能である．

5.3. 有能な他者による支援

図 5 において，困惑深度 3 における「幸福」が正の方向に影響を与えていることがわかった．Russell[20]の感情の円環モデルにおいて「幸福」と「困惑」は，感情価の軸において対照的な位置に存在するとされている．これは，他者とのコミュニケーションによる安堵や笑顔から「幸福」が多く見られた．一方で，表 4 および図 7(c)より，被験者 A の困惑深度 3 が全体より 0.1 スコアが上昇した理由は困惑深度 3 において「驚き」という固有の感情が現れたためである．

5.4. 時間的変化の有効性

SHAP の図において重要な特徴量を見るとスライドウィンドウで統計的特徴量増加した特徴量が上位を占めていることがわかった．これは関連研究[10]が示す通り，困惑検出する際，一瞬の画像から得られる単なる分類モデルでは分類できず，機械学習を適用する際は時間的変化を特徴量の中に持たせ，分類することが必要であることが明らかになった．

5.5. 非対面学習における有効性

図 5，7，9 において，多くの特徴量において正の方

向または負の方向に赤い点が集中していることがわかった．ここから基本感情は表情表出が明確であると示唆している．顔の表情という日常的な情報から内面を可視化することで教育現場に心理的な安全性を保ちながら，データに基づいた教員が考案した個別指導を導入できる道筋を立てることができる．教育者に対して学習者の内面世界への深い洞察を与え，非対面学習における作業停滞の早期発見することが可能となる．

6. おわりに

本研究では，困惑は複数の基本感情の要素を含んでいることに着目し，基本感情の複合感情からプログラミング作業時の困惑を推定する手法を提案した．結果から，構築した困惑推定モデルは 0.8882 の F 値となった．また，SHAP を用いた分析から被験者毎に困惑推定が可能であること，時間的変化が重要であること，基本感情は表情から視認可能であることが明らかになった．しかし，個人モデルにおいて，データ数が極めて少ない困惑深度が存在し，不均衡データによる学習の偏りが生じていることが明らかになった．

今後は，学習データの質を担保するための実施課題の再選定および，被験者数の拡充によるデータセット全体の拡充を図る．また，スマートリング等の生体センシング機器を導入し，継続的かつ詳細なデータ収集を実現することも検討していく．

参 考 文 献

[1] 総務省,“令和 6 年 通信利用動向調査書(企業編)”, https://www.soumu.go.jp/johotsusintokei/statistics/pdf/HR202400_002.pdf, 2024 （参照 2026-01-05）.

[2] 日本私立大学連盟，“私立大学における遠隔授業等の実施状況等に関する調査結果について”, https://www.shidairen.or.jp/topics_details/id=4151.pdf, 2021 （参照 2026-01-05）.

[3] ソニー損保，“Web 会議と安心に関する調査”, https://prtimes.jp/a/?c=53388&r=8&f=d53388-8-1b2973afd75fa874698521938823dc45.pdf, 2021 （参照 2026-01-05）.

[4] 国立情報学研究所，“遠隔授業に関するアンケート調査の概要”, https://www.nii.ac.jp/event/upload/20200914_Report.pdf, 2020 （参照 2026-01-05）.

[5] 布施泉，“初学者を主対象とする大学の一般プログラミング教育のオンライン授業による実施”, 高等教育ジャーナル 高等教育と生涯学習, Vol. 28, pp. 65-72, 2021.

[6] P. Rozin and A. B. Cohen, “High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans”, Emotion, Vol. 3, pp. 68-75, 2003.

[7] S. D'Mello and A. Graesser, “Dynamics of affective states during complex learning”, Learning and Instruction, Vol. 22, No. 2, pp. 145-157, 2012.

[8] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo and A. C. Graesser, “Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments”, International Journal of Human-Computer Studies, Vol. 68, No. 4, pp. 223-241, 2010.

[9] F. I. Yasser, B. H. Abd and S. M. Abbas, “Detection of confusion behavior using a facial expression based on different classification algorithms”, Engineering and Technology Journal, 2021.

[10] M. Pachman, et al., “Eye tracking and early detection of confusion in digital learning environments: Proof of concept”, Australasian Journal of Educational Technology, Vol. 32, No. 6, 2016.

[11] M. S. Hussain, R. A. Calvo and F. Chen, “Automatic Cognitive Load Detection from Face, Physiology, Task Performance and Fusion during Affective Interference”, Interacting with Computers, Vol. 26, pp. 256-268, 2014.

[12] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions”, Proc. of NIPS 2017, pp. 4765-4774, 2017.

[13] A. Vaswani, et al., “Attention is all you need”, Proc. of NIPS 2017, pp. 5998-6008, 2017.

[14] L. Cai, M. M. Msafiri and D. Kangwa, “Exploring the impact of integrating AI tools in higher education using the Zone of Proximal Development”, Education and Information Technologies, Vol. 30, No. 6, pp. 7191-7264, 2025.

[15] L. Breiman, “Random forests”, Machine Learning, Vol. 45, No. 1, pp. 5-32, 2001.

[16] G. Ke, et al., “Lightgbm: A highly efficient gradient boosting decision tree”, Proc. of NIPS 2017, pp. 3147-3157, 2017.

[17] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial”, Frontiers in Neurorobotics, Vol. 7, p. 21, 2013.

[18] S. Suthaharan, “Support Vector Machine”, Machine Learning Models and Algorithms for Big Data Classification, Springer, pp. 207-235, 2016.

[19] R. E. Schapire, “Explaining Adaboost”, Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, Springer, pp. 37-52, 2013.

[20] Russell, J. A.: A circumplex model of affect, Journal of Personality and Social Psychology, 39(6), pp.1161-1178 (1980).

スマートリングによる指先生体情報を用いたプログラミング学習者の困惑状態検知

一色 夢香† 石川 昂樹† 寺田 憲司† 遠藤 雅樹† 大野 成義†

†職業能力開発総合大学校 電子情報専攻 情報通信ネットワークユニット 〒187-0035 東京都小平市小川西町 2-32-1

E-mail: †{b22305, b22302, k-terada, endou, ohno}@uitec.ac.jp

あらまし

近年，教育現場において個別支援の不足による学習意欲の低下が課題となっている．特に，プログラミング学習のような自律的かつ試行錯誤を要する分野では，学習者が過度な困惑に陥り学習を停滞させる前に，適切なタイミングで支援を行うことが極めて重要である．そこで本研究では，指の精密な生体センシングによる困惑状態の推定に着目し，スマートリングを用いた人の困惑深度推定を目的とする．提案するスマートリングは，人の動作を阻害せず，実作業中に違和感なく常時運用できる点が大きな利点である．指先における生体情報を活用した困惑推定手法は未だ確立されていないものの，装着部位である指先は，手首と比較して光電容積脈波法による計測におけるノイズの影響を受けにくいデータ取得が可能である．さらに指の動きの加速度も抽出する．本実験では，被験者にプログラミング課題を課しスマートリングによる計測を実施した．取得データから統計的特徴量を生成してデータセットを構築し，アンケートによる主観的な困惑度を正解ラベルとして分類検証を行った．その結果，被験者ごとの学習において困惑状態を推定できることが示唆された．

キーワード 感情，機械学習，生体センシング，感情分類，学習支援

1. はじめに

近年，学びの場において学習意欲の低下が深刻な課題として指摘されている[1]．特に個別支援が行き届いていない状況が散見され，こうした支援不足により，学習者が過度の困惑状態に陥り続けると，思考停止に陥り学習意欲が低下する恐れがあることが指摘されている．この問題に対処するため，困惑状態を適切に把握することは学習意欲の向上に極めて有効であると考えられている[2]．実際，困惑の解消は学習意欲を向上させる効果があり，また短期間の困惑状態は，意欲的に問題解決に取り組むための好機となり得ることも報告されている[3]．このような背景から，学習過程における個人の困惑状態を正確に推定し，適切なタイミングで支援を提供できる手法が求められている．プログラミング学習者が言語化できない「静かなつまずき」を生体情報等から可視化し，適切なタイミングで共感的なフィードバックを提供することは，試行錯誤のプロセスを前向きな体験へと変容させ，自律的な深い学びへと導くために極めて有効である．そこで本研究では，多種の生体センサが困惑推定に有効である[4]ことからスマートリングから取得可能なデータを用いて，プログラミング作業中の学習者の困惑度を推定することを目的とする．本研究においてスマートリングを測定デバイスとして選定した理由は，実験環境下における検証に留まらず，本研究の将来的な実用化及び社会実装において高い適合性を有しているためである．

2. 関連研究

Peter ら[5]は，大学生のコホートが長期間着用したスマートリングからの生理学的データを分析し，学業上の試験，就職活動のプレッシャーに関連する「周期的ストレス」を特定した．研究の結果，起床時心拍数と最大起床時心拍数の偏差を見ることで，ストレス期間を特定できることが分かった．これは，困惑推定が無菌的な実験室条件を必要とせず，日々の生理学的パターンの逸脱から推論可能であることを示唆している．

Visuri ら[6]はスマートリングとスマートフォンを用い，睡眠の質が日中の認知能力に与える影響を 2 ヶ月間にわたり調査した．覚醒度を測る能動的なテストである精神運動警戒課題に加え，スマートフォンのキーボード入力を「受動的な認知指標」として定義し，分析した．結果として，単なる睡眠時間だけでなく，睡眠潜時，夜間の心拍といった特定の質的指標を分析することで，日常的なタイピング動作から認知機能の変動を推定できる可能性を示し，スマートリングから得られるデータセットによって認知認識型システムの実現性を裏付けた．

南部ら[7]は感情ラベル付きデータがない新規ユーザに対し，日常行動の生体信号から個人特性を抽出し，パーソナライズされた感情認識モデルを構築するメタ学習手法を提案した．Leave-One-Out Cross Validation を適用することで，追加のラベル収集コストなしに，従来の教師あり学習を上回る精度を達成した．ただし，提案手法は従来手法を上回ったものの，4 クラス分類

の平均正解率は 41.5%程度であり，実用レベルとしては認識精度のさらなる向上が求められている．

工藤ら[8]は，接客現場での表情モニタリングに向け，プライバシーの課題があるカメラを使用せず，心拍・GSR・体温・筋電の 4 種の生体情報を取得するウェアラブルデバイスを開発した．ニューラルネットワークを用いた評価では，複数センサの併用が単独使用より有効であることが示され，試作機において目標に近い 65.69%の感情推定精度を達成した．ただし，デバイスに課題が残る．試作デバイスではセンサと処理部をつなぐ配線がノイズの原因となっている可能性を指摘している．精度の高い信号を取得するためには，配線を短くする工夫や，センサと信号処理部を物理的により近づけるようなデバイス設計の改良が必要であるとしている．また，現在の試作機での精度は 65.69%であり，接客現場で必要とされる目標精度 70%には到達していない．

以上，スマートリングは個人の困惑推定を実現するデバイスとして十分に期待できるデバイスである．

3. 提案手法

3.1. 概要

本研究では，プログラミング作業中の学習者の困惑状態を推定するため，スマートリング（SOXAI RING 1.1[9]）を用いた機械学習による推定システムを提案する．本システムは，学習者の指に装着したデバイスから生体情報及び挙動データを取得し，機械学習モデルを用いて困惑深度を推定するものである．システムの全体像を図 1 に示す．処理の流れは以下の通りである．

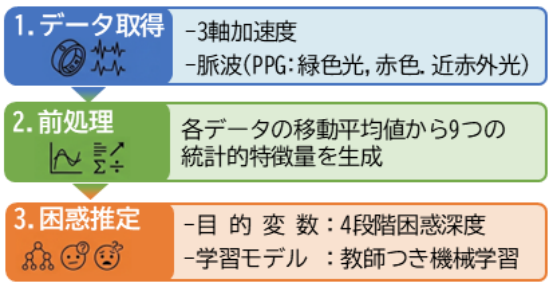


図 1　システム全体像

3.2. スマートリング

3.2.1. 機能

スマートリングは指に装着するデバイスである．指輪と同じく径別にデバイスが揃っており，密着性の高い装着が可能となる．本研究で利用するスマートリングは 3 軸方向の加速度及び，Photo-plethysmography（PPG）による脈波測定が可能である．

手指の動態を計測するグローブ型センサ等は，その物理的形状から教育現場やリモートワーク環境における常時装着が困難であり，被験者の自然な活動を阻害する心理的・身体的負荷が課題である．これに対し，指輪型デバイスのスマートリングは装着負荷が極めて低く，日常生活における連続計測に適している．

図 2　スマートリング

3.2.2. Photo-plethysmography(PPG)

PPG は皮膚表面から脈拍計測を可能とする非侵襲の生理学的計測光学技術[10]である．血液中の酸素化ヘモグロビンは 550nm 付近の光に対して強い吸収特性を持つ．測定では主に LED を利用し，緑色光 537nm，赤色光 660nm 及び近赤外光 880nm を利用する．

緑色光は特に含まれるヘモグロビンが緑色の光を吸収する特性に着目し，血管の収縮と拡張に伴う皮膚表面の反射光の微小な輝度変化を捉えることで，脈波信号の抽出が可能である．赤色光と近赤外光は両光を比較してヘモグロビンによる吸収のバランスが良く，脈動に伴う容積変化を感度良く計測する．

一方で PPG センサはモーションアーチファクトと呼ばれる体動による乱れや歪みの影響が発生するため，装着時には体動の影響が出にくい箇所に設置する必要がある．よって，スマートリングは手首にベルト装着する時計タイプに比べ密着性が高い．

3.3. 特徴量

本システムでは，スマートリングから取得した脈波と加速度に対し，ウィンドウ処理を用いた統計的特徴量の算出を行い，機械学習モデルの入力とする．

時刻 t における取得データセットを M_t と定義する．M_t は以下の 6 種類のローデータから構成される(式 1)．

$$M_t = \{axis\_x, axis\_y, axis\_z, led\_G, led\_R, led\_IR\} \text{(式 1)}$$

ここで，各変数の定義は以下の通りである．

$axis\_x_t$:時間 t における加速度 X 軸[m²/s]
$axis\_y_t$:時間 t における加速度 Y 軸[m²/s]
$axis\_z_t$:時間 t における加速度 Z 軸[m²/s]
led_G_t:時間 t における緑色光センサの電流値[A]
led_R_t:時間 t における赤色光センサの電流値[A]
led_IR_t:時間 t における近赤外光センサの電流値[A]

続いて，瞬時値だけでは捉えきれない挙動特性を反映させるため，ウィンドウサイズ w ごとのデータセット Es を構築する(式 2).

$$E_s = \{M_t, M_{t-1}, \cdots, M_{t-(w-1)}\} \quad (式\ 2)$$

データセット Es に対し，各センサ値の時系列的な変動特性を表す統計的特徴量セット F_d を算出する(式 3).

$$F_d = \{F_{ma}, F_{mi}, F_{me}, F_{mo}, F_{md}, F_{va}, F_{st}, F_{ku}, F_{sk}\} (式\ 3)$$

算出する統計量は以下の通りである.

F_{ma}：Es の最大値
F_{mi}：Es の最小値
F_{me}：Es の平均値
F_{mo}：Es の最頻値
F_{md}：Es の中央値
F_{va}：Es の分散
Fst：Es の標準偏差
F_{ku}：Es の尖度
F_{sk}：Es の歪度

これらの統計量を各センサ軸（加速度 3 軸，光学 3 種）ごとに算出し，最終的な特徴量として構築する．これを機械学習モデルに入力することで，分類結果を得る（式 4).

$$P_e = M_{RF}(M_t, F_d) \ (式\ 4)$$

P_e：分類結果
M_{RF}：教師付き機械学習

3.4. 教師付き機械学習

本研究における教師付き機械学習は Random Forest(RF)[11]を適用する．RF は，決定木を基底学習器として構築するアンサンブル学習アルゴリズムであり，異種混合的なデータセットに対しても高い汎化性能を発揮する．本手法の特筆すべき利点の一つは，単位系や数値尺度が大きく異なる複数のセンサ値を，事前の正規化や標準化処理を施すことなく直接入力として扱える点にある．これは，決定木のノード分割が各特徴量の単調増加関数に対して不変であるという性質に起因しており，異なる物理量間の絶対的な数値差が予測精度を阻害しないため，複雑なマルチセンサーデータの統合解析において有効な特性となっている．

また，各センサ値が不純度の減少に寄与した度合いを定量化する「特徴量の重要度」を算出することで，多次元的な情報の中でどの変数が重要であるかを客観的に評価することが可能である．したがって，ノイズを含む多種多様な時系列センサデータが混在する実環境の解析においても，本手法は高い解釈性とロバストネスを同時に提供する強力な解析枠組みといえる．

3.5. 評価方法

機械学習のモデル評価指標は F 値（F-score）を適用する．F 値は混同行列から求めることができる「適合率（Precision)」と「再現率（Recall)」のバランスを評価するための重要な指標である．特にデータセットのクラスに偏りがある場合や，誤検出と見逃しの両方を考慮したい場合に利用する．混同行列とは機械学習の分類において，学習モデルの分類結果と実際の正解値を対照させた表のことである．モデルが「どのクラスをどの程度正しく予測できたか」，あるいは「どのクラスと間違えやすいのか」を評価するための指標である．例えば二値分類（Positive / Negative）の場合，行に実際の正解クラス，列に予測クラスを配置する(表 1).

表 1　混同行列

		予測	
		Positive	Negative
実際	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

混同行列を構成する 4 つの要素は,以下の通りである.

- TP (True Positive): 実際も正で，予測も正
- TN (True Negative): 実際も負で，予測も負
- FP (False Positive): 実際は負だが，予測は正
- FN (False Negative): 実際は正だが，予測は負

これら 4 つの要素を利用した Precision はモデルが「正（Positive)」と予測したもののうち，実際に正であったものの割合である．すなわち,「予測の正確さ」に焦点を当てた指標である(式 5).

$$Precision = \frac{TP}{TP+FP} \quad (式\ 5)$$

Recall は，実際に「正」であるもののうち，モデルが正しく正と予測できたものの割合である．すなわち，「見逃しの少なさ」に焦点を当てた指標である(式 6).

$$Recall = \frac{TP}{TP+FN} \quad (式\ 6)$$

F-score は Precision と *Recall* により求めることができる．適合率と再現率はトレードオフの関係にある．これら 2 つの指標を統合し，一つの値でモデルの性能を評価する指標が F 値である(式 7).

$$F\text{-}score = \frac{2\times Precision\times Recall}{Precision+Recall} \ (式\ 7)$$

4. 実験

4.1. データ収集方法

本実験では，スマートリングによりプログラミング作業中における学習者の 3 軸方向の加速度及び，PPG による脈波を収集した．18 歳から 23 歳の大学生 4 名が参加した．C 言語によるプログラミング課題を実施した．課題内容は事前に履修したことがあるものに限定した．スマートリングは被験者の両手人差し指に装着し，「3.3. 特徴量」に示すローデータを取得した．ローデータの総数は 442,025 となった．

本研究は，職業能力開発総合大学校倫理審査委員会の承認を受けた．

4.2. 実験環境

実験は，学習者が普段通りにプログラミング作業が行えるデスク環境にて課題，コードエディタ，配布資料を提示し，実施した．被験者は PC に向かい，キーボード及びマウス操作によって課題に取り組んだ．実験中の様子を図 3 に示す．被験者の両手人差し指にはスマートリングが装着されており，作業の邪魔にならない状態で加速度及び脈波の常時測定が行われた．

データ収集には，SOXAI 社から提供された Android 端末向け専用アプリケーション「SET_RAW」を使用した．本アプリは，スマートリングと Bluetooth Low Energy(BLE)を介して接続される．

図 3　実験環境

式 4 に示す分類結果 P_e は困惑深度アンケートにより取得する．本研究における困惑は教育心理学における「発達の最近接領域(ZPD)」の範囲[12]を採用した．ZPD は学習支援の最適化において，学習者が独力では解決に至らないものの，教育的介入によって達成可能となる心理領域を指す．ZPD を通過する学習過程は「自動化」,「自己による支援」,「有能な他者による支援」の段階があるとされている．そこで，本研究では困惑深度アンケートの評価基準を表 2 のようにまとめた．本実験では，この主観評価に基づいてデータへのラベル付けを行った．なお，ボタン「0」は課題の開始・終了のトリガー及び「困惑が終わった」状態を示すものとして使用した．

表 2　困惑深度の評価基準

困惑深度	評価基準
0	困惑がおわった
1	手が止まり，思い出そうとする状態
2	配布資料や教材で調べる状態
3	他者に質問する状態

4.3. 実験手順

実験は以下の手順で実施した．

1. 両手人差し指にスマートリングを装着
2. データ収集用のスマートフォンアプリケーションにてデータ取得開始ボタンを押し，リングからのセンサデータ受信を開始
3. 被験者は PC 上に表示された困惑深度アンケート画面の「困惑深度 0」ボタンを押し，2 問の C 言語のプログラミング課題を開始
4. 被験者はプログラミング作業中，自身の状態に合わせてアンケート画面のボタンを押下し，困惑状態を記録
5. 課題が終了した時点で被験者は再度 PC 上の困惑深度アンケート画面の「困惑深度 0」ボタンを押し，実験を終了

4.4. 結果

「3.3. 特徴量」に示す特徴量を説明変数とし，表 2 に示す困惑深度を記録したアンケート結果を目的変数とする教師付き機械学習による困惑推定モデルを構築した．機械学習は RF で被験者全員の両手，左手，右手の分類評価を表 3 から表 5 に，混同行列を図 4 から図 6 に，特徴量重要度を図 7 から図 9 に示す．

表 3　被験者全員の両手の分類評価

困惑深度	適合率	再現率	F 値
0	0.95	0.99	0.97
1	0.89	0.99	0.94
2	1.00	0.96	0.98
3	0.87	1.00	0.93
平均値	0.92	0.98	0.95

表 4　被験者全員の右手の分類評価

困惑深度	適合率	再現率	F 値
0	0.84	0.84	0.84
1	0.53	0.95	0.68
2	0.98	0.81	0.89
3	0.97	1.00	0.98
平均値	0.78	0.87	0.80

表 5　被験者全員の左手の分類評価

困惑深度	適合率	再現率	F 値
0	0.92	0.90	0.91
1	0.84	0.98	0.90
2	0.99	0.93	0.96
3	0.65	1.00	0.79
平均値	0.92	0.94	0.92

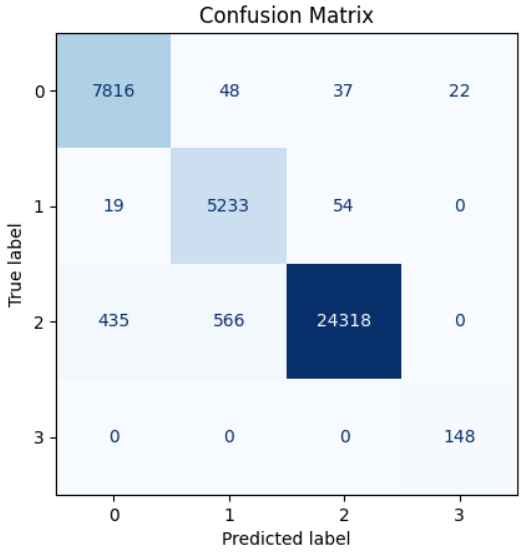


図 4　被験者全員の両手による困惑推定の混同行列

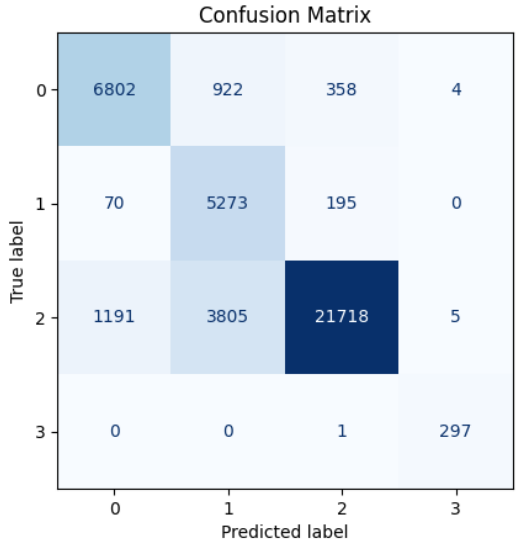


図 5　被験者全員の右手による困惑推定の混同行列

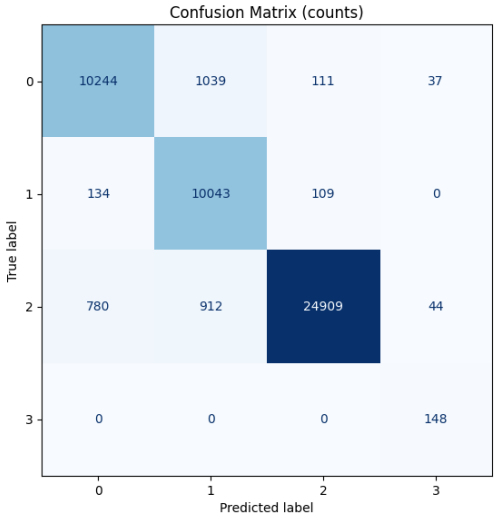


図 6　被験者全員の左手による困惑推定の混同行列

No	特徴量	重要度
1	左の近赤外光	0.054371
2	左の近赤外光F_{mo}	0.050514
3	左の近赤外光F_{me}	0.047232
4	左の近赤外光F_{ma}	0.046353
5	左の近赤外光F_{md}	0.040396
6	左の近赤外光F_{mi}	0.038748
7	左の赤色光F_{me}	0.033214
8	左の赤色光F_{mi}	0.03173
9	左の緑色光F_{mi}	0.031525
10	左の赤色光F_{md}	0.030337
11	右の近赤外光F_{ma}	0.030187
12	左の赤色光F_{ma}	0.030124
13	右の近赤外光F_{md}	0.030017
14	左の緑色光F_{ma}	0.028868
15	左の赤色光F_{mo}	0.0285
16	左の赤色光	0.027677
17	左の緑色光F_{me}	0.026673
18	左の緑色光	0.026574
19	右の緑色光F_{md}	0.026078
20	右の緑色光F_{mo}	0.026027
21	右の近赤外光F_{me}	0.025304
22	右の近赤外光	0.019668
23	右の近赤外光F_{mo}	0.019135
24	右の近赤外光F_{mi}	0.018857
25	右の赤色光	0.015818
26	右の赤色光F_{ma}	0.015548
27	右の赤色光F_{mi}	0.012732
28	右の緑色光F_{ma}	0.011516
29	右の緑色光F_{mi}	0.011038
30	右の緑色光F_{mo}	0.010686
31	右の赤色光F_{me}	0.01067
32	右の緑色光	0.010011
33	右の緑色光F_{md}	0.009767
34	右の赤色光F_{mo}	0.008927
35	右の赤色光F_{md}	0.008905
36	右の緑色光F_{me}	0.008266
37	左の加速度Y軸F_{md}	0.006756

図 7　被験者全員の両手による困惑推定の特徴量重要度

No	特徴量	重要度
1	近赤外光F_{mi}	0.084062
2	近赤外光F_{ma}	0.075843
3	近赤外光F_{mo}	0.072018
4	近赤外光	0.069049
5	近赤外光F_{md}	0.055123
6	近赤外光F_{me}	0.05084
7	赤色光F_{mo}	0.041553
8	赤色光	0.040982
9	赤色光F_{mi}	0.039904
10	赤色光F_{me}	0.037683
11	赤色光F_{md}	0.034586
12	赤色光F_{ma}	0.034258
13	緑色光F_{md}	0.031771
14	緑色光F_{mi}	0.030349
15	緑色光F_{me}	0.029167
16	緑色光	0.027938
17	緑色光F_{ma}	0.027336
18	緑色光F_{mo}	0.026078
19	加速度X軸F_{mi}	0.018367
20	加速度Y軸F_{me}	0.01749
21	加速度X軸F_{me}	0.016915
22	加速度X軸F_{mo}	0.01595
23	加速度X軸F_{md}	0.015387
24	加速度Y軸F_{md}	0.012301
25	加速度Y軸F_{mo}	0.012113
26	加速度Z軸F_{ma}	0.01072
27	加速度Y軸F_{mi}	0.009707
28	加速度X軸F_{ma}	0.009646
29	加速度Z軸F_{md}	0.007914
30	加速度X軸	0.006987
31	加速度Y軸	0.005733
32	加速度Y軸F_{ma}	0.005567
33	加速度Z軸F_{me}	0.005262
34	加速度Z軸F_{mo}	0.004423
35	加速度Z軸F_{mi}	0.00348
36	加速度Z軸	0.002231
37	加速度X軸F_{st}	0.001788

図 8　被験者全員の右手による困惑推定の特徴量重要度

No	特徴量	重要度
1	近赤外光F_{ma}	0.080536
2	近赤外光F_{md}	0.07546
3	近赤外光F_{mo}	0.064404
4	近赤外光	0.058334
5	近赤外光F_{mi}	0.057385
6	近赤外光F_{me}	0.05601
7	赤色光F_{me}	0.050272
8	赤色光F_{mo}	0.048556
9	赤色光	0.046071
10	赤色光F_{ma}	0.044977
11	赤色光F_{md}	0.044847
12	緑色光F_{mi}	0.043991
13	赤色光F_{mi}	0.043883
14	緑色光F_{md}	0.04352
15	緑色光F_{mo}	0.038721
16	緑色光F_{me}	0.038317
17	緑色光	0.03635
18	緑色光F_{ma}	0.035551
19	加速度Y軸F_{ma}	0.010343
20	加速度Y軸F_{md}	0.008073
21	加速度Y軸F_{me}	0.007265
22	加速度Y軸F_{mo}	0.006491
23	加速度Y軸F_{mi}	0.005962
24	加速度X軸F_{md}	0.005474
25	加速度X軸F_{me}	0.005469
26	加速度Y軸	0.005128
27	加速度X軸F_{ma}	0.004731
28	加速度X軸F_{mi}	0.004571
29	加速度X軸F_{mo}	0.004448
30	加速度X軸	0.003658
31	加速度Z軸F_{ma}	0.002961
32	加速度Z軸F_{me}	0.002821
33	加速度Z軸F_{md}	0.002566
34	加速度Z軸F_{mo}	0.002181
35	加速度Z軸F_{mi}	0.002142
36	加速度X軸F_{va}	0.001622
37	加速度Z軸F_{st}	0.001193

図 9　被験者全員の左手による困惑推定の特徴量重要度

5. 考察

本研究では，スマートリングから取得される生体情報に基づいたプログラミング学習者の困惑深度推定モデルを構築し，その有効性を検証した．実験の結果，表 3 及び図 4 より各困惑深度の平均 F 値は 0.95 という高い推定精度を記録した．特筆すべきは，従来の困惑の有無に留まらず，0 から 3 までの 4 段階におよぶ「困惑深度」の識別において高精度を達成した点である．この多段階評価の実現は，学習者の「つまずき」の度合いを客観的に定量化することを可能にし，指導者が介入の緊急性や優先度を判断するための科学的根拠を提供する点で，極めて高い学術的・実用的意義を有する．

デバイス装着部位の左右比較分析の結果，表 4，表 5 図 5 及び図 6 より右手に比べて左手の分類精度が一貫して高い，あるいはクラスごとの精度のバランスが良い傾向が確認された．プログラミング作業等の PC 操作において，右手はマウス操作やテンキー入力などで頻繁に使用されるため，微細な筋電位や，把持動作に伴う血流変化がノイズとして機能している可能性が示唆される．一方で，左手はホームポジションから大きく動く頻度が比較的少なく，ノイズの影響が低減されたことで，心理的な困惑状態由来の生体信号をより純粋に捉えられていると考えられる．この発見から，マウス操作を伴う PC 作業においては，マウス操作を行わない手へのデバイス装着が，心理状態をより正確に反映する可能性を見出した．

図 7，図 8，図 9 の結果より，本研究の独創的な点は，緑色・赤色・近赤外線の 3 波長を用いた PPG 測定を統合し，これを困惑深度識別の鍵として確立した点にある．特徴量重要度の分析により，深部の組織情報や酸素飽和度を反映する赤色・近赤外光が推定に大きく寄与していることが明らかとなった．赤色や近赤外光は生体組織の深部まで到達するため，表面的な心拍変動だけでなく，集中やストレスに伴う深層の血流動態の変化を捉えている可能性が高い．また，表層の血流を捉える緑色光の検出は，加速度値の変位よりも寄与度の高い特徴量として分析されており，脈波は指の動きよりも有意であることが示された．この多角的なデータ統合こそが，従来の単一波長センサでは到達し得なかった高精度な感情推定を実現しているポイントである．また，加速度値については，本研究における重要度は PPG に劣るものの，学習者が作業中か否かといった実作業の状態を推察する上では有益な情報であるため，引き続きマルチモーダルな要素として利用する価値がある．

本手法を用いて学習者の困惑深度を教員に可視化し提示することで，学習意欲の向上につながる質の高い教育環境の提供が可能となる．今回の結果は，オンライン作業のみならず，対面授業においても，静かな環境下で講師が各生徒の状態を把握し，適切なタイミングで個別の声掛けを行う「適応型学習支援システム」への応用が期待できる．学習者が言語化できない「つまずき」をリアルタイムで可視化できる点に，本研究の社会的意義がある．以上の結果から，本研究の成果は，日常生活に馴染むスマートリングという形態でありながら，専用の医療機器や大型装置に匹敵する精度で人間の心理状態を可視化できるという，高い社会的・技術的優位性を示していると結論付ける．

6. おわりに

本研究では，スマートリングを用いたプログラミング学習時の困惑推定手法を提案し，RF により F 値 0.95 超の高精度を達成した．特徴量分析の結果，加速度よりも脈波情報の寄与が高く，動作が静止していても内面的な困惑を検知可能であることが明らかになった．また，片手装着でも実用上十分な精度が得られることを確認した．最終的に，本技術は学習者が安心して試行錯誤を繰り返し，自律的な深い学びへと到達するための強力な支援手法となることが期待される．

7. 謝辞

本研究を進めるにあたり，スマートリングのファームウェアの変更にご協力いただいた株式会社 SOXAI 様に深く感謝の意を表する．

参 考 文 献

[1] ベネッセ教育総合研究所, “子どもの生活と学びに関する親子調査 2023”. https://issnews.iss.u-tokyo.ac.jp/3a1f33158fc1c43f511ea673eac7410f637a8c7e.pdf 最終アクセス 2025/8/29.

[2] A. Arguel, L. Lockyer, G. Kennedy, J. M. Lodge and M. Pachman, “Seeking Optimal Confusion: A Review on Epistemic Emotion Management in Interactive Digital Learning Environments”, Interactive Learning Environments, vol. 27, no. 2, pp. 200-210, 2019.

[3] M. Kapur, “Productive Failure”, Cognition and Instruction, vol. 26, no. 3, pp. 379-424, 2008.

[4] K. Kurata, A. Tsuji and K. Fujinami, “Uh-Huh Duck: Self-Problem-Solving Support During Programming Through Interaction with a Doll Duck”, Proc. of IEEE GCCE, pp. 676-677, 2024.

[5] P. Neigel, A. Vargo, B. Tag and K. Kise, “Unobtrusive Stress Detection Using Wearables: Application and Challenges in a University Setting”, Frontiers in Computer Science, vol. 7, 1575404, 2025.

[6] A. Visuri, H. Koskimäki, N. van Berkel, A. Alorwu, E. Peltonen, S. Abdullah and S. Hosio, “Cognitive Performance Measurements and the Impact of Sleep Quality Using Wearable and Mobile Sensors”, arXiv preprint arXiv:2501.15583, 2025.

[7] 南部優太，幸島匡宏，岩田具治，片岡春乃，望月

理香, 山本隆二, “日常生活時の生体信号を用いたパーソナライズされた感情認識モデルの学習”, 第 38 回人工知能学会全国大会論文集 2024.

[8] 工藤悠佑, 高松誠一, 伊藤寿浩, “接客現場における感情推定のためのマルチモーダルウェアラブルデバイスの研究”, 第 36 回エレクトロニクス実装学術講演大会 2022.

[9] SOXAI , “ SOXAI RING 1 ” , https://soxai.co.jp/products/soxai-ring-2 （2025 年 7 月 1 日アクセス）.

[10] 中嶋一喜, 江田英雄, “指尖脈波のみから計算した PTT の検証”, 生体医工学 2024.

[11] L. Breiman, “Random Forests”, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[12] L. Cai, M. M. Msafiri and D. Kangwa, “Exploring the Impact of Integrating AI Tools in Higher Education Using the Zone of Proximal Development”, Education and Information Technologies, vol. 30, no. 6, pp. 7191-7264, 2025.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique”, Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

動画に基づく教本参照型コーチングエージェントの構築

川田　拓朗†　　藤若　雅也††　　ジショウトン††　　劉　　健全††

† 法政大学大学院理工学研究科　〒 184–8584 東京都小金井市梶野町
†† NEC ビジュアルインテリジェンス研究所　〒 211–8666 神奈川県川崎市中原区下沼部
E-mail: †takuro.kawada.3g@stu.hosei.ac.jp, ††{fujiwaka, xiaotong-ji, jqliu}@nec.com

あらまし　コーチングとは, 観測された行動を参照基準と比較し, その差分に基づいて改善を促す教育的支援である. 動画に基づくコーチングの既存研究の多くは, 学習者の動画に対してお手本動画を参照基準としてきたが, 高品質なお手本動画を大規模に用意することは困難である. 本研究では, 教本などのドキュメントを参照基準とし, 学習者の行動が, いつ・どの規範から・どの程度, 基準から逸脱したかを推定するコーチングエージェントを提案する. 提案手法では, ドキュメントをルーブリックとして構造化し, 動画中の行動と整列させることで, 各時刻の逸脱度を定量化する. これにより, 参照動画や追加学習を必要としない解釈可能なコーチング支援を実現する.

キーワード　マルチモーダル検索, 動画ベースコーチング, 映像解析

1　はじめに

コーチングとは, 観測された行動を何らかの参照基準と比較し, その差分に基づいて学習者の理解や行動の改善を促す教育的支援である [2]. この考え方は, スポーツ指導や医療教育, 専門職教育など多様な分野において, 効果的なフィードバックや指導の在り方として広く議論されてきた [3], [9], [7]. 近年, 大規模言語モデル (Large Language Models; LLM) や視覚言語モデル (Vision Language Models; VLM) の進展に伴い, 人間の指導プロセスを自動化あるいは支援するコーチングモデルへの関心が高まっている. これらの研究では, 学習者の行動を入力として解析し, 適切な助言や改善点を提示することで, 人手による指導を補完または代替することが目指されている.

スポーツや技能学習の分野では, 学習者の動作を動画として入力し, 熟練者のお手本動画との比較に基づいてフィードバックを生成する参照動画ベースのコーチング手法が提案されている [1], [10]. このような参照動画に基づくアプローチでは, 学習者と熟練者の動作を時間的に対応付けた上で, 姿勢や運動の時間変化に基づく表現を用いて両者の動作差分を明示的に捉えることができるため, 改善すべき箇所を直接的に特定できるという利点がある. しかし, 多くの既存手法では技能や運動種目ごとに収集された熟練者動画を用いたモデルの学習を前提としているため, 新たな技能や指導内容に適用する際には追加のデータ収集や再学習が必要となる. 特に, 専門性の高い技能や限定的な指導領域において, その内容を十分に反映した高品質なお手本動画を大規模に収集することは容易ではなく, これら手法の適用範囲や拡張性は限られている.

一方, 実際の技能学習や指導の現場では, 書籍や教本といった文書資料が参照基準として広く用いられている. このようなドキュメントには, 技能の手順や注意点, 評価観点が文章や図解として体系的に整理されており, 技能に関する規範的知識を明示的に提供している. しかし, これらの記述は人間による理解を前提として構成されており, 実世界の学習者の動画中に含まれる連続的な行動に対して, ドキュメント内のどの記述を対応付けるべきかはあらかじめ明示されていない. また, 技能の記述方法や粒度, 図解の有無といった文書の構成やスタイルはドキュメントごとに大きく異なるため, 熟練者動画との比較のように時系列的な対応関係や動作差分を直接導出することは容易ではない. したがって, ドキュメントを参照基準として動画中の行動を評価するためには, 記述された技能知識を動画から判定可能な構造へと整理した上で, それらと動画中の行動との対応関係を体系的に推定する枠組みが求められる.

本研究では, ドキュメントを参照基準とするコーチングエージェントの構築を目的とし, ドキュメントに記述された技能知識を手順構造および評価規範からなるルーブリックに変換し,

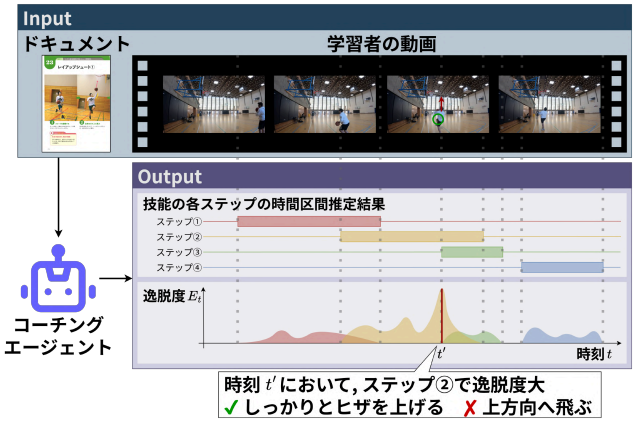


図 1: 学習者動画と教材ドキュメントを入力とし, 技能遂行を時間構造と品質を可視化する提案手法の出力例. 参照動画や追加学習を必要とせず, 入力動画における行動が「いつ・どの行動が・どの行動規範から・どの程度」ドキュメントに記述された規範を参照基準から逸脱しているかの特定を可能とする.[1]

1：ドキュメントのスクリーンショットは文献 [11] より引用.

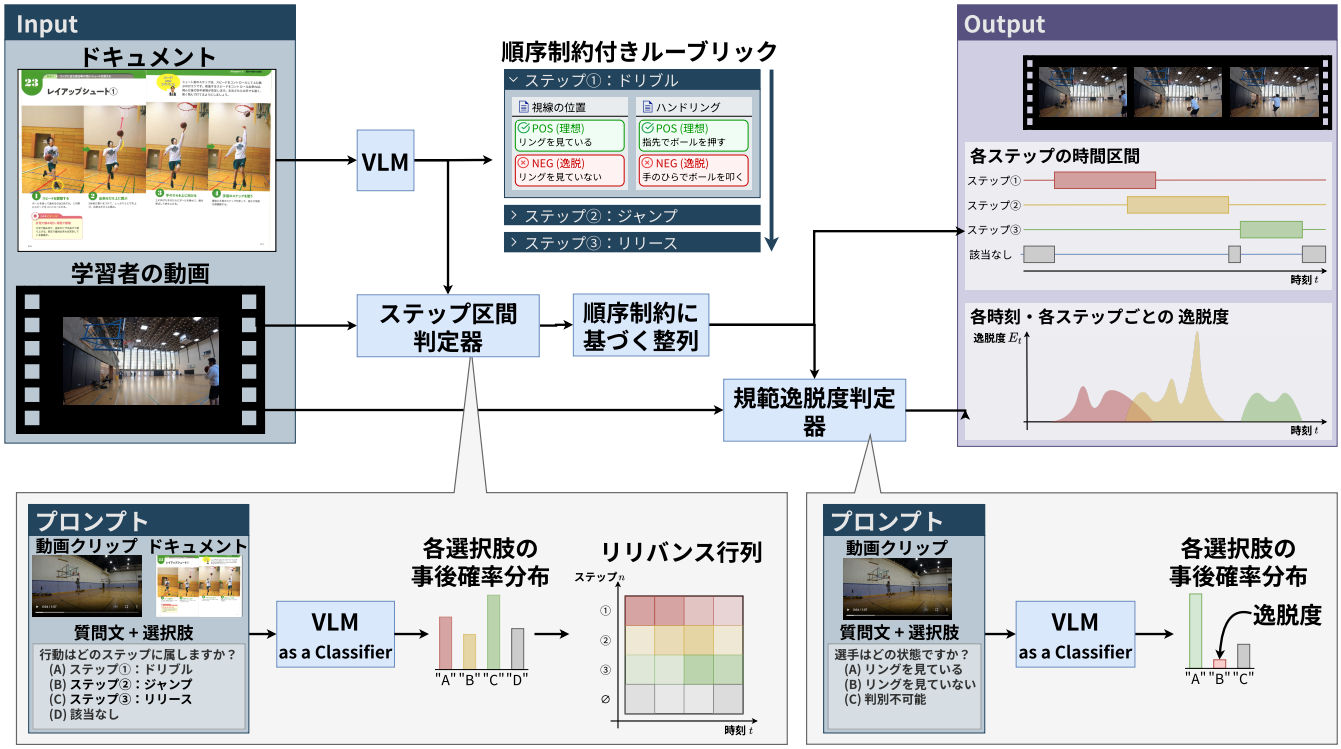


図 2: 提案するコーチングエージェントの概要図．ドキュメントと学習者の動画を入力とし，(1) 文書記述を手順ステップおよび評価規範からなる順序制約付きルーブリックへ変換し，(2) 各ステップが動画中のどの区間に対応するかを順序制約の下で推定・整列し，(3) 各ステップに定義された規範に基づいて行動の逸脱度を時系列的に算出する．ドキュメント内の知識と動画中の行動を一貫した枠組みで対応付けることで，技能遂行の時間構造と品質を同時に把握可能な解析基盤を実現する．[2]

学習者の動画中の行動がそれらの規範からどの程度逸脱しているかを定量的に推定する枠組みを提案する．我々が提案する枠組みは以下の 3 段階の処理から構成される: (1) まず，ドキュメント内の記述を解析することで当該技能の一連の動作をステップに分解し，各ステップに対応する評価規範を抽出・構造化することで動画中の行動を評価可能なルーブリックを構築する．(2) 次に，各ステップが動画内のどの区間に該当するかを順序制約を考慮しながら推定することで，動画全体に対するステップ区間を整列する．(3) 最後に，各ステップごとに，学習者がドキュメント内に記載された評価規範を満たしているか否かを判定し，その結果を参照基準からの逸脱度として定量化する．この枠組みにより，入力された動画に対して，「いつ・どの行動が・どの行動規範から・どの程度」逸脱しているのかを推定し，フィードバックとしてユーザに提供することが可能となる (図 1)．我々は収集が困難なお手本動画を用いることなく，ドキュメントに基づく学習不要なコーチングエージェントの構築した．

本研究の貢献は以下の通りである:

- ドキュメントに記述された技能知識を構造化された順序付きのルーブリックへ変換し，動画内の行動を評価するために利用可能とする枠組みを示した．
- ドキュメントと動画の間に時系列的な対応関係が明示されていない状況において，技能の手順ごとの時間区間推定を導入し，後段の行動評価と整合的な形で動画中の行動と文書知識を対応付ける方法を示した．
- ドキュメントを参照基準として用いることで，参照動画や追加学習を必要とせず，動画中の各時間区間における行動の逸脱度を定量的かつ解釈可能に推定できることを示した．

2 提 案 手 法

本研究では，ドキュメントを参照基準として用い，学習者の動画中の行動をステップ単位で評価する手法を提案する．入力として学習者の動画およびドキュメントの該当ページを与え，出力として動画中の各時間区間に対応するステップと，各ステップに定義された規範に対する逸脱度を推定する．これにより，学習者は入力した動画の各時刻において，どのような規範をどの程度逸脱したかを明示的に把握することが可能となる．提案手法は，図 2 に示すように，以下の 3 つの処理から構成される: (1) ルーブリック構築: ドキュメント内の記述をステップおよび評価規範からなるルーブリックへ変換する; (2) ステップ区間推定: 動画中の各クリップと各ステップとの対応区間を推定する; (3) 逸脱度推定: 各ステップに定義された規範に基づき，動画中の行動の逸脱度を算出する．

2.1 ルーブリック構築

我々はドキュメントに記述された技能知識を，動画中の行動を評価可能な参照基準として利用するため，ドキュメントの内容をルーブリックとして構造化する． ルーブリックとは，技能の遂行過程を構造化された手順と評価観点として明

2：ドキュメントのスクリーンショットは文献 [11] より引用．

示的に定義した評価基準を指す．本研究においては，技能の遂行過程を表す順序付きのステップ集合 $S = \{s_i \mid i \in \mathcal{I}\}$, $\mathcal{I} = \{1, 2, \ldots, N\}$ と，各ステップに対応する規範項目の集合 $R^{(i)} = \{r_j^{(i)} \mid j \in \{1, 2, \ldots, M^{(i)}\}\}$ から構成されるチェックリスト形式の評価基準としてルーブリックを定義する．ここで s_i は技能手順における第 i ステップに関するドキュメント内の記述を表し，N はステップ数である．また，$r_j^{(i)}$ は第 i ステップにおいて，ある行動がドキュメントの記述に沿っているかを判定するための評価項目を表し，$M^{(i)}$ は規範項目の数である．

ステップ抽出． ステップ S は，情報抽出器として VLM を用いてドキュメント内に記述された技能手順を解析することで構築される．この処理は，ドキュメント内の記述を変更することなく，技能の遂行過程を複数の手順段階に分割し，各段階に対応する説明文を対応付けることを目的とする．VLM への入力には，該当技能について記載された該当ページ群のスクリーンショット画像および PDF テキストを用いる．これらの入力を基に，各ステップを簡潔に表すタイトルと関連する全てのドキュメント内の文の集合を抽出する．このように構築された各ステップ s_i は，後続の動画解析において単一の時間区間に割り当て可能な技能遂行上の動作単位として構成される．

規範項目への極性付与および補完． 各ステップ s_i に対応する規範項目集合 $R^{(i)}$ は，ステップ抽出段階で得られた各ステップに関連する全ての説明文を VLM に入力し，評価可能な命題へと整形することで構築する．各規範項目 $r_j^{(i)}$ は，行動がドキュメントに記述された規範を満たしている状態を表す命題文 $r_{j,\mathrm{POS}}^{(i)}$ と，満たしていない状態を表す命題文 $r_{j,\mathrm{NEG}}^{(i)}$ のペアからなる二値的な表現として定義される．これら命題は第 i ステップを実行している学習者の行動が，どちらの状態に該当するかを動画から判定可能な形式で記述される．VLM は，ドキュメント内の記述を基に，これら 2 つの状態を区別するための命題文を生成・整形し，両者が意味的に補集合となるよう規範表現を構成する．ここで，ドキュメントの記述形態によっては，これら 2 つの状態のいずれか一方のみが明示されている場合がある．すべての規範項目を一貫して二値的に扱うため，記載されている命題を基に，対応するもう一方の状態を VLM を用いて補完する．この補完処理は，ドキュメントの記述内容を逸脱しない範囲で行われ，新たな評価基準や知識は付加されない．

以上の処理により，ドキュメントは各ステップの順序構造と，各ステップに対応する二値的な評価規範集合からなるルーブリックへと変換される．各ステップおよび規範項目は，ドキュメント内の記述に基づいて構成されており，評価の根拠を原文に対応付けて保持している．本ルーブリックは，ドキュメントに記述された技能知識を手順構造と評価観点の両面から構造化したものであり，後続のステップ区間推定および逸脱度推定における参照基準として用いられる．

2.2 ステップ区間推定

続いて，学習者動画中において，各ステップが実行されている時間区間を推定する．この問題は，クエリとなる各ステップがあらかじめ順序を持つという制約の下で，対応する動画中の時間区間を検索する Video Moment Retrieval タスクとして捉えられる．本研究では，以下の 2 つの処理でステップ区間推定を行う: (1) リリバンス行列の構築: 動画中の各時間区間と各ステップとの関連度を推定する; (2) 順序制約に基づく整列: 各ステップの順序制約に基づき，リリバンス行列を用いて動画中の区間を一貫した形で整列する．

リリバンス行列の構築． まず，学習者動画 V を固定フレーム長 τ の短いクリップ列 $V = \{c_t \mid t \in \{1, 2, \ldots, T_{\mathrm{clip}}\}\}$ に分割する．ここで c_t は動画中の時刻 t に対応するクリップ，T_{clip} はクリップ数を表す．動画の総フレーム数を T_{frame} とすると，$T_{\mathrm{clip}} = \lceil T_{\mathrm{frame}}/\tau \rceil$ となる．

次に，各動画クリップ c_t と，各ステップ s_i の関連度 $p_{t,i}$ あるいは，いずれのステップにも属さない状態との関連度 $p_{t,\emptyset}$ を推定し，これらを要素として持つリリバンス行列 $P \in \mathbb{R}^{T_{\mathrm{clip}} \times (N+1)}$ を構築する．ここで，$N+1$ 列目はいずれのステップにも属さない状態に対応する．我々は動画クリップとドキュメント中の技能記述という異なるモダリティ間の対応関係を VLM を用いて推定する．動画クリップ c_t，該当ページのスクリーンショット画像，「このクリップがどのステップを実行しているか」を問う質問文，選択肢となるステップ集合 S および，いずれのステップにも属さない状態における選択肢を VLM に入力し，$N+1$ 択の QA タスクとして推論させ，各選択肢に対応する識別子トークンを 1 文字生成する．逐次的なトークン生成に基づいて推論を行う VLM が最初に出力する各選択肢識別子トークンの対数尤度に softmax 関数を適用することで各クリップと各ステップの関連度 $p_{t,n}$ を定義する．

このように VLM を確率推定可能な分類モデルとして用いることで，ドキュメントに含まれるテキスト，図表などの複雑な視覚的情報，VLM が持つ行動理解能力を活用し，異なるモダリティ間の対応関係を連続値として柔軟に推定できる．

順序制約に基づく整列． リリバンス行列 P は各クリップごとに独立に推定されるため，局所的な誤りやノイズを含む可能性がある．また，各ステップはドキュメントに定義された順序に従って単調に進行するという構造的制約を持つ．そこで，ステップ順序の制約を明示的に考慮するため，動画全体に対するステップ区間を推定する．

まず，各時刻 t における動画中の行動に対応するステップを，ステップ番号の集合列 $\mathcal{A} = \{A_t \subseteq \mathcal{I}\}$ として表す．$i \in A_t$ であるとき，時刻 t において，第 i ステップ s_i がアクティブであることを意味する．また，$A_t = \{\emptyset\}$ であるとき，時刻 t において，いずれのステップにも該当しない区間であることを表す．ステップ番号が時間と共に単調に進行するという制約の下，$A_t \neq \{\emptyset\}$ ならば，開始ステップ $\alpha_t \in \mathcal{I}$ と同時にアクティブなステップ数 $L_t \coloneqq |A_t|$ を用いて，$A_t = \{\alpha_t, \alpha_t + 1, \ldots, \alpha_t + (L_t - 1)\}$ となる．ただし，$\alpha_t + (L_t - 1) \leqq N$ である．

次に，各時刻 t におけるスコア ϕ_t を，割り当てられたステップ集合に含まれる確率の最大値として次のように定義する:

$$\phi_t = \max_{n \in A_t} p_{t,n} \tag{1}$$

そして，全時刻におけるスコアの和が最大となる区間系列を推

定するため, 以下の最適化問題を解く:

$$\mathcal{A}^* = \underset{\mathcal{A}}{\operatorname{argmax}} \sum_{t=1}^{T} \phi_t. \tag{2}$$

ここで, ステップ境界付近における曖昧性を表現するため, 連続する時刻間で共有されるステップ数が高々 d 個となるよう, $|A_t \cap A_{t+1}| \leqq d$ という制約を課す. ここで, $d \geqq 1$ である. これにより, 各ステップの時間的独立性を考慮するとともに, 全ての時刻において $A_t = I$ となるような自明解を防ぐ. また, 各ステップは動画中で 1 つの連続区間として出現するものとし, $\max A_t \leqq \min A_{t+1}$ という制約を課す. これにより, 各ステップの進行の単調性を担保し, 一度終了したステップが再度出現するような遷移を防ぐ. 本研究では, この最適化問題を動的計画法で解き, 各ステップの時間区間を推定する. A^* は後続の逸脱度推定において参照される.

2.3 逸脱度推定

最後に, 2.1 節で構築したルーブリックと 2.2 節で推定したステップ区間に基づき, 学習者動画中の行動が各規範をどの程度逸脱しているかを定量的に評価する. 各時刻 t においてアクティブなステップ群 A_t において, 各ステップで定義された規範項目集合 $R^{(i)}$ を用いて行動評価を行い, どの規範がどの程度満たされていないかを定量化する.

我々は 2.1 節のリリバンス行列の推定と同様に, VLM を分類モデルとして用い, 各クリップが規範項目を満たしているか否かを推定する. 動画クリップ c_t, 「このクリップにおける行動はどの状態に該当するか」を問う質問文, 選択肢となる命題 $r^{(i)}_{j,\mathrm{POS}}$, $r^{(i)}_{j,\mathrm{NEG}}$ および, いずれの状態にも該当しない場合の選択肢 r_{UNK} を VLM に入力し, 3 択の QA タスクとして推論させ, 各選択肢に対応する識別子トークンを 1 文字生成する. VLM が最初に出力する各選択肢識別子トークンの対数尤度を取得し, softmax 関数を適用することで, 動画クリップ c_t が各状態に属する確率を取得する. ここで, 規範を満たしていない状態 $r^{(i)}_{j,\mathrm{NEG}}$ が選択される確率を, 時刻 t における規範項目 $r^{(i)}_j$ に対する 逸脱度 $e^{(i,j)}_t$ として定義する. この定義により, 動画の品質や撮影アングル, 遮蔽などの要因によって規範の判定が困難な場合には, r_{UNK} に対応する確率が大きくなり, 結果として逸脱度 $e^{(i,j)}_t$ は小さく抑えられる. これにより, 観測情報が不十分な状況において誤って逸脱を検出することを避け, 評価の不確実性を反映させる. 最終的に, 時刻 t における行動全体の逸脱度 E_t は, アクティブなステップ群 A_t に含まれる全ての規範項目に対する逸脱度の総和として定義する:

$$E_t = \sum_{s_i \in A_t} \sum_{r^{(i)}_j \in R^{(i)}} e^{(i,j)}_t. \tag{3}$$

このように定義された逸脱度は, 動画中の各時間区間において, どの程度ドキュメントに記述された規範から逸脱しているかを連続値として表現するものであり, 後続のコーチング支援において定量的かつ解釈可能な指標として利用可能である.

3 評価実験

3.1 実験設定

本研究では, スポーツの練習動画とそれに対応する指導書を用い, 提案手法により動画中の行動が指導書に記述された規範からどの程度逸脱しているかを推定する実験を行った. 我々は動画データとして, Ego-Exo4D データセット [4] に含まれるバスケットボール練習動画 280 件を使用した. 本データセットには, 学習者が実際にバスケットボールの基礎的な技能練習を行っている様子が三人称視点で収録されており, 撮影環境や被写体のばらつきを含む実環境に近い条件が含まれている. また, 参照基準となるドキュメントとして, バスケットボールの基礎技能を解説した指導書 [11] を用いた. 本教本は文章および図解を用いて技能の手順や注意点を説明しており, Ego-Exo4D データセットの動画が対象とするジャンプシュートおよびレイアップに関する全てのページを対象とした.

本研究において用いる VLM は, ルーブリック構築, ステップ区間推定, 逸脱度推定のすべての段階において GPT-5.2 [6] を使用した. また, 各動画は固定フレーム長 $\tau = 4$ の短いクリップ列に分割し, 提案手法の各段階においてこれらのクリップを基本単位として処理を行った. ステップ区間推定の順序制約に基づく整列において, 隣接する時刻間で共有されるステップ数を $d = 1$ とし, ステップの切り替わりに伴う短時間の曖昧さを表現しつつ, 不自然な長時間の重なりを抑制した.

ステップ区間推定におけるリリバンス行列の構築については, 提案手法である VLM に QA を解かせて対数尤度を取得する手法に加えて, 既存の Video Moment Retrieval 手法である InternVideo2 [8] および R^2-Tuning [5] を比較手法として用いた. これらの手法では, 各ステップの説明文と動画クリップとの類似度に基づいてリリバンス行列を構築し, 提案手法と同一の動的計画法による整列処理を適用した.

3.2 評価指標

ルーブリック構築の妥当性評価. 提案手法により構築されたルーブリックが, ドキュメントに記述された内容を過不足なく構造化できているかを検証するため, 人手による妥当性評価を行った. 提案手法では, ルーブリック構築を以下の 2 段階に分けて行っている: (1) ステップ抽出; (2) 規範項目への極性付与および補完. そこで, 我々はそれぞれの段階に対応した観点を設定し, 定性的にその妥当性を評価した. まずステップ抽出について, 次の 3 つの観点から妥当性を評価した: (i) 原文忠実性: 各ステップがドキュメント原文から直接導出可能な記述に基づいて構成されているか; (ii) ステップ混在の不存在: 1 つのステップに, 異なる技能段階における動作が混在していないか; (iii) 手順的曖昧性の不存在: ステップが技能遂行のどの段階に対応するかが不明確でないか. 次に, 規範項目への極性付与および補完について, 次の 3 つの観点から妥当性を評価した: (iv) 意味逸脱の不存在: 規範項目がドキュメントに記載されていない新たな評価基準や解釈を含んでいないか; (v) 極性の妥当性:

命題 $r^{(i)}_{j,\mathrm{POS}}$ および $r^{(i)}_{j,\mathrm{NEG}}$ が意味的に自然な補集合として構成されているか; (vi) 動画判定可能性: 動画を観察することで, 当該規範を満たしているか否かを判断可能な形式になっているか. 各観点に関して, 構築されたルーブリックが条件を満たすか否かを手動で二値的に判断し, 全評価対象項目のうち条件を満たした割合を指標とする.

ステップ区間推定の妥当性評価. 提案手法におけるステップ区間推定の妥当性を検証するため, 推定された各ステップの時間区間が, 人間の直感的な理解と整合した大まかな区切りを捉えているかを評価した. 我々は動画全体を観察し, 各ステップが実行されていると考えられる時間区間を手動で指定した. そして, モデルによって推定されたステップ区間と手動でアノテーションされた区間との一致度を, 各ステップごとの Intersection over Union (IoU) により評価し, 提案手法および比較手法について結果を比較した. ここで, IoU は推定区間と正解区間の重なり長を, それらの和集合長で割った値として定義される.

熟練度に基づく逸脱度の比較. 提案手法により推定される逸脱度が技能の熟練度の違いを反映しているかを検証するため, 熟練者の動画と初心者の動画をそれぞれ入力した際の逸脱度の差の比較を行った. 初心者と熟練者との分割には, コーチングモデルに関する先行研究 ExpertAF [1] における拡張データセットのラベルを用いた. 我々は, 各手法について初心者群と熟練者群の時間方向の平均逸脱度の差 Δ(初心者 − 熟練者) を比較した. この差は, 提案する逸脱度が技能レベルの違いをどれだけ明確に反映しているかを表す量であり, 値が正の方向に大きいほど初心者の動画においてより高い逸脱度が付与されていることを意味し, 熟練度の違いを適切に捉えられていることを示す.

逸脱度の定性評価. 提案手法により推定される逸脱度が, 人間の直感的な良否判断と整合した挙動を示すかを検証するため, 定性的な評価を行った. 我々は少数の動画サンプルを対象とし, 動画全体の中から明らかに上手くできている区間, および明らかに不適切である区間を手動で選択した. その上で, 時系列方向における逸脱度 E_t の推移を可視化し, 不適切と判断されたクリップが出現する区間において, 逸脱度が相対的に高くなる傾向が見られるかを定性的に評価した.

4 結果と考察

ルーブリック構築の妥当性評価. 表 1 に, 各教材ページから構築されたルーブリックに対する妥当性評価結果を示す. ステップ抽出に関する観点である (i) 原文忠実性 および (iii) 手順的曖昧性の不存在 は, すべての対象において高い達成率を示した. また, 多くの場合において (ii) ステップ混在の不存在 も高い値を示し, 構築されたルーブリックにおいて, 技能手順が概ね明確な単位として分割されていることが確認された. 規範項目への極性付与および補完に関しては, (v) 極性の妥当性 が一貫して高い値を示し, POS/NEG が意味的に自然な補集合として構成されていることが確認された. さらに, (iv) 意味逸脱の不存在 も高い達成率を示しており, 構築された規範項目が原文に含まれない新たな評価基準を過剰に導入していないことが示された. 一方で, (vi) 動画判定可能性 は他の観点と比較して相対的に低い値となった. すなわち, 構築された規範項目の一部は, 動画から直接二値判定するには曖昧さを含む形式となっていることが確認された.

表 1: ルーブリック構築の妥当性評価結果. ステップ抽出および規範項目の極性付与・補完という二段階の処理に対して, 各評価観点を満たすと判断された項目の割合を示す.

評価する処理段階	評価観点	妥当な項目の割合
ステップ抽出	原文忠実性	1.000
	ステップ混在の不存在	1.000
	手順的曖昧性の不存在	1.000
極性付与・補完	意味逸脱の不存在	0.967
	極性の妥当性	1.000
	動画判定可能性	0.833

これらの結果は, いずれも提案手法の設計方針に起因するものである. 提案手法は, 技能手順を再解釈して再編成するのではなく, 原文中に現れる節構造や記述単位を保持したまま, ドキュメントの構造をルーブリック形式へ忠実に写像することを目的として設計されている. そのため, 原文においてステップが明示的かつ時間的に整合的に分割されている場合には, 高い (i) 原文忠実性 と (iii) 手順的曖昧性の不存在 が得られる. 一方で, 原文側の分割が曖昧であったり, 複数の動作段階が一つの節に混在している場合には, その構造的曖昧さや時間的非整合性もまたそのままルーブリックに反映される. また, 原文に含まれる抽象的・感覚的表現も忠実に写像されるため, 一部の規範項目は (vi) 動画判定可能性 の観点からは不十分な形式となる. このように, 本手法は原文忠実性を最大化する設計であるがゆえに, 評価基準としての可観測性や, 手順構造を再編成する柔軟性との間にトレードオフを内包している.

以上より, 本手法はドキュメントに記載された知識を過不足なく構造化するという目的に対して有効であり, とりわけ技能手順がステップとして明示的に記述された教材に対しては妥当なルーブリックを自動的に構築できることが確認された. 一方, 原文中の手順構造が曖昧であったり, 感覚的・抽象的な表現に依存する教材に対しては, その曖昧さ自体がルーブリックに反映されるという限界も明らかとなった. このようなドキュメントに対しても適用可能な枠組みとするためには, 原文構造に依存せずに潜在的な手順境界を推定する仕組みや, 抽象的な記述を動画上で観測可能な行動表現へと変換する補正処理を組み込むことが重要である.

ステップ区間推定の妥当性評価. 表 2 に, ステップ区間推定の妥当性評価結果を示す. 提案手法は 0.415 ± 0.230 の IoU を達成し, InternVideo2（0.199 ± 0.157）および R^2-Tuning（0.230 ± 0.095）を大きく上回った. この結果は, 提案手法が各ステップの位置と広がりを, 人手で指定した区間と整合する形で安定して推定できていることを示す. 従来の Video Moment Retrieval 手法では, 各ステップの説明文を固定的なテキスト表

表 2: ステップ区間推定手法ごとの妥当性評価および熟練度に基づく逸脱度の比較結果. 各手法が利用する入力情報 (説明テキスト／スクリーンショット) とともに, ステップ区間推定精度を IoU により評価する. 逸脱度は, 初心者群および熟練者群における平均値と標準偏差, 両者の差 Δ（初心者 − 熟練者）を報告する. Δ が正で大きいほど, 初心者に対して一貫して高い逸脱度が付与されており, 推定結果が熟練度の違いをより明確に反映していることを示す.

ステップ区間推定手法	入力情報		ステップ区間 IoU	逸脱度の平均		
	説明テキスト	スクリーンショット		初心者	熟練者	Δ(初心者 − 熟練者)
InternVideo2 [8]	✓	✗	0.199 ± 0.157	2.748 ± 1.007	2.704 ± 1.019	0.044
R^2-Tuning [5]	✓	✗	0.230 ± 0.095	6.590 ± 1.378	6.400 ± 1.370	0.190
ours (GPT-5.2 [6])	✓	✗	0.312 ± 0.154	0.848 ± 0.321	0.732 ± 0.400	0.116
	✓	✓	$\mathbf{0.415 \pm 0.230}$	0.928 ± 0.440	0.695 ± 0.388	**0.233**

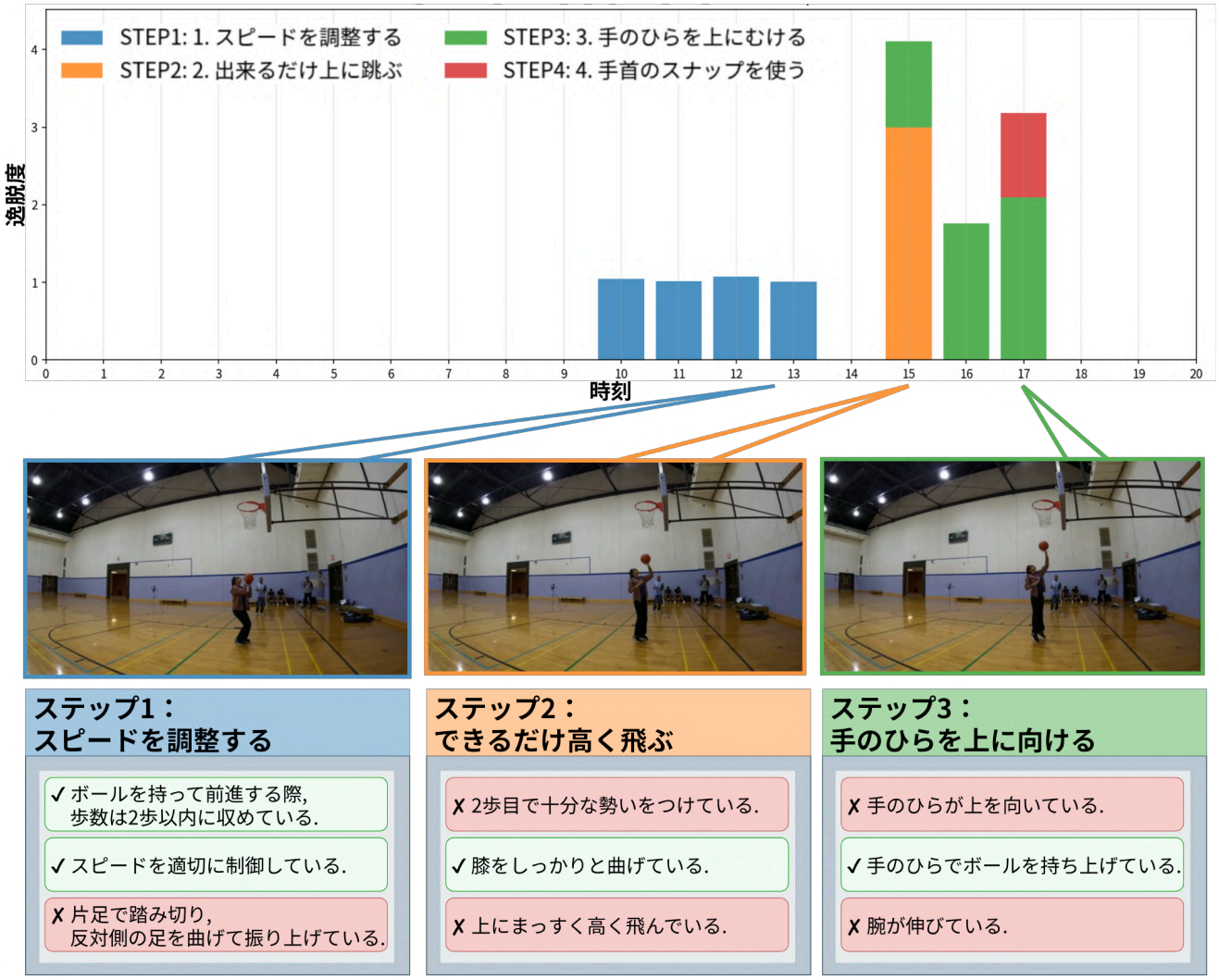


図 3: 逸脱度推定結果の可視化例. 上段は各時刻における各ステップごとの逸脱度を示し, 下段は対応する動画フレームと, 各ステップで定義された評価規範を満たしているか否かを手動で判定結果を示している. 逸脱度が大きく推定された区間では, 対応するステップにおいて複数の評価規範が満たされていないことが確認できる.

現へと埋め込み, 動画クリップとの類似度に基づいて対応付けを行う. 一方, 提案手法ではテキストと図解を含むドキュメント全体と動画クリップを VLM に与え, 「このクリップはどのステップに対応するか」という QA 形式の問題として定式化することで, VLM を確率出力可能な分類器として利用している. この枠組みにより, 本来は対話や記述理解を目的として学習された VLM を追加学習を行うことなく Video Moment Retrieval の文脈へと適用できる. 入力情報に関するアブレーション結果を見ると, 説明テキストのみの場合でも IoU は 0.312 ± 0.154 と既存手法を上回っており, この定式化自体が zero-shot かつ

専門的な技能に対して有効に機能していることが分かる．さらに，スクリーンショットを併用すると IoU は 0.415 ± 0.230 まで向上し, 図やイラストに含まれる視覚的文脈をそのままクエリに取り込むことで曖昧な記述や技能特有の表現と動画中の状態との対応付けがより安定することが確認できる．このように, 提案手法は教本に含まれるテキストだけでなく, 図やイラストといった視覚情報を含む文脈全体をクエリとして扱うことができる．これは VLM を基盤とすることで初めて可能となる性質であり, 固定的なテキスト表現に基づく従来の Video Moment Retrieval 手法では原理的に扱えない情報である．教本が本来備えているマルチモーダルな表現を歪めることなく検索過程に持ち込める点は, 専門的かつ曖昧な技能記述に対しても頑健に対応できるという, 本手法の重要な特徴である．

熟練度に基づく逸脱度の比較． 表 2 に, 熟練者動画群および初心者動画群における逸脱度の比較結果を示す．いずれの手法においても, 初心者群の逸脱度平均は熟練者群より大きい傾向を示すが, その差の大きさには手法間で顕著な違いが見られた．InternVideo2 の差は 0.044 と小さく，熟練者と初心者の分布がほぼ重なっていることが示唆された．R^2-Tuning では差が 0.190 と一定の分離が見られ, 提案手法では差が 0.233 と最も大きく, 熟練度差をより明確に反映できていることが確認できた．特筆すべき点は, 本手法が熟練度ラベルを直接用いることなく, 手順に基づく規範への逸脱という観点から算出された指標のみで, 結果として初心者と熟練者を分離できていることである．これは, 提案手法によって構築されたルーブリックおよびステップ区間推定により, 技能遂行に内在する質的な差異を適切に捉え, それを逸脱度として定量化できていることを示唆している．この傾向は, 提案手法が推定する逸脱度がステップ依存の規範集合に基づき計算される点と整合的である．すなわち, ステップ区間推定が妥当であるほど各クリップは適切なステップ文脈の下で評価され, 初心者に特有の違反や規範未達が逸脱度として顕在化する．一方, ステップ割当てが不安定な場合, クリップが無関係なステップ規範と照合されることで逸脱度が過度に増減し, 熟練度差がノイズに埋もれやすい．したがって, ステップ区間推定の精度が逸脱度を熟練度指標として安定に機能させる上で重要であることが示された．

逸脱度の定性評価． 図 3 は，提案手法によって推定された各時刻における逸脱度と，対応するステップの評価規範に基づく逸脱判定の例を示している．ステップ 1 およびステップ 4 に対応する区間では逸脱度が相対的に小さい一方で，15 クリップ目付近ではステップ 3，17 クリップ目付近ではステップ 4 に対応する逸脱度が大きくなっていることが確認できる．各規範項目に対する人手による判定結果と比較すると, 逸脱度が大きく推定された区間では対応するステップにおいて満たされていない規範項目の数が相対的に多いことが確認できる．すなわち, 提案手法によって推定された逸脱度の増減は, 人手判断と整合した挙動を示すことを示している．

5 おわりに

本研究では, 書籍や教本に記述された技能知識を参照基準として用い, 学習者動画中の行動を評価するコーチングエージェントの構築手法を提案した．提案手法では, ドキュメント内の記述を手順ステップと評価規範からなるルーブリックとして構造化し, ステップの順序制約を考慮した区間推定を行った上で, 各時刻における行動の逸脱度を定量的に算出する．評価実験の結果, 提案手法は参照動画や追加学習を用いることなく, 動画中の行動をドキュメントに基づく規範と対応付けられることが確認された．また, 推定されたステップ区間は人手による大まかな区切りと整合する傾向を示し, さらに算出された逸脱度は熟練度の違いや, 人手で確認した規範未達の区間と対応した挙動を示した．これらの結果は, 提案手法が技能遂行の時間構造と品質を同時に捉えるための基盤として有効であることを示す．

一方, 動作が極めて短時間で生じるステップや, 空間的に局所的な運動に依存する規範に対しては, 区間推定や逸脱度推定の精度に課題が残ることも確認された．今後の方向として, より細粒度な時間分解能での解析や, 文書記述の曖昧さを補正する仕組みを導入することで, より幅広い技能や教材に適用可能なコーチング支援へと拡張していくことが有望である．

文　献

[1] Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. ExpertAF: Expert actionable feedback from video. In *CVPR*, 2025.

[2] Adelle Atkinson, Christopher J. Watling, and Paul L. P. Brand. Feedback and coaching. *European Journal of Pediatrics*, Vol. 181, pp. 441–446, 2022.

[3] Phillip Dawson, Michael Henderson, Patrick Mahoney, Michael Phillips, Tracii Ryan, David Boud, and Elizabeth Molloy. What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education*, Vol. 44, No. 1, pp. 25–36, 2019.

[4] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *CVPR*, 2024.

[5] Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. R^2-Tuning: Efficient Image-to-Video Transfer Learning for Video Temporal Grounding. 2024.

[6] OpenAI. GPT-5, 2025. `https://platform.openai.com/docs/models/gpt-5`.

[7] Stephanie J. Sohl, Deborah Lee, Heather Davidson, Blaire Morriss, Rebecca Weinand, Katherine Costa, Edward H. Ip, James Lovato, Russell L. Rothman, and Ruth Q. Wolever. Development and validation of the Health Coaching Index. *Patient Education and Counseling*, Vol. 104, No. 3, pp. 642–648, 2022.

[8] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. InternVideo2: Scaling Foundation Models for Multimodal Video Understanding. 2024.

[9] Benedikt Wisniewski, Klaus Zierer, and John Hattie. The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology*, Vol. 10, , 2020.

[10] Wei-Hsin Yeh, Yu-An Su, Chih-Ning Chen, Yi-Hsueh Lin, Calvin Ku, Wenhsin Chiu, Min-Chun Hu, and Lun-Wei Ku. CoachMe: Decoding Sport Elements with a Reference-Based Coaching Instruction Generation Model. In *ACL*, 2025.

[11] 森圭司. 目で学ぶシリーズ 3 見るだけでうまくなる！バスケットボールの基礎. ベースボール・マガジン社, 東京, 2020. 第 1 版 第 1 刷.