スマートグラスを想定した講義補助を目的とした板書文字認識

金　承憲　竹内　孔一
岡山大学大学院環境生命自然科学研究科
pacf0axm@s.okayama-u.ac.jp
takeuc-k@okayama-u.ac.jp

概要

スマートグラスによる講義支援では，黒板・スライドの OCR 結果を翻訳・要約などの下流処理へ接続するが，日本語文と数式が混在する講義資料では OCR 誤りが多く，下流処理の安定性を損なう．本研究は準リアルタイム利用を想定し，OCR 結果に講師音声（ASR）を文脈ヒントとして付与して LLM で後処理するパイプラインを提案する．提案法は信頼度に基づくルーティングと，数式／混在テキスト向けの分岐，phase 別音声ヒント，出力制限ポリシを統合する．理工系日本語講義の静止画 21 枚で評価した結果，PaddleOCR 単独に対して最大 32.6%の CER 改善，41.3%の WER 改善，83.2%の BLEU-1 上昇を確認し，残る失敗要因も分析した．

1　はじめに

近年の大規模言語モデル（LLM）の発達により，携帯機器上でも OCR/AR を介して実世界の文字情報を翻訳・要約などへ接続する応用が現実的になりつつある．中でもスマートグラスは視界内提示とハンズフリー性により学習支援への応用が期待される [1]．一方，教育分野の LLM 活用はオンデマンド用途が中心で，講義進行に追従する準リアルタイム支援は計算資源・遅延・誤情報リスクの点で依然難しい [2, 3]．このため講義場面では，下流処理の安定性を左右する一次 OCR 誤りを抑える ***OCR* 後処理**（*post-correction*）の堅牢化が重要となる．

本研究は，スマートグラスを想定した教育支援アプリケーション開発の一要素として，一次 OCR テキストを LLM で補正することに焦点を当てる．本研究の特徴は次の 2 点である．第一に，講師音声（ASR 書き起こし）を**文脈ヒント**として LLM へ提示し，専門用語・表記ゆれ・誤分割の修正を促す．第二に，講義資料に多い**数式混在**を考慮し，文と数式を同一方針で扱わず，軽量な判定に基づいて処理経路を切り替えることで，LLM の効率と精度の両立を図る．

本稿の貢献は次の 3 点である．

- OCR 信頼度と形式判定を組み合わせたルーティング（pass/text/image/math_image/mixed_math）を設計し，準リアルタイム利用で問題となる計算資源を難例に集中させる枠組みを示す．
- 講義音声を phase 単位に分割し，対象静止画に対応する区間ヒントを与えることで，LLM 後処理に必要な文脈を**局所化**して提示する手法を提案する．
- 21 枚の講義静止画で評価し，BLEU-1/CER/WER の改善とともに，失敗パターン（低 OCR 精度・mixed_math の誤判定）を分析する．

本稿では OCR 後処理に焦点を当てるが，将来的には本成果を利用して，(1) 数式の逐次説明，(2) 用語の語釈提示，(3) 音声と板書の対応付け，など多様な講義支援機能へ拡張する予定である．

2　関連研究

OCR 後処理は，辞書・言語モデル・編集距離に基づく訂正から，深層学習による系列変換型の訂正へと発展してきた．従来は，OCR 誤りを文字列変換として扱い，(i) 誤りの尤度推定，(ii) 候補生成，(iii) 文脈に基づく再順位付け，という枠組みで研究されることが多かった．近年では Transformer を用いた誤り訂正や，マルチモーダル入力による訂正が提案されている [4]．

一方，LLM の登場により，OCR 出力を自然言語として「校正」し直すアプローチが現実的になった．LLM は追加学習なしに多様なドメインへ適用できる利点があるが，幻覚や過剰補完，すなわち入

力に存在しない情報を生成してしまう危険性が指摘される [7, 8]．したがって，教育現場における LLM 利用では，**出力の制約とフォールバックを設計する**ことが重要となる [5, 6]．

本研究は，**音声ヒントを利用した *LLM* 後処理と文／数式の形式に応じたルーティング**を統合し，スマートグラスでの準リアルタイム利用を意識した「軽量処理（pass/text）＋必要時のみ高コスト（image/math）」という設計方針を具体化する点に特徴がある．

3　提案手法

図 1 に，本研究で提案する OCR 後処理パイプラインの全体像を示す．本パイプラインは，(i) 講義静止画から得た一次 OCR 結果と，(ii) 講義音声の ASR 書き起こし（音声ヒント）を同時に入力として与え，MLLM/LLM によって OCR 誤りを補正することを目的とする．

まず入力として，講義動画から抽出した静止画（黒板の画面）を OCR にかけ，図 1 左側に示すように，一次 OCR テキストと信頼度（confidence）を得る．この信頼度は後述のルーティングの判定に用いられる．

次に，講義音声から得た ASR 書き起こしを音声ヒントとして用意する．音声ヒントは，該当区間に対応する語彙や説明文脈を含み得るため，OCR だけでは曖昧になりやすい専門用語・表記ゆれの補正を支援する．また，本研究では Microsoft の Clipchamp が提供する自動生成字幕を用い，書き起こされたテキストデータを音声ヒントとして利用することで，疑似的なリアルタイムでの音声入力環境を構築している．

最後に，一次 OCR 結果と音声ヒントを MLLM/LLM へ入力し，図 1 右側に示すように補正結果を得る．ここで 1 次 OCR で得られたテキストの信頼度，および形式により，機械的に 5 つのルートに分けられて LLM の処理が行われる．

最終的に，出力は `corrected_text`（補正後テキスト）に加え，どの処理経路で生成されたかを示す `route` を併せて記録する．

3.1　phase 別音声ヒント生成

講義は時間的に進行し，扱う用語や数式は区間ごとに偏る．そこで講義全体を phase 別に分割し，対象静止画が属する phase に応じて音声ヒントを付与

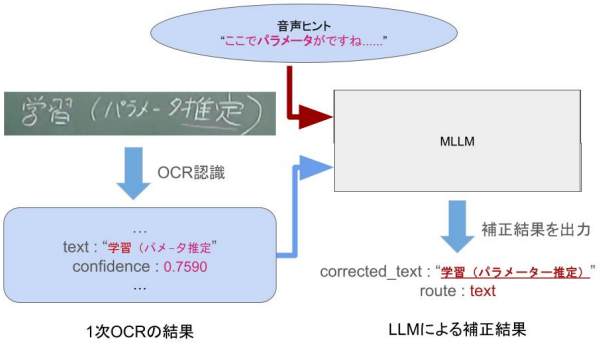

図 1　LLM による後補正の流れ

する．本実験では phase は手動で付与したが，将来的にはスライド遷移や話題境界検出による自動分割も可能である．

音声ヒントは次の 2 種類を用意する．

- **全体ヒント（global）**：対象 phase 内の ASR 文を集計し，頻出語彙 Top-K を抽出する．専門用語・固有表現の表記ゆれ抑制に用いる．
- **局所ヒント（local）**：時刻 t 近傍の ASR 文を連結し，当該画面に対応する説明文脈（関連語彙・導入語）を与える．

ただし，音声ヒントを「正解」あるいは「正解の候補」として考慮させるのではなく，OCR の曖昧なところに対する候補の絞り込みに限定して利用する点である．この方針は 3.3 のプロンプト制限として明示し，過剰補完を抑制する．

3.2　OCR 信頼度に基づくルーティング

OCR 行信頼度を $c_i \in [0, 1]$ とし，サンプル全体の信頼度を

$$\mathrm{conf}(X) = \min_i c_i$$

で定義する．最小値を用いるのは，「一部の行のみが著しく崩れる」講義 OCR の典型的な失敗を検出し，難例に高コスト処理を確実に割り当てるためである．LLM/MLLM の処理タスクは閾値 $\tau_{\mathrm{high}}, \tau_{\mathrm{mid}}$ により次の段階的ルートへ分岐する．

- **pass ルート**：$\mathrm{conf}(X) \geq \tau_{\mathrm{high}}$ なら OCR をそのまま採用し，LLM 呼び出しを回避する．
- **text ルート**：$\tau_{\mathrm{mid}} \leq \mathrm{conf}(X) < \tau_{\mathrm{high}}$ ならテキストのみを入力として LLM 補正を行う．
- **image ルート**：$\mathrm{conf}(X) < \tau_{\mathrm{mid}}$ なら画像も入力し，視覚情報を含めて補正する（高コスト）．

text ルートではテキスト入力のみの LLM 推論を行い，image/math_image ルートでは画像を併用するマ

ルチモーダル推論を行う．スマートグラス運用では電力・ネットワーク・遅延が制約となるため，難例にのみ高コスト処理を投入することが重要である．本ルーティングはこの要請に沿って推論コストを制御する枠組みを与える．

3.2.1 math_image ルートの数式判定

上記の pass ルート，text ルート，image ルートに加え，さらに math_image ルートを設ける．講義資料では，数式のみからなる行が頻出する．一般的に，別途の学習やプロンプト指示なしで LLM に数式を認識させることは困難であり，特に，同一画面に一般文と数式が混在する場合は性能が大きく低下する．

そこで各行が数式であるか否かを軽量に判定し，数式のみの行は **math_image ルート**へ送る．**math_image** では「単一行 LaTeX（$...$）のみを出力」と強く制約し，説明文や余計なトークン生成を禁止することで，数式構造の保存を優先する．

判定には（例として）(i) 演算子や等号の出現，(ii) 英数字と記号の比率，(iii) 括弧のパターン，(iv) ひらがな連続比率の低さ，(v) 添字風パターン（英字＋数字）などを用いる．これらの特徴量から単純なスコアを計算し，所定の閾値を超えた場合に数式であると判定する．本稿では詳細な閾値の設定，および判定基準は省略する．

3.2.2 mixed_math ルートの判定

一方，文と数式が同一行または同一スライド内に混在する場合，全体を数式として扱うと日本語文が不自然になり，逆に文として扱うと数式が崩れることが多く見られる．そこでサンプル単位で「混在」の兆候（数式行と日本語行の混在，あるいは数式記号の局所的集中）を検出した場合，**mixed_math ルート**として処理する．

mixed_math では，まず text/image ルートで全体の自然言語校正を得る．次に行単位で再度，数式行の判定を行う．数式行のみを math_image で再生成して置換することで，全体の画像を LLM に入力する費用を節約できる．最後に OCR の行順を保持しながら結合（merge）する．ここでは「どの数式がどの行に対応するか」という行対応付けが重要であり，text/image 側で改善された日本語行を壊さず，数式行のみを局所的に置換する点に狙いがある．

また，OCR が一つの数式を複数行へ誤分割した場合，LLM が同一式を複数回出力することがあるため，merge 時には近似一致（文字 n-gram 類似度など）などの簡易フィルタを用いて重複候補を抑制する．

3.3 プロンプトと制限ポリシ

LLM による OCR 後処理では，入力に含まれない語を補完してしまう「過剰補完」や，音声ヒントの誤りをそのまま転記してしまう危険がある．そこで本研究では，プロンプトに以下を明示して生成を制約する：(1) ヒントは参照のみで本文にコピーしない，(2) OCR に存在しない情報を追加しない，(3) 言い換え・要約をせず最小修正に徹する，(4) 行構造を保持する，(5) 数式ルートでは$...$のみを出力し，説明文を禁止する．さらに出力が不自然に長い／空である／ヒントへの依存が過度な場合は補正を棄却し，OCR または text/image 結果を保持するフォールバックを設けた．

4 実験

4.1 実験環境

実験データは，パターン学習に関する約 1 時間の日本語講義動画，および当動画から抽出した静止画 21 枚である．この静止画はリアルタイムで「スマートグラスユーザが選択した画面」を想定したものである．各静止画には，抽出した時点での時間が記録されたタイムスタンプが付与される．動画から音声を抽出し ASR でテキスト化したのち，講義進行に沿って phase1〜4 に分割した．LLM 推論時には，各静止画が属する phase に対応する音声ヒント（global/local）を与える．

OCR は PaddleOCR（padOCR）[10]，LLM は gemini-3-flash を使用した．いずれも本タスク向けの追加学習は行っていない．また，閾値は開発データで経験的に調整し，本実験では $\tau\ high = 0.87$, $\tau\ mid = 0.71$ を用いた。

結果では，(i) 一次 OCR のみのテキスト（pad_text）と，(ii) 提案パイプラインで補正した結果テキスト（gemini_text）を比較して，補正によって改善された程度の評価を行う．

4.2 評価指標

BLEU-1（char）は文書レベルでの文字ユニグラム精度を測る．講義 OCR は短い用語や記号が多く，単語分割に依存しない文字ベース評価が有用であ

表 1　Macro スコア（21 タスクの平均）

手法	BLEU-1↑	CER↓	WER(morph)↓
pad_text	0.589526	0.446191	0.617739
gemini_text（提案）	0.832889	0.311364	0.362665

表 2　Global スコア（21 タスクの平均）

手法	BLEU-1↑	CER↓	WER(morph)↓
pad_text	0.472801	0.553753	0.671642
gemini_text（提案）	0.866290	0.373225	0.402985

る．**CER（char）**は編集距離に基づく文字誤り率であり，OCR 後処理の基本指標として用いる．

一方，講義支援では「文として意味が通るか」「用語が正しいか」も重要であるため，**WER** も導入する．本研究では次の 2 種類を区別した．

- **WER(line)**：行単位（改行単位）での一致を評価する．これは**構造**（行分割，箇条書き等）の保全度を点検する指標である．
- **WER(morph)**：形態素単位（mecab 等）での誤り率を評価する．これは**内容**（語彙・助詞・用語）の一致度を測る．

本研究では両者を併用し，構造と内容を分けて分析する．（表の幅制約により，本稿では WER(line) は本文数値掲載を省略し，補足として提示している）．

4.3　実験結果

表 1 と表 2 に結果を示す．Macro は各サンプルの平均，Global は全文連結での集計である．提案法は BLEU-1 を大きく改善し，CER と WER(morph) も一貫して低下した．

全体として Global 集計のほうが改善幅が大きく，最大で 32.6%の CER 改善，41.3%の WER 改善，83.2%の BLEU-1 上昇を確認した．一方，一部ケースでは改善幅が小さく，低精度 OCR がボトルネックになる傾向が見られた．このような傾向については考察で詳細を述べる．また 21 タスクの処理時間は平均で 7 分 22 秒であり，1 枚あたり約 21 秒を要した．ルーティング内訳については，`pass` ルート 2 件，`text` ルート 6 件，`image` ルート 4 件，**math_image**3 件，`mixed_math` ルート 6 件である．

以下では，結果について 2 つの考察点を挙げている．考察に関連する具体例は，別途の資料にて提示する．

4.4　考察：math ルートの効果

数式を含むケースでは，BLEU-1 スコアが平均 28%（pad_text）から約 92%（gemini_text）までと，約 64%の向上を見せ，math_image ルートによって記号・添字の崩れが抑制され，改善が顕著であった．これは，数式出力を$...$に限定し，説明文生成を禁止することで，LLM の過剰生成を抑えられたためと考えられる．特に，math ルートの分岐により，LLM に画像とテキストを渡して推論させる高性能処理を，数式のような高難度のタスクに集中させることができたことに，本実験の意義がある．今後の課題としては，判定特徴の追加（ひらがな比率，等号周辺のパターン，括弧対応の安定化等）や，bbox 単位の数式領域検出との統合がある．

4.5　考察：ルーティングエラーと mixed_math

一方，mixed_math では，BLEU-1 スコアが平均 49%（pad_text）から約 66%（gemini_text）までと，約 17%p のわずかな向上を見せた．

この問題は，math_image 自体の能力というより，入力側の行対応付けが崩れたことに起因する．例えば，1 つの数式が描かれた黒板の画像に対して，OCR が「3 つの文」と認識した場合，LLM が数式（1 文）を復元したとしても，どの分に当てはまるかは判断不可能となる．さらに，mixed_math では，LLM が数式を完璧に補正する一方で，一般文を歪曲してしまうケースも見られた．この問題の解決のために，数式領域検出に基づく領域単位処理や，行再整列の導入が考えられる．

5　まとめ

本研究はスマートグラスを想定した教育支援 OCR アプリケーション開発の一環として，講義 OCR の一次結果を音声ヒントで補強し，LLM で後処理するパイプラインを提案した．文／数式の形式に応じたルーティング（pass/text/image/math_image/mixed_math）と，phase 別音声ヒント付与により，追加学習なしでも低コストで文字認識精度を大きく改善できることを示した．

今後は，(1) 数式領域検出と連携した bbox 単位処理，(2) mixed_math での alignment 強化，(3) 最適な OCR メソッドの再検討を進める．これにより，数式説明・用語解説・音声と板書の対応付けなど，講義中の多様な支援機能へ拡張し，スマートグラスを通じた日本語教育支援への貢献を目指す．

参考文献

[1] Dawon Kim and Yosoon Choi. Applications of Smart Glasses in Applied Sciences: A Systematic Review. *Applied Sciences*, 11(11):4956, 2021.

[2] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. Large Language Models for Education: A Survey and Outlook. arXiv:2403.18105, 2024.

[3] Y. Wu et al. Enabling Interactive Education with Low-Latency Large Language Models. *Concurrency and Computation: Practice and Experience*, 2025.

[4] Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

[5] James Zhang, Wouter Haverals, Mary Naydan, and Brian W. Kernighan. Post-OCR Correction with OpenAI's GPT Models on Challenging English Prosody Texts. In *Proceedings of the ACM Symposium on Document Engineering (DocEng' 24)*, 2024.

[6] Emanuel Boroş et al. Post-correction of Historical Text Transcripts with Large Language Models. In *Proceedings of the LaTeCH-CLfL Workshops*, 2024.

[7] OpenAI. Why Language Models Hallucinate. Technical report, 2025.

[8] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of ACL*, 2020.

[9] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. Image-to-Markup Generation with Coarse-to-Fine Attention. In *Proceedings of ICML*, 2017.

[10] Yuning Du et al. PP-OCR: A Practical Ultra Lightweight OCR System. arXiv:2009.09941, 2020.