講義スライドのコンテンツ解析に基づく講義動画の自動生成に関する予備検討

水上 皓太[1]　出口 大輔[1]　久徳 遙矢[2]　峰松 翼[3]　村瀬 洋[1]

概要：オンライン教育の普及に伴って講義スライドを講師が読み上げる形式の講義動画の需要が高まっている．しかし，講義動画の作成には労力やコストの面で大きな負担がかかる．そこで本研究では，講義スライドを入力として音声やハイライト表示などを含んだ講義動画を自動生成する手法を提案する．具体的には，まず講義スライドのコンテンツを大規模言語モデルや光学文字認識を用いて解析する．そして，その情報を用いてテキスト読み上げ音声を生成し，講義スライド中の重要なテキストをハイライトすることによって学修者の理解を促す講義動画を生成する手法を提案する．さらに，実際のスライドを用いた実験によって，提案手法が講義動画を自動生成できることを確認した結果について報告する．

1. はじめに

近年，オンライン教育やハイブリッド授業の普及に伴い，映像教材の需要が高まっている．特に，大学や企業の研修現場において，講義スライドを講師が読み上げる形式の動画（以下，講義動画）を用いた授業やセミナーが行われている．また，言語情報と視覚情報を組み合わせることで，学修者の理解度が向上することが示されており [1]，講義動画の利用によって教育の質を向上できる可能性がある．しかし，講義動画を制作するには，教員や制作者が長い時間をかけて録画や編集を行う必要があり，労力やコストの面で大きな負担となる．こうした状況を踏まえ本研究では，スライド中の文章や図表などのコンテンツを自然言語処理や画像解析などを用いて抽出・分析し，学修者の理解を促すような視覚効果を加えたスライドショー映像と，スライドの内容を説明する音声を組み合わせた講義動画を自動生成する手法を提案する．このような自動生成技術が実現すれば，従来は手作業で行っていたスライド説明の録画・編集作業を自動化することが可能となり，教育現場や企業研修などさまざまな場面において貢献が期待できる．

本発表では，講義動画の自動生成にあたり必要となる，スライドの内容を説明する文章（以下，読み上げ文）の自動生成，その文章を読み上げた音声（以下，読み上げ音声）の自動生成，そして読み上げ文に対応するスライド中テキストのハイライト表示方法について検討した結果について報告する．そのために，まず 2 章で関連する先行研究や既存技術の動向を概観し，3 章では本研究で提案する自動生成手法について示す．そして，4 章で提案手法の有効性を確認するための予備的な実験を行い，5 章でその考察を述べる．最後に，6 章でむすびと今後の展望について述べる．

[1] 名古屋大学
Nagoya University
[2] 愛知工科大学
Aichi University of Technology
[3] 九州大学
Kyushu University

2. 関連研究

2.1 と 2.2 では，コンテンツ解析に用いる光学文字認識とテキスト埋め込みについて，2.3 と 2.4 では，読み上げ文と読み上げ音声の生成に用いる大規模言語モデルとテキスト読み上げについて，2.5 では，AI で生成された教育ビデオに関する先行研究について説明する．

2.1 光学文字認識

光学文字認識（Optical Character Recognition，OCR）は，印刷物や手書き文書，画像データに含まれる文字や記号を，コンピュータが読み取ってテキストデータに変換する技術である．テンプレートマッチングやパターン認識等の古典的な画像処理を用いた手法が主流であったが，近年は CRAFT[2] のように，機械学習や深層学習を取り入れることで，精度を向上させている．

2.2 テキスト埋め込み

テキスト埋め込みは，ニューラルネットワーク等を用いてテキストを特徴空間にマッピングし，特徴ベクトルを得る手法である．これにより，テキスト同士の意味的な近さ

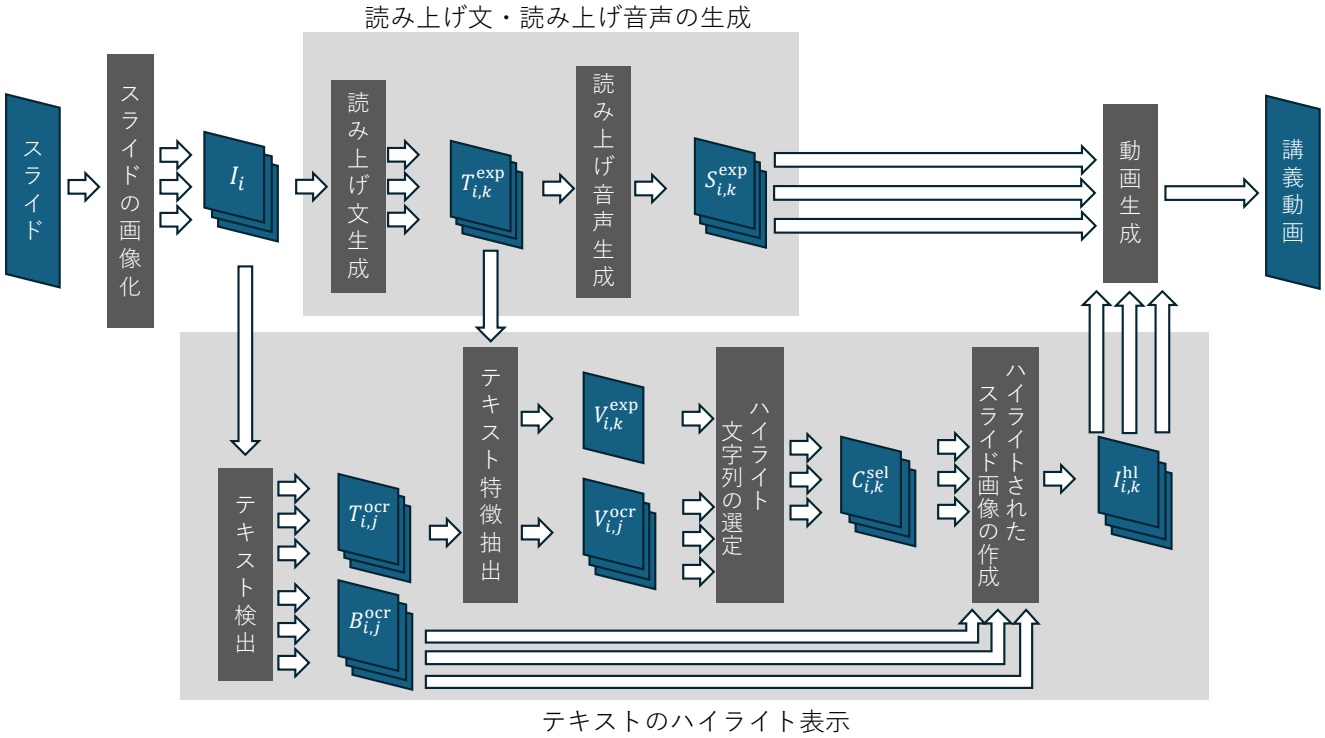

図 1 提案手法による講義動画作成の処理フロー

などをベクトルの計算によって求めることが可能である．手法の一つとして，Sentence-BERT[3] が挙げられる．

2.3 大規模言語モデル

大規模言語モデル（Large Language Models，LLM）は，大量のテキストデータを用いて学習した大規模なニューラルネットワークによって，文章生成，翻訳，要約，質問応答など多様な自然言語処理タスクに対応できるモデルである．近年はマルチモーダル化が進んでおり，ChatGPT[4] 等では画像とテキストを統合した処理が可能になっている．これらのモデルは視覚言語モデル（Vision-Language Models，VLM）とも呼ばれる．

2.4 テキスト読み上げ

テキスト読み上げ（Text-To-Speech，TTS）は，入力されたテキストをコンピュータが音声として合成し出力する技術である．ルールベースの手法や統計的な手法，そして機械学習や深層学習を用いた手法などがある．広く使われるサービスとして，Google Text-To-Speech(gTTS)[5] がある．

2.5 AI 生成教育ビデオ

Xu らは，AI で生成された教育ビデオの学習者への影響を調査するために，OCR，LLM，TTS を組み合わせて，講師のアバターが自動音声で講義スライドの内容を話す動画を生成する連続的な処理を構築した [6]．その上で，AI で生成した講義動画と従来の講義の録画を被験者実験によって比較し，AI で生成した講義動画が英語学習における学習者の記憶力を向上させることを示唆した．しかし，スライド中の要素をハイライトする仕組みは取り入れられていない．

3. 提案手法

本研究では，スライドを入力として，読み上げ文と読み上げ音声の生成，テキストのハイライト表示，そして講義動画の自動生成を行う手法を提案する．全体の処理の流れを図 1 に示す．ここで，スライドは基本的に pptx 形式や pdf 形式のファイルであるが，データ構造等の問題から，これらの形式のデータからテキストの位置情報を取得することは難しい．そのため，提案手法ではスライドのそれぞれのページ全体を 1 枚の画像（以下，スライド画像）に変換し，OCR によるテキスト検出を行う．i ページ目のスライドから変換された画像を I_i として，以降の処理の入力とする．

読み上げ文・読み上げ音声の生成では，まず ChatGPT に特定のプロンプトとスライド画像を入力することで，読み上げ文のテキストを得る．その後，TTS 技術によって読み上げ文から読み上げ音声を得る．テキストのハイライト表示では，OCR 技術によってスライド画像から学修者が注目するべきテキストの候補を探す．その後，テキスト埋め込みを活用して読み上げ文と対応するテキストを選択し，ハイライト表示する．3.1，3.2，3.3 で，それぞれの処理の詳細について述べる．

> あなたは大学教授です。あなたはスライド資料を用いて学生に授業をします。その時にしゃべる内容を考えてください。ただし、以下の条件を満たしてください。
> 1. しゃべる内容のみを返す
> （略）
> 7. まず、全ページのスライド画像が与えられた後、各ページのスライド画像が入力として与えられていくので，スライドのつながりや内容の偏りに注意してしゃべる内容を考えてください

図 2 前提の確認と全体的なタスクの指示を行うプロンプト

> まず、スライド全体を入力するので、内容を要約してください。その後、これからしゃべる内容の戦略を立ててください。具体的には、しゃべる内容の偏りや冗長さを防いでください。

図 3 スライド全体を把握させるプロンプト

> これは i 番目のページです。このページを見ながら実際に発表してください。

図 4 各ページの読み上げ文を生成させるためのプロンプト

3.1 読み上げ文・読み上げ音声の生成

3.1.1 読み上げ文生成

スライド画像を ChatGPT に入力することで，講義動画における読み上げ文を生成する．このとき，ChatGPT には次のようなプロンプトを与える．

前提の確認と全体的なタスクの指示

まず，読み上げ文生成を行う上で重要となる情報や，タスク全体を把握させるための，図 2 のようなプロンプトを入力する．

スライド全体の把握

次に，全てのスライド画像を入力することで，スライド全体の流れをつかませ，自然な読み上げ文を生成するための戦略を立てさせる．図 3 はプロンプトの具体例である．

各ページの読み上げ文の生成

最後に，各ページのスライド画像を順に入力し，読み上げ文のテキストを生成する．図 4 はプロンプトの具体例である．

以上の処理により，スライドの i ページ目の画像 I_i から生成された読み上げ文のテキスト T_i^{exp} を得る．

3.1.2 読み上げ音声生成

3.1.1 で生成した読み上げ文のテキスト T_i^{exp} を句読点で区切り，得られた k 番目のテキストを $T_{i,k}^{\mathrm{exp}}$ とする．そしてこの $T_{i,k}^{\mathrm{exp}}$ を TTS 技術によって音声化し，音声 $S_{i,k}^{\mathrm{exp}}$ を得る．

3.2 テキストのハイライト表示

3.2.1 テキスト検出

スライド画像から OCR 技術によって，テキスト位置のバウンディングボックスとテキストの内容を検出する．I_i から j 番目に検出されたテキストのバウンディングボックスとテキストの内容を，それぞれ $B_{i,j}^{\mathrm{ocr}}$，$T_{i,j}^{\mathrm{ocr}}$ とする．

3.2.2 テキスト特徴抽出

$T_{i,k}^{\mathrm{exp}}$ と $T_{i,j}^{\mathrm{ocr}}$ を，Sentence-BERT[3] を用いてそれぞれ特徴空間に埋め込むことで，ベクトル $V_{i,k}^{\mathrm{exp}}$，$V_{i,j}^{\mathrm{ocr}}$ を得る．

3.2.3 ハイライト文字列の選定

全ての $V_{i,k}^{\mathrm{exp}}$ に対して，対応する i ページ目の $V_{i,j}^{\mathrm{ocr}}$ とのコサイン類似度 $C_{i,j,k}$ を計算する． その後，すべての $T_{i,k}^{\mathrm{exp}}$ に対して以下の操作を行う．

(1) 類似度の集合 $C_{i,k}^{\mathrm{all}} = \{C_{i,1,k}, \ldots, C_{i,j,k}, \ldots\}$ の中から，しきい値 θ 以上のものをすべて選び，集合 $C_{i,k}^{\theta}$ として抽出する．

(2) 集合 $C_{i,k}^{\mathrm{all}}$ の中から，類似度が大きい順に n^{sel} 個選び，集合 $C_{i,k}^{n^{\mathrm{sel}}}$ として抽出する．

(3) 和の集合 $C_{i,k}^{\mathrm{sel}} = C_{i,k}^{\theta} \cup C_{i,k}^{n^{\mathrm{sel}}}$ を得る．これにより，最低 n^{sel} 個のハイライト文字列を選定する．

3.2.4 ハイライトされたスライド画像の作成

$C_{i,k}^{\mathrm{sel}}$ の全ての要素 $C_{i,j,k}$ に対して，対応するバウンディングボックス $B_{i,j}^{\mathrm{ocr}}$ を画像 I_i 上に描画することによって，テキストを視覚的にハイライトする．$C_{i,k}^{\mathrm{sel}}$ の全要素に対応するバウンディングボックスが描画された画像を $I_{i,k}^{\mathrm{hl}}$ とする．

3.3 動画生成

音声 $S_{i,k}^{\mathrm{exp}}$ と画像 $I_{i,k}^{\mathrm{hl}}$ をすべて結合して動画を生成する．これにより，読み上げ音声とテキストのハイライト表示が組み合わさった講義動画を得る．

4. 実験

提案手法の有効性を確認するため，実際の研究発表スライドを用いて実験を行った．

4.1 実験条件

ハイライトするテキストの決定を行うしきい値 θ は 0.7 に設定した．数合わせによるハイライトを行うためのパラメータ n^{sel} は 1 に設定した．実験データとして，4 本の情報科学系の研究発表スライドを収集し，全 96 枚のスライド画像で実験を行った．

4.2 実験結果

実際に生成された動画の一部を図 5，図 6，図 7，図 8 に示す．ハイライトの色は，赤に近いほど類似度が高いことを示している．図 5 は，読み上げ文の生成とハイライトの生成がともに上手くいった例である．生成された読み上げ文は，「まず、BigColor を用いてモノクロ画像をカラー化します。」である．ハイライトされたテキストは，「BigColor」と「モノクロ画像を BigColor でカラー化し」の 2 つである．図 6，図 7，図 8 は，ハイライト生成が上手くいかな

かった例である．図 6 の読み上げ文は「ピクセル値の調査と被験者実験により、」であり，ハイライトされたテキストは「モノクロ画像から」と「画素値の変化の調査による評価と被験者実験により」の 2 つである．図 7 の読み上げ文は「MLP では、」であり，ハイライトされたテキストは「サンプリング」，「座標」，「視線方向」，「MLP」，「密度」，「ボリューム」，「レンダリング」，「画素値 $\hat{C}$」，「画素値の取得」，「二乗」，「L_{nerf}」，「画素値 C_{GT}」の 12 個である．図 8 の読み上げ文は「このページでは、」であり，ハイライトされたテキストは「画像の入力」，「Representing」，「レンダリング」の 3 つである．

5. 考察

図 5 において，読み上げ文とハイライトされたテキストは意味的に近いものであることが分かる．そのため，テキストが適切にハイライトされているといえる．図 6 においては，「画素値の変化の調査による評価と被験者実験により」がハイライトされるのは適切だが，「モノクロ画像から」がハイライトされるのは適切ではない．不適切なハイライトが生じた理由としては，「画素値」と「モノクロ画像」が 2 つとも画像処理分野でよく使われる用語であり，似たような特徴ベクトルが得られてしまったからだと考えられる．図 7 においては，読み上げ文とハイライトされたテキストに意味のつながりが無いように見える．そのため，ハイライトが不適切であるといえる．この不適切なハイライトが生じた理由としては，読み上げ文が短すぎることによって，どのような文章とも意味が近くなってしまうような特徴ベクトルが得られてしまったからだと考えられる．図 8 においては，読み上げ文がスライド内容とは直接関係のない「このページでは、」であり，ハイライトするべきテキストが無いにもかかわらず，3 つのテキストがハイライトされている．これは，図 7 の例と同様に，読み上げ文が短すぎることに起因していると考えられる．図 6，図 7 から観察できるこれらの問題は，ハイライト生成の処理が，テキストの類似度を個別に計算し，文脈を無視した単語レベルでの類似度のみを測る指標に基づいていることに起因していると考えられるため，文脈を加味した類似度計算手法を用いることで解決できる可能性がある．また，図 8 から観察できる問題に関しては，n^{sel} を動的に変化させる仕組みを取り入れることで解決できる可能性がある．

6. むすび

講義スライドを入力として，LLM や OCR によってコンテンツ解析した情報に基づく読み上げ音声の生成やテキストのハイライトを行い，講義動画を自動生成する手法を提案した，また，実際の研究発表スライドを用いた実験により提案手法で講義動画の自動生成が可能であることを確認

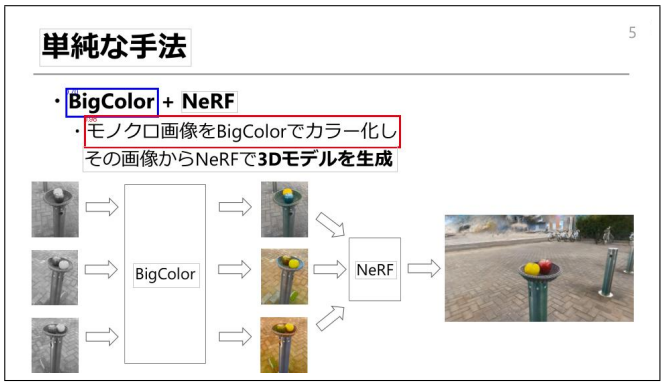

図 5 実験結果 1: 読み上げ文は「BigColor を用いてモノクロ画像をカラー化します。」

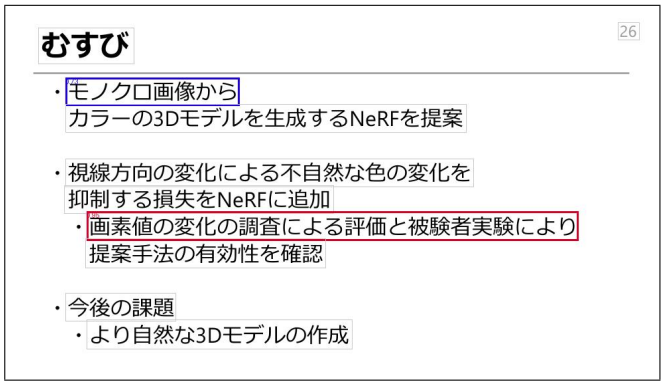

図 6 実験結果 2: 読み上げ文は「ピクセル値の調査と被験者実験により、」

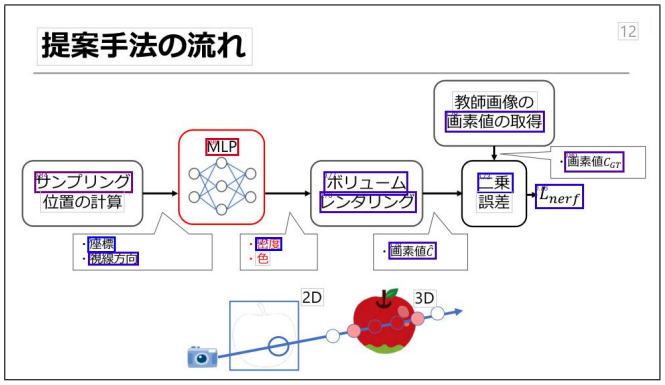

図 7 実験結果 3: 読み上げ文は「MLP では、」

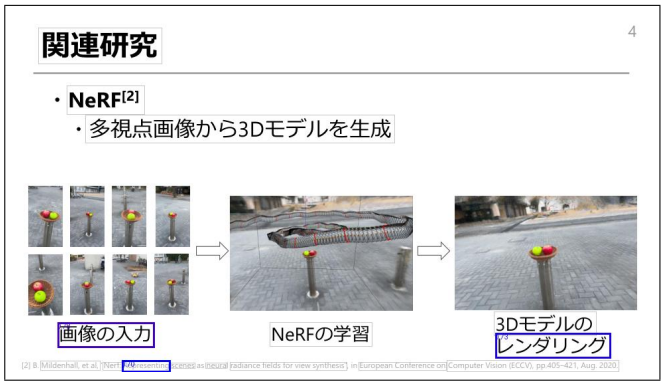

図 8 実験結果 4: 読み上げ文は「このページでは、」

した．今後の課題としては，ハイライト生成処理の改善がある．現時点ではテキストにのみ対応しているが，将来的には図表への対応を行いたい．また，読み上げ文の文脈や，スライドの全体を見られるようなモデルの構築も検討している．そのために，テキストや画像を統一的に処理でき，複数の特徴を用いた計算が可能なニューラルネットワークによる複雑な類似度計算の手法が求められる．

謝辞 本研究の一部は，CREST，JPMJCR22D1 の支援を受けたものである．

参考文献

[1] Mayer, R. E. and Anderson, R. B.: Animations need narrations: An experimental test of a dual-coding hypothesis, *Journal of Educational Psychology* (1991).
[2] Baek, Y., Lee, B., Han, D., Yun, S. and Lee, H.: Character Region Awareness for Text Detection, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).
[3] Reimers, N. and Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics (2019).
[4] OpenAI: ChatGPT: Optimizing Language Models for Dialogue (online), got at $\langle https://openai.com/blog/chatgpt \rangle$ (2022). Accessed 2025-02-16.
[5] Google: Google Cloud Text-to-Speech (online), got at $\langle https://cloud.google.com/text-to-speech \rangle$. Accessed 2025-02-16.
[6] Xu, T., Liu, Y., Jin, Y., Qu, Y., Bai, J., Zhang, W. and Zhou, Y.: From recorded to AI-generated instructional videos: A comparison of learning performance and experience, *British Journal of Educational Technology* (2024).