音声認識導入による議論評価システムの逐次的分析

李 昂[1]　嶋田 和孝[2]

概要：近年，試験において，受験者のコミュニケーション能力を評価する手段としてグループディスカッションが用いられている．参加者の能力向上には，適切なフィードバックが必要不可欠であるが，フィードバックの提供には多大な労力を必要とする．このような背景から，複数人議論の品質評価や分析結果を提供するシステムが求められている．議論のフィードバックは，議論の内容が参加者の記憶に強く残っている間に提供されることで，より効果的な改善を実現できるが，そのためには議論の分析を逐次的に行う必要がある．そこで，本論文では音声認識を利用し，逐次的に得られる議論内容をシステムの入力とすることにより，議論の逐次分析を実現する．

キーワード：議論，音声認識，品質評価，議論分析

Real-time Analysis with Speech Recognition for Quality Assessment of Debate

Abstract: In recent years, group discussions have been used in exams to assess participants' communication skills. Effective feedback is essential for participants' improvement but requires significant effort. Therefore, a system for evaluating and analyzing discussion quality is needed. Providing feedback during the discussion is still fresh in participants' minds enhances effectiveness, necessitating real-time analysis. This paper proposes a system that utilizes speech recognition to sequentially analyze discussions by processing spoken content as input.

Keywords: Debate, Speech Recognition, Quality Assessment, Debate Analysis

1. はじめに

近年，大学入試や就職試験において，受験者のコミュニケーション能力を評価する手段としてグループディスカッションが広く活用されている．また，教育現場においてもコミュニケーション能力の向上を目的として，グループディスカッションの取り入れられる機会が増加している．しかし，グループディスカッションでは正解が存在しない課題を扱うことから，評価者が定量的に評価を行うのは非常に困難である．さらに，試験や教育現場では，複数のグループによるディスカッションが行われるため，全てのグループの評価をするためには膨大な労力を要する．

また，議論参加者のコミュニケーション能力向上を目的とする場合，単に評価を行うだけでなく，議論の具体的な評価点や改善点を示すフィードバックを提供することが不可欠である．しかし，議論のフィードバックを行うためには，議論内容を適切に分析し，評価点・改善点を明確にする必要がある．そのため，フィードバックの提供もまた，評価と同様に多大な労力を要し，効率的に実現することが困難である．このような背景から，複数人による議論の品質評価や分析結果を提供するシステムの構築が重要な課題となっている．そこで，本論文では実用的な議論の評価・分析システムの構築を行う．

橋口ら [1] は，複数人議論データセットに含まれている議論音声の人手書き起こしを議論評価モデルの入力として使用し，議論の品質評価を実施している．しかし，実際には議論音声を正確に人手で書き起こす作業には多大な時間

[1] 九州工業大学 大学院情報工学府
Department of Creative Informatics, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
[2] 九州工業大学 大学院情報工学研究院 知能情報工学研究系
Department of Artificial Intelligence, Kyushu Institute of Technology 680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN

と労力を要する．さらに，議論評価モデルの入力に人手書き起こしを用いる場合，リアルタイムでの議論評価が不可能となり，実用性が下がる．そのため，人手で議論音声を書き起こすコストを抑えつつ，リアルタイム性のある議論評価を実現するためには，議論音声の音声認識結果を議論評価モデルの入力として利用することが望ましい．

一般に，音声認識による書き起こしは不完全であり，誤りを含むことが多い．しかし，我々は音声認識結果を議論評価モデルの入力に用いた場合でも，人手による書き起こしを用いた場合と比較して，評価精度が低下しないことを確認している [2]．一方で，精度の面では十分とはいえない．そこで，本論文では音声認識結果を入力とする評価モデルの精度向上を目指して，2 つのアプローチを提案し，その有効性を検証する．

1 つ目のアプローチは，言語以外の情報の利用である．一般に，グループディスカッションにおける発話者の意見や主張は，言語から解釈できる情報のみではなく，発言者の非言語的要素からも大きな影響を受けるとされている [3]．そのため，議論の品質評価では，議論参加者の発言内容だけでなく，動作・音声情報や感情情報などの様々な情報を考慮することが重要である．そこで，本論文では，議論内容に加えて，議論参加者の動作・音声情報や感情情報をモデルに与えることにより，評価モデルの改良を試みる．

2 つ目のアプローチは，不均衡データへの対策である．機械学習に用いられるデータは，クラス間の分布が不均衡な場合が存在する．しかし，機械学習で分類問題を解く場合，データの少ないクラスにおける予測精度が低下する傾向にある．そのため，不均衡データは議論評価の精度低下を招く可能性がある．したがって，議論評価の精度を向上させるためには，不均衡データへの対策を講じることが重要である．そこで，本研究では不均衡データの対策として，LLM による議論評価，LLM を用いたデータ拡張，損失関数の変更を行う．

さらに，実用的なシステムには，議論の評価に加えて，フィードバックを提供する機能が求められる．議論のフィードバックを行う場合，議論中の多様な情報を分析し，分析結果を可視化することが重要である．また，議論のフィードバックは，議論の内容が参加者の記憶に強く残っているうちに提供されることで，より効果的な改善を実現できる．しかし，即座にフィードバックを行うためには，議論の分析を逐次的に行う必要がある．そこで，本論文では音声認識を利用し，逐次的に得られる議論内容をシステムの入力とすることにより，議論の逐次分析を実現する．

2. データセット

本節では，本論文で使用するデータセットについて説明する．本研究では，複数人による議論データセットとして，Shiota ら [4] が作成した Kyutech Debate Corpus を使用する．Kyutech Debate Corpus は，表 1 に示した 5 つの議論テーマに基づき，大学生・大学院生 4 人 1 組で行われた討論と合意形成の計 10 対話の議論データが収録されている．議論データには，実際の議論の映像や音声に加えて，映像データから得られた顔特徴点の座標，視線，表情などの顔情報や上半身の骨格特徴点である動作情報，音声データから得られた声の大きさや高さなどの音声情報が含まれている．顔情報には各議論参加者の顔特徴点の座標，視線方向，頭の向き・回転角，Facial Action Units[*1](AUs) が含まれている．動作情報には各議論参加者の各フレームにおける骨格，手の特徴点が含まれている．音声情報には各議論参加者の 13 次元 MFCC，RMS，基本周波数，スペクトル重心，ジッタ，シマが含まれている．13 次元 MFCC は発話時の口や喉の形を表現する声の特性，RMS は声の大きさ，基本周波数は声の高さ，スペクトル重心は声の明るさ，ジッタは声の高さのゆらぎ，シマは声の大きさのゆらぎを示す．また，議論データの各対話は，トピックセグメンテーション手法により，トピックごとに分割されており，10 対話の議論データは合計 178 個の議論セグメントに分割されている．

表 1 議論テーマ

(小中高の) 生徒は制服を着用すべきである
成人の拳銃所持・携帯の権利を認めるべきである
小中高の教材はタブレットに置き換えるべきである
飲酒可能年齢は 20 歳から下げられるべきである
未成年の暴力的ゲームのプレイを禁止すべきである

表 2 各評価軸におけるスコア分布

評価軸	Low	Middle	High
有効性	9	97	72
合理性	13	89	76

各議論セグメントには，Wachsmuth ら [5] が提案した議論品質評価基準に基づき，有効性と合理性という 2 つの評価軸に沿ってスコアがつけられている．有効性の高い論・議論とは，その主張が聞き手に対して納得や同意を促す効果を有する論・議論を表す．合理性の高い論・議論とは議題の解決に対して十分に受容可能な形で寄与する論・議論を意味する．スコアは有効性と合理性という 2 つの評価軸のそれぞれで，「Low」「Middle」「High」の 3 つに分類されている．2 つの評価軸における 178 個の議論セグメントの「Low」「Middle」「High」の分布は表 2 のようになっている．本論文では，Kyutech Debate Corpus における 2 つの評価軸のうち，有効性のラベルが付与された議論セグメントを対象に議論の品質評価・分析を行う．

3. 音声認識を用いた議論評価

本論文では，音声認識を用いた議論評価システムが基盤

[*1] https://www.cs.cmu.edu/ face/facs.htm

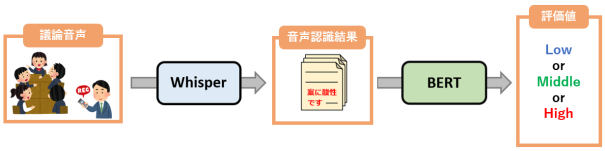

図 1 議論評価システムの概要

となる．これは我々によって既に構築されている [2]．本節では，その基盤システムについて概略と精度を示す．3.1 節では，議論評価システムの構成について説明する．3.2 節では，実験設定及び実験結果について述べる．

3.1 議論評価システムの構成

図 1 に，先行研究で用いた議論評価システムの概略を示す．議論評価モデルには，橋口ら [1] が提案した既存モデルを使用し，評価値の推定を行う．既存モデルには，BERT (Bidirectional Encoder Representations from Transformers) [6] が使用されている．BERT は，Masked Language Model および Next Sentence Prediction という 2 つの事前学習タスクを通じて文脈情報を効果的に捉えることができるモデルである．また，議論音声の書き起こしに使用する音声認識モデルには，OpenAI が公開した Whisper [7] の large モデルを使用する．Whisper は多言語対応の音声認識モデルであり，large モデルにおける日本語の単語誤り率は 6.4 %と，高精度の書き起こしを実現している．

3.2 実験

本節では音声認識結果を用いた議論評価モデルの実験について述べる．3.2.1 節では実験設定について述べる．3.2.2 節では実験結果と考察を述べる．

3.2.1 実験設定

議論評価モデルの評価方法について説明する．評価方法については，橋口ら [1] と同様に，Kyutech Debate Corpus に含まれる 10 対話のデータを，訓練データ 8 対話，検証データ 1 対話，テストデータ 1 対話に分割し，データセット中の全体がテストデータとして用いられるように設定し，10 対話交差検証を行った結果を 1 実験の評価とする．評価指標には分類タスクの指標として用いられる F 値を使用する．結果の頑健性保証のため，10 対話交差検証を 5 回繰り返し，その重み付き平均を報告する．

また，実験に使用する議論評価モデルの設定について説明する．BERT は東北大学が公開しているモデルを使用する[*2]．損失関数は CrossEntropy，最適化関数は AdamW，学習率は 1e-5，バッチサイズは 8，エポック数は 50 に設定して実験を行う．

3.2.2 実験結果および考察

実験結果を表 3 に示す．Ave. は Low，Middle，High における F 値の重み付き平均である．また，表内に存在する数字の太字は，各ラベルの F 値および重み付き平均において最も高い精度を表している．実験結果より，議論評価の入力における書き起こし手法として音声認識を用いた場合，人手での書き起こしを用いた場合に比べ，精度の低下は見られなかった．これにより，議論音声の音声認識結果を評価モデルに使用した場合，議論評価の精度に悪影響を及ぼさないことが確認された．したがって，議論評価モデルの入力における音声認識結果の有用性が示された．

[*2] https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking

表 3 各書き起こし手法による評価精度

書き起こし手法	Low	Middle	High	Ave.
人手 [1]	0.000	0.634	**0.466**	0.534
音声認識	0.000	**0.650**	0.452	**0.537**

表 4 学習・テストに異なる書き起こしを用いた場合の評価精度

学習 / テスト	Low	Middle	High	Ave.
人手 / 音声認識	0.000	0.635	0.367	0.494
音声認識 / 人手	0.000	0.646	0.246	0.451

また，人手書き起こしと音声認識の両書き起こし手法において，Low ラベルの F 値が 0 となっている．これは，表 2 に示されている有効性ラベルにおいて，Low ラベルのデータ数が他のラベルに比べて極端に少ないことが原因となり，議論評価モデルが正確に予測できなかった可能性があると考えられる．

なお，本論文では，訓練・検証データとは異なる書き起こし手法を用いたデータをテストデータに使用した場合の議論評価精度を調査する追加実験も行っている．追加実験の結果を表 4 に示す．表 4 の Ave. の数値は Low，Middle，High における F 値の重み付き平均である．実験結果より，訓練・検証データとテストデータに異なる書き起こし手法で得られたデータを使用した場合，表 3 の同一書き起こし手法によるデータを使用した場合に比べ，精度の低下が見られた．したがって，BERT を用いた議論評価モデルでは，訓練・検証データとテストデータに同一の書き起こし手法で得られたデータを使用することが重要であると考えられる．

4. 議論評価モデルへの特徴量追加と不均衡データ対策

本節では，音声認識結果を入力とする議論評価モデルの評価精度向上を目的とした 2 つのアプローチとその実験について述べる．議論の品質評価では，議論参加者の発言内容だけでなく，動作・音声情報や感情情報などの言語以外の情報が影響すると考えられる．そこで，4.1 節では言語以外の情報を利用した議論評価を行う．また，機械学習に用いられるデータは，クラス間の分布が不均衡である場合が存在する．実際に，表 2 のように，本研究で使用する複数人議論データセットにおいても，データの偏りが生じて

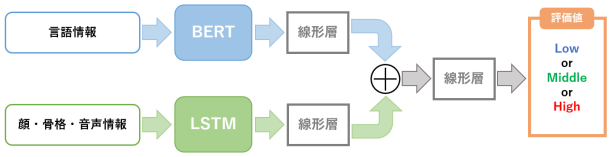

図 2 動作や音声の情報を利用するモデルの概要

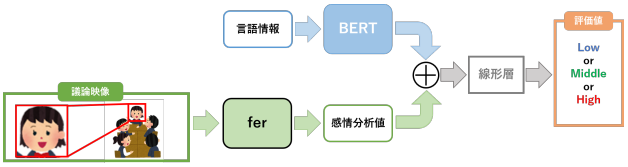

図 3 感情特徴量を利用するモデルの概要

いる．しかし，機械学習を用いた分類タスクでは，低頻度クラスの予測精度が低下する傾向が見られる．そこで，4.2 節では不均衡データへの対策を講じる手法を導入する．4.3 節では提案する 2 つのアプローチの実験について述べる．

4.1 言語以外の情報の利用

本節では言語情報に加えて，言語以外の情報を利用するアプローチについて説明する．一般に，議論では身振りや手ぶり，表情，発声の仕方により，聞き手への説得力や話し手への共感，不満が表現される．そのため，議論参加者の動作や音声の情報は議論の質に影響を与えていると考えられる．そこで，4.1.1 節では，議論中の発言内容に加えて，動作や音声の情報を利用する．また，議論では発話者の発話内容に対する聞き手の感情が，議論内容への賛同や反対を示している．したがって，議論参加者の感情情報も，議論の質に影響していると考えられる．そこで，4.1.2 節では，議論中の発言内容に加えて，議論中の映像から推定した各議論参加者の感情分析値を利用する．

4.1.1 動作や音声の情報の利用：Multi

本節では，議論中の発言内容に加えて，動作や音声の情報を利用する議論評価手法について説明する．Shiota ら [4] らは議論評価モデルに，Attention 機構 [8] を用いた LSTM (Long Short-Term Memory) [9] を利用し，議論の発言内容に加えて議論参加者の顔や体の動作，音声の情報をモデルに与えることにより，発言内容のみを利用した場合よりも高い評価精度を獲得している．一方で，発言内容の学習には文脈の理解が可能な BERT の方が適していると考えられる．そこで，本手法では BERT と Attention 機構有りの LSTM を統合したモデルを構築し，議論の発言内容，議論中の顔や体の動作，音声の情報をモデルに入力として与え，議論の評価値を推定する手法を提案する．以降，この手法を手法 $_{Multi}$ と呼ぶ．

図 2 に提案するモデルの構成について説明する．まず，言語情報を BERT，顔・骨格・音声情報を LSTM に入力し，各モデルの出力をそれぞれ線形層に入力する．その後，線形層から出力された 2 つのベクトルを結合し，結合したベクトルを線形層に入力することにより，評価ラベルを推定する．

4.1.2 感情特徴量の導入：Emotion

一般に，議論では発話者の発話内容に対する聞き手の感情は，議論内容への賛同や反対を示している．例えば，聞き手が発話内容に対して共感や好意的な感情を抱いている場合，議論内容がより説得力を持つ可能性が高い．一方，聞き手が議論中に不快感や抵抗感を覚える場合，議論内容に不服を示していると判断できる．また，話し手の感情表現は，議論の信頼性の補強や相手の注意を引きつける役割を持つ．例えば，熱意や情熱を示すことにより，聞き手に対して主張の重要性を印象づけ，主張の受容性を向上させる効果を果たす．さらに，適切な感情表現は，話し手が自分の主張に対して深い理解や関心を持っていることを示し，聞き手の心に響きやすくなる．さらに，感情豊かな表現は単調な話し方に比べ，聞き手の興味を引きつける効果があるため，議論の内容がより効果的に伝わる可能性が高まる．このように，議論参加者の感情は議論の品質に影響を与える要素であると考えられる．そこで，本手法では BERT の出力に，議論映像から得られた感情分析値を組み合わせることにより，議論参加者の感情を考慮した議論の評価値推定手法を提案する．以降，この手法を手法 $_{Emotion}$ と呼ぶ．

図 3 に提案するモデルの構成について説明する．まず，各議論参加者を撮影した上半身カメラの映像に対し，fer[*3][10] を利用して感情分析を行う．fer は入力された顔の画像や映像に対する恐怖，中立，幸せ，悲しい，怒り，嫌悪の 6 つの感情の信頼度を計算可能なオープンソースの Python ライブラリである．次に，fer を用いて得られた 6 つの感情の感情分析値と，BERT に言語情報を入力して得られた出力を結合する．最後に，結合したベクトルを線形層に入力することにより，評価ラベルを推定する．

4.2 不均衡データへの対策

本節では，不均衡データに対するアプローチについて説明する．LLM は，追加学習を行わずに言語処理タスクに適用可能なため，データの不均衡による影響を受けにくい．そこで，4.2.1 節では，LLM を用いた議論評価を行う．また，不均衡なデータの対応策として，LLM を用いた少数派クラスのデータを人工的に生成するデータ拡張手法が広く用いられている．そこで，4.2.2 節では，LLM を使用したデータ拡張を行う．さらに，モデルの損失関数にクラスの重要度を反映させることにより，モデルの学習における少数派クラスの影響を増大する手法も，不均衡データへの対策として活用されている．そこで，4.2.3 節では，損失関数の変更を行う．

[*3] https://github.com/justinshenk/fer

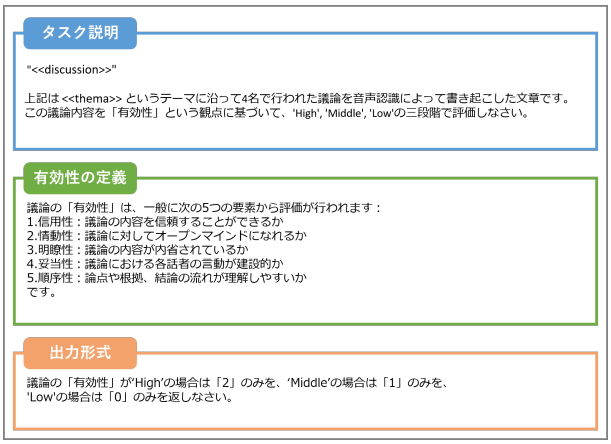

図 4　議論評価を行うプロンプト

4.2.1　LLM：GPT-4o，GPT-4o-mini

近年，自然言語処理分野において，大規模言語モデル（LLM: Large Language Models）が広く用いられている．LLM は，追加学習無しで言語処理タスクを解くことが可能であり，不均衡なデータの影響を受けない．そのため，LLM を用いた議論の評価値推定はデータ数の少ないクラスに対して，有効性が高いと考えられる．そこで，本手法では LLM を利用し，追加学習せずに議論の評価を行う．

議論の有効性という基準に基づいて議論評価を行うプロンプト ($\text{Prompt}_{\text{Evaluation}}$) を図 4 に示す．図 4 中の`<<discussion>>`は評価対象の議論，`<<theme>>`は議論のテーマを表している．

$\text{Prompt}_{\text{Evaluation}}$ はタスク説明，有効性の定義，出力形式の 3 つの要素から構成されている．タスク説明では，LLM に対して実施させる具体的なタスク内容を指示する．ここでは，議論の有効性に基づいて議論の評価をする指示に加えて，議論のテーマや議論音声の書き起こし手法，議論の参加人数等もプロンプトとして提示する．有効性の定義では，Wachsmuth ら [5] による有効性の定義に従い，有効性を決める 5 つの要素（信用性：Credibility, 情動性：Emotional Appeal, 明瞭性：Clarity, 妥当性：Appropriateness, 順序性：Arrangement）をプロンプトに明記する．また，LLM がどのような形式でレスポンスを返すべきかについても，具体的な出力形式を指示する．

本手法では議論評価を行う LLM として，GPT-4o[*4] と GPT-4o-mini[*5] を利用し，議論の評価値を推定する．以降，GPT-4o により議論評価を行う手法を手法 $_{GPT-4o}$，GPT-4o-mini により議論評価を行う手法を手法 $_{GPT-4o-mini}$ と呼ぶ．

4.2.2　データ拡張：Augmentation

橋口 [1] らは，GPT-4 [11] によるデータ拡張を低頻度クラスに対して行うことにより，人手書き起こしを入力とする議論評価モデルの精度を向上させている．そこで，本手法では音声認識結果を入力とする議論評価モデルに対し，GPT-4 によるデータ拡張手法を導入し，その有効性を検証する．以降，この手法を手法 $_{Augmentation}$ と呼ぶ．

本手法では，橋口らの手法に倣い，低頻度クラスである議論の有効性が Low の疑似データを 100 件生成する．その後，BERT に対してデータ拡張手法を適用し，議論の評価ラベルの推定を行う．

4.2.3　損失関数の変更：ICF，FL

機械学習における不均衡データの対策として，損失関数を変更する手法も広く採用されている．本手法では，損失関数として Inverse Class Frequency Loss (ICF) や Focal Loss (FL) [12] を使用することにより，低頻度ラベルの予測精度を向上させ，全体の予測精度向上を図る．

ICF は式 (1) で表され，クラスのデータ数の逆数を重みとして Cross Entropy に掛けた損失関数である．ただし，n はデータ数，C_n はクラスのデータ数，p_n は予測確率を示している．

$$ICF = -\sum_{i=1}^{n} \frac{1}{C_n} \log(p_n) \tag{1}$$

一方で，FL は式 (2) で表され，正解クラスの予測確率を 1 から引いた数を重みとして Cross Entropy に掛けた損失関数である．ただし，γ はハイパーパラメータ，p_n は予測確率を示している．本論文ではパラメータ γ について，$\gamma = 2.0$ とする．

$$FL = -\sum_{i=1}^{n} (1 - p_n)^{\gamma} \log(p_n) \tag{2}$$

本アプローチでは，ICF と FL の 2 種類の損失関数を適用した BERT を利用し，議論の評価値を推定する．以降，ICF を損失関数に用いる手法を手法 $_{ICF}$，FL を損失関数に用いる手法を手法 $_{FL}$ と呼ぶ．

4.3　実験

本節では，言語以外の情報の利用と不均衡データへの対策の 2 つのアプローチによるモデルの評価精度を調査する実験について述べる．具体的には 3.1 節で述べた橋口ら [1] の BERT モデルを使用する手法をベースライン手法とし，提案する各手法と議論評価の精度を比較する．以降，ベースライン手法を手法 $_{BERT}$ と呼ぶ．

4.3.1 節では，言語以外の情報の利用を行ったモデルの実験設定と実験結果について述べる．4.3.2 節では，不均衡データへの対策を行ったモデルの実験設定と実験結果について述べる．

4.3.1　言語以外の情報の利用

手法 $_{Multi}$ と手法 $_{Emotion}$ における実験設定について説明する．議論の品質評価の評価方法，BERT のモデルについては 3.2 節の実験設定に倣う．手法 $_{Multi}$ では，LSTM の隠れ層は 768 次元，損失関数は CrossEntropy，最適化関

*4　https://platform.openai.com/docs/models#gpt-4o
*5　https://platform.openai.com/docs/models#gpt-4o-mini

表 5 言語以外の情報を利用したモデルの評価精度

手法	Low	Middle	High	Ave.
手法 $_{BERT}$	0.000	**0.650**	**0.452**	**0.537**
手法 $_{Multi}$	0.000	0.637	0.204	0.429
手法 $_{Emotion}$	0.000	0.616	0.398	0.497

数は AdamW，学習率は 1e-5，バッチサイズは 8，エポック数は 50，ドロップアウトは 0.2 に設定して実験を行う．手法 $_{Emotion}$ では，損失関数は CrossEntropy，最適化関数は AdamW，学習率は 1e-5，バッチサイズは 8，エポック数は 50 に設定して実験を行う．

手法 $_{Multi}$ と手法 $_{Emotion}$ の実験結果表 5 に示す．Ave. は Low，Middle，High における F 値の重み付き平均である．また，表内に存在する数字の太字は，各ラベルの F 値および重み付き平均において最も高い精度を表している．

手法 $_{Multi}$ と手法 $_{Emotion}$ の実験結果表 5 に示す．Ave. は Low，Middle，High における F 値の重み付き平均である．また，表内に存在する数字の太字は，各ラベルの F 値および重み付き平均において最も高い精度を表している．

今回提案した手法 $_{Multi}$i と手法 $_{Emotion}$ について実験結果と考察を述べる．手法 $_{Multi}$ は手法 $_{BERT}$ と比較して精度の低下が確認された．BERT と LSTM の 2 つのモデルは，それぞれ異なる意味を持つベクトルを出力する．これらの異なる意味を持つベクトルを単純に結合した場合，後続の線形層が両方のベクトルの情報を十分に学習できなかったと考えられる．

また，手法 $_{Emotion}$ も手法 $_{BERT}$ と比較して精度が低下していた．これは fer により得られた感情情報と BERT の出力を単純結合しただけでは，異なる特徴ベクトルがモデルにとってノイズとなり，線形層がうまく学習できなかったと考えられる．

4.3.2 不均衡データへの対策

手法 $_{Augmentation}$，手法 $_{ICF}$，手法 $_{FL}$ における実験設定について説明する．議論の品質評価の評価方法, BERT のモデルについては 3.2 節の実験設定に倣う．手法 $_{Augmentation}$ では，損失関数は CrossEntropy，最適化関数は AdamW，学習率は 1e-5，バッチサイズは 8，エポック数は 50 に設定して実験を行う．手法 $_{ICF}$ では，損失関数は Inverse Class Frequency Loss，最適化関数は AdamW，学習率は 1e-5，バッチサイズは 8，エポック数は 50 に設定して実験を行う．手法 $_{FL}$ では，損失関数は Focal Loss，最適化関数は AdamW，学習率は 1e-5，バッチサイズは 8，エポック数は 50 に設定して実験を行う．

実験結果を表 6 に示す．Ave. は Low，Middle，High における F 値の重み付き平均である．また，表内に存在する数字の太字は，各ラベルの F 値および重み付き平均において最も高い精度を表している．

不均衡データに対する手法について実験結果と考察を述べる．まず，手法 $_{GPT-4o}$ と手法 $_{GPT-4o-mini}$ については，いずれも Low ラベルの F 値が向上していた．これは，LLM は追加学習が不要であるため，低頻度クラスの影響を受けなかったと考えられる．一方で，両手法において，Middle ラベルの F 値や High ラベルの F 値が低下し，全クラスにおける予測精度も手法 $_{BERT}$ に比べて低下する結果となった．これは，LLM が追加学習を行わないことにより，議論評価タスクに特化した特徴が不足したためであると考えられる．

表 6 不均衡データへの対策を行ったモデルの評価精度

手法	Low	Middle	High	Ave.
手法 $_{BERT}$	0.000	**0.650**	0.452	**0.537**
手法 $_{GPT-4o}$	0.119	0.330	0.000	0.184
手法 $_{GPT-4o-mini}$	**0.180**	0.603	0.000	0.337
手法 $_{Augmentation}$	0.033	0.596	0.466	0.403
手法 $_{ICF}$	0.000	0.500	**0.528**	0.486
手法 $_{FL}$	0.000	0.614	0.481	0.529

手法 $_{Argumentation}$ は Low ラベルの F 値が 0 近傍の値となっており，低頻度クラスにおける予測精度の向上はほとんど見られなかった．また，手法 $_{BERT}$ に比べ，全クラスにおける予測精度が低下していた．手法 $_{Augmentation}$ では，GPT-4 によって生成された議論データを疑似データとしてモデルの学習に使用した．しかし，生成データが実データである議論音声の音声認識結果と比較すると認識誤りがなく，整ったデータであるため，学習データ間に差異が生じたことが影響していると考えられる．

手法 $_{ICF}$ と手法 $_{FL}$ では，低頻度クラスである Low ラベルの F 値は 0 となり，手法 $_{BERT}$ と同様の結果であった．この要因として，Low ラベルのデータが少なすぎることにより，モデルがそのクラスの特徴を十分に学習できなかったためであると考えられる．また，手法 $_{BERT}$ に比べて，Middle ラベルの精度が低下しており，High ラベルの精度は向上していた．これは，損失関数を変更したことにより，高頻度クラスの影響を抑制し，低頻度クラスの影響を増大させたことによるものと考えられる．

5. 議論の逐次的分析

本節では議論の逐次的分析について述べる．議論参加者に対し，議論のフィードバックを行う場合，議論中の様々な情報を分析し，分析結果を可視化する必要がある．また，議論のフィードバックは，議論の内容や進行状況が参加者の記憶に鮮明に残っている間に提供されることにより，具体的かつ効果的な改善を促進できる．しかし，即時的なフィードバックを行うためには，議論の分析を逐次的に行う必要がある．そのため，議論の分析結果の可視化を逐次的に行うフィードバックシステムの実現が重要な課題となっている．そこで，本論文では議論音声の音声認識結果を利用した逐次的分析を行う．

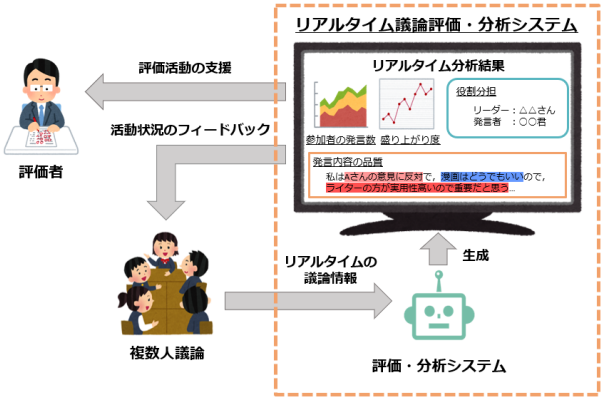

図 5　リアルタイムで議論のフィードバックを行うシステムの概要

5.1 節では本論文で行う分析の内容について説明する．5.2 節では分析手法について，5.3 節では分析結果について述べる．なお，4 節で提案した種々の手法は精度向上に貢献しなかったため，本節での逐次的分析の入力としては，3 節で述べた音声認識結果を利用したモデルの出力を利用する．

5.1　分析の内容

図 5 にリアルタイムで議論を評価・分析し，フィードバックを行うシステムの概要を示す．議論のフィードバックシステムでは，参加者の発言数や議論の盛り上がり度，発言内容の品質などの様々な情報を分析し，その結果を可視化する必要がある．そこで，本論文では発言内容の品質について分析し，議論品質に良い影響を与えた発言内容と悪い影響を与えた発言内容を明確化する．

また，議論のフィードバックでは，逐次的に分析結果が提供されることで，より効果的なシステムを実現できる．そこで，本分析では音声認識を活用し，逐次的に得られる発言内容をシステムの入力とすることにより，議論の逐次分析を実現する．

5.2　分析手法

本論文では議論内容の分析として，議論品質に良い影響を与えた内容と悪い影響を与えた内容を明確化する．分析には，3.1 節で述べた音声認識結果を入力とする議論評価モデルに対し，SHAP (SHapley Additive exPlanation) [13] を用いる．SHAP は機械学習モデルを解釈するための手法の一つであり，各特徴量の寄与度を示す SHAP 値を用いて，機械学習モデルの出力を線形和として表現する方法である．本論文では，BERT によりトークン化された音声認識結果の各トークンに対して，High と Low の SHAP 値を計算する．その後，式 (3) のように，High の SHAP 値から Low の SHAP 値を引き，その値を 2 で割ることで，各トークン i が議論の品質に良い影響を与えたか悪い影響を与えたかを示す．ただし，$\mathrm{SHAP}_{i,\mathrm{High}}$ はトークン i における High の SHAP 値，$\mathrm{SHAP}_{i,\mathrm{Low}}$ はトークン i における Low の SHAP 値を示している．

$$C_i = \frac{\mathrm{SHAP}_{i,\mathrm{High}} - \mathrm{SHAP}_{i,\mathrm{Low}}}{2} \tag{3}$$

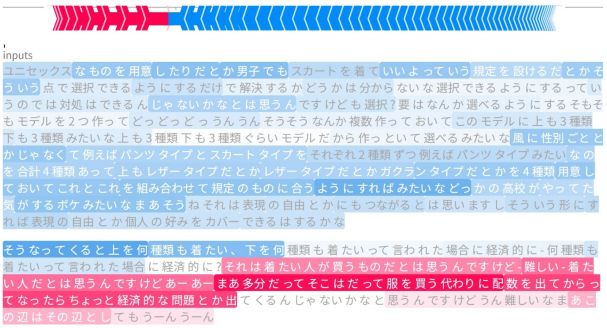

図 6　議論の分析結果

5.3　分析結果

図 6 に，分析結果の一例を示す．この例は，「(小中高の) 生徒は制服を着用すべきであるという」議論の一部に対して，議論品質に良い影響を与えた内容と悪い影響を与えた内容を分析した結果である．赤いハイライト部分のテキストは議論に良い影響を与えた議論内容を示し，青いハイライト部分のテキストは議論に悪影響を及ぼした議論内容を示している．さらに，ハイライトの明度は議論の質に与えた影響の大きさを示し，明度が低いほど議論品質に与えた影響が大きくなっている．また，テキストの上側にあるグラフは議論内容全体における議論品質への影響を示しており，赤と青の部分はそれぞれ，議論内容全体における良い内容と悪い内容の比を表している．

本分析内容により，議論における評価点や改善点が明確になり，議論参加者に対して適切なフィードバックを行うことが可能になる．また，音声認識を活用することにより，本分析を逐次的に行うことができ，即時的なフィードバックを提供することが可能となる．したがって，複数人議論の分析を逐次的に行えるという点で，音声認識結果は議論評価システムの入力として有用であることがいえる．

6.　おわりに

本論文では，まず議論評価モデルの精度向上を目的として，言語以外の情報の利用と不均衡データへの対策という 2 つのアプローチを提案し，その有効性を調査した．1 つ目の言語以外の情報の利用では，議論中の発言内容に加えて，動作や音声の情報を利用するモデルや，議論参加者の感情情報を利用するモデルを構築し，議論の品質評価を行った．結果として，提案した 2 つの手法は，いずれもベースラインに比べ，精度の向上は見られなかった．異なる情報を表現するベクトルを単純に結合しただけではモデルの学習は困難であるといえる．また，2 つ目の不均衡データへの対

策として，LLM による議論評価や少量データの拡張，損失関数の変更を行った．LLM による議論評価では，低頻度クラスにおける評価精度が向上した一方で，高頻度クラスの評価精度は低下する結果となった．少量データの拡張では，GPT-4 を用いて低頻度クラスの疑似データを生成し，学習データの拡張を行った．結果として，低頻度クラスの精度向上，全クラスにおける精度向上のどちらも実現できなかった．損失関数の変更では，不均衡データの学習に有効とされる損失関数を期待される導入した．しかし，本手法においても低頻度クラスの精度向上，全クラスにおける精度向上のどちらも確認できなかった．

さらに，議論の即時的フィードバックを目的とし，音声認識結果を利用した議論の逐次的分析を行った．議論の逐次的分析では，音声認識結果を入力に用いる議論評価モデルに対して，SHAP を用いることにより，議論品質に良い影響を与えた内容と悪い影響を与えた内容を明確化した．

今後の課題として，議論評価モデルの精度向上のため，LLM と BERT を組み合わせて議論を評価する手法を導入することが挙げられる．低頻度クラスの議論評価を LLM，高頻度クラスの議論評価を BERT で行うことにより，低頻度クラスと高頻度クラスの両方の精度を向上できるようになると期待される [14]．また，複数人議論の分析内容を多様化させることも課題として挙げられる．本論文においては，議論の分析として，議論の品質に好影響を与えた内容と悪影響を与えた内容を明確化した．しかし，複数人議論においては，各参加者の発言数や議論の盛り上がり度などの情報もフィードバックに重要である．そのため，各参加者の発言数や議論の盛り上がり度などをリアルタイムで可視化できれば，評価者の負担をさらに軽減することができるようになると考えられる．さらに，近年では LLM によるフィードバック生成が注目されている．実際に，LLM を用いることにより，文章による質の高い小論文のフィードバックが行われている [15][16]．本論文で行った SHAP による議論分析に加えて，LLM によるフィードバック生成を行うことで，より良いフィードバックを実現できると考えられる．

謝辞

本研究は科研費 23K11368 の一部です．

参考文献

[1] 橋口駿亮, 嶋田和孝．複数モデルの統合とデータ拡充による議論評価．電気情報通信学会-信学技報, vol-123, no. 416, NLC2023-23, pp. 1-6, 2024.

[2] 李昂, 嶋田和孝．議論評価システムの入力における音声認識の利用．電子情報通信回九州支部，第 32 回学生会講演会，D-33, 2024.

[3] 武川直樹, 中山知大, 徳永弘子, 大和淳司, 山下直美．グループディスカッションにおける発言者の言語/非言語の表出と評価者評価の関係の分析．電子情報通信学会論文誌 D 情報・システム, Vol. J101-D, No. 2, pp. 284–293, 2018.

[4] Tsukasa Shiota and Kazutaka Shimada. Annotation and multi-modal methods for quality assessment of multi-party discussion. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pp. 175–182, 2022.

[5] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 176–187, 2017.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28492–28518, 2023.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, Vol. 9, No. 8, p. 1735–1780, 1997.

[10] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, Vol. 64, pp. 59–63, 2015.

[11] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report, 2024.

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[13] Scott M Lundberg and Su-ln Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 2017.

[14] 竹尾匡貴, 嶋田和孝．教師有り学習モデルと大規模言語モデルを組み合わせた低評価レビューを考慮したレビュー文書の評価値推定．言語処理学会 第 31 回年次大会発表論文集，P9-17, 2025.

[15] Sayaka Nakamoto, Yoshi Okamoto, Takashi Nakakouchi, and Kazutaka Shimada. Towards human-level evaluation: Assessing the potential of gpt-4 in automated evaluation and feedback generation on japanese essays. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 156–161, 2024.

[16] 中本さや香, 嶋田和孝．日本語小論文に対する LLM を用いた建設的フィードバックの生成と分析．電子情報通信学会-信学技報, 言語理解とコミュニケーション研究会, 2025-03-NLC-NL, 2025.