

講義スライドのコンテンツ解析に基づく ハイライト表示付き講義動画の自動生成に関する検討

水上皓太^{1,a)} 出口大輔^{1,b)} 久徳遙矢² 峰松翼³ 村瀬洋^{1,c)}

概要

オンライン教育の拡大により、講師が講義スライドを解説する形式の動画コンテンツへのニーズが高まっている。一方で、このような講義動画の制作には多大な労力とコストがかかるという課題がある。そこで本発表では、講義スライドを入力とし、音声やハイライト表示を含む講義動画を自動的に生成する手法を提案する。具体的には、講義スライドの内容を大規模言語モデルおよび光学文字認識技術によって解析し、得られた情報から音声合成による読み上げ音声を生成する。さらに、読み上げ音声に対応するスライド内のテキストをハイライト表示することで、学修者の理解を支援することを目指す。加えて、実際の講義スライドを用いた実験を通じて、本手法が自動的に講義動画を生成できることを検証し、その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広がりにより、映像教材への需要が一層高まっている。中でも、大学や企業の研修現場では、講義スライドを講師が解説する形式の動画（以下、講義動画）を用いた授業やセミナーの導入が進んでいる。さらに、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を促進できることが示されており [3]、講義動画の活用は教育の質の向上にも寄与すると期待される。

しかしながら、講義動画の制作には、教員や制作者が長時間をかけて録画・編集を行う必要があり、その作業負担やコストは無視できない。

こうした背景のもと、本研究では、スライド中の文章や図表といったコンテンツを自然言語処理および画像解析を通じて抽出・分析し、理解促進のための視覚的効果を付加したスライドショー映像と、スライド内容を説明する音声

を組み合わせることで、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確立されれば、従来手作業で行われていたスライド説明の録画・編集工程を自動化でき、教育現場や企業の研修など、多様な場面での活用が見込まれる。

本発表では、講義動画の自動生成に必要な三つの要素、すなわちスライド内容を解説する文章（以下、読み上げ文）の自動生成、その読み上げ音声の生成、および読み上げ文と対応するスライド上のテキストに対するハイライト表示について検討した結果を報告する。

まず 2 節において関連する先行研究と既存技術の動向を整理し、続く 3 節では提案手法の詳細を述べる。4 節では提案手法の有効性を評価するために実施した予備実験の内容と結果を示し、5 節にてその考察を行う。最後に、6 節において本発表のまとめと今後の展望を述べる。

2. 関連研究

2.1 画像解析

光学文字認識（Optical Character Recognition, OCR）は、画像中の文字や記号を、機械が処理可能な文字列データへと変換する技術である。代表的な手法としては、文字領域の検出と認識を高精度に行う CRAFT がある [1]。

また、テキスト埋め込みは、ニューラルネットワークなどを用いてテキストを特徴空間に射影し、その意味内容を表現する特徴ベクトルを得る手法である。このベクトル表現により、テキスト間の意味的類似度をベクトル演算によって定量的に評価することが可能となる。代表的な手法の一つとして、Sentence-BERT が挙げられる [5]。

2.2 テキストおよび音声の生成

大規模言語モデル（Large Language Models, LLM）は、大量のテキストコーパスを用いて事前学習された大規模なニューラルネットワークであり、テキスト生成や要約、翻訳など、幅広い自然言語処理タスクに対応可能である。代表的な LLM の利用例として、対話型 AI サービスである ChatGPT がある [4]。

¹ 名古屋大学

² 愛知工科大学

³ 九州工業大学

a) mizukamik@vislab.is.i.nagoya-u.ac.jp

b) ddeguchi@nagoya-u.jp

c) murase@nagoya-u.jp

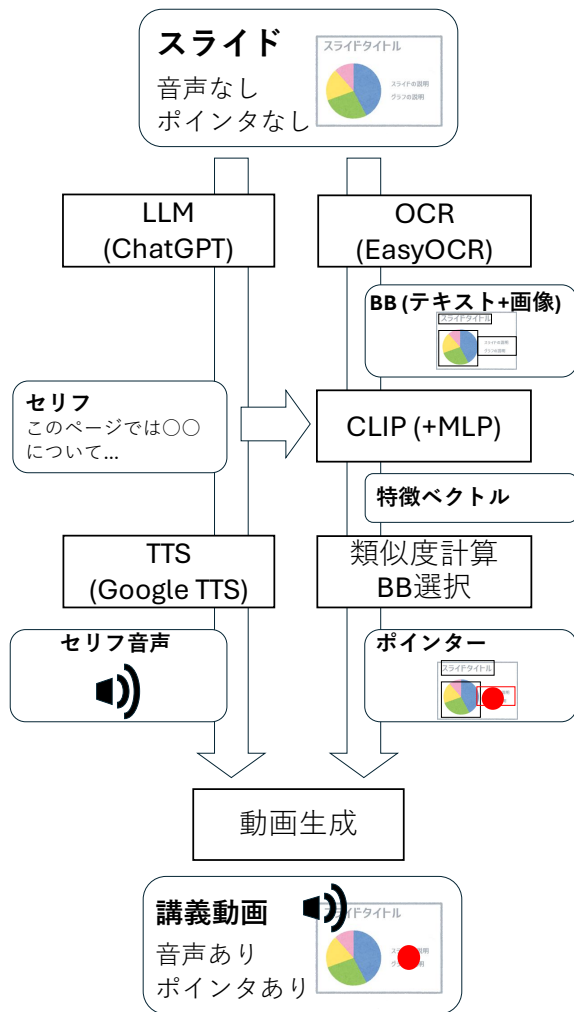


図 1 提案手法による講義動画作成の処理フロー

テキスト読み上げ (Text-to-Speech, TTS) は、入力されたテキストから自然な音声を合成する技術である。広く使われるサービスとして、Google Text-To-Speech (gTTS)[2] がある。

2.3 講義動画の自動生成

Xu ら (2024) は、OCR, LLM, および TTS を統合したパイプラインを構築し、入力されたスライドからナレーション音声を生成し、それにリップシンクされたアバター映像を合成する講義動画自動生成手法を提案した [6]。被験者実験においては、自動生成された講義動画と従来型の講義動画とを比較し、英語学習において前者の方が学修者の記憶定着に寄与する可能性があることが示された。ただし、この研究ではレーザーポインターなどによってスライド要素をハイライトする機能は扱われていないという限界がある。

3. 提案手法

本発表では、スライドを入力として、読み上げ文および読み上げ音声の生成、テキストのハイライト表示、さらに講義動画の自動生成を行う手法を提案する。全体の処理フローを図 1 に示す。

本手法では、スライドは基本的に pptx 形式または pdf 形式のファイルとして与えられるが、これらの形式から直接テキストの位置情報を取得することは、データ構造の不統一などの理由により困難である。そこで、各スライドページを 1 枚の画像（以下、スライド画像）に変換し、光学文字認識 (OCR) を用いてテキスト検出を行う。 i ページ目のスライドを変換した画像を I_i とし、これを以降の処理の入力とする。

読み上げ文および読み上げ音声の生成においては、まず ChatGPT に特定のプロンプトとスライド画像を与え、その内容に基づいた読み上げ文を生成する。続いて、生成された読み上げ文に対して TTS (Text-to-Speech) 技術を適用し、対応する音声を合成する。

テキストのハイライト表示では、OCR によって抽出されたスライド画像中のテキスト群から、学修者にとって重要と考えられる箇所を候補として抽出する。その後、テキスト埋め込み技術を用いて読み上げ文との意味的な類似度を計算し、対応関係のあるテキストを選択してハイライト表示を行う。

各処理の詳細については、3.1 で読み上げ文・音声の生成、3.2 でハイライト処理、3.3 で講義動画の生成についてそれぞれ述べる。

3.1 読み上げ文・読み上げ音声の生成

3.1.1 読み上げ文の生成

本手法においては、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自動生成する。このとき、生成される文が講義として自然なものとなるよう、ChatGPT には段階的にプロンプトを与える。各ステップの概要を以下に示す。

前提の確認およびタスクの概要提示 まず、読み上げ文生成の目的や背景など、タスク全体を理解させるために、図 2 に示すようなプロンプトを用いる。これにより、ChatGPT は今後与えられる入力に対して適切な出力を行う準備が整う。

スライド全体の構成把握 次に、全てのスライド画像を ChatGPT に入力することで、スライド全体の流れや文脈を把握させる。このステップにより、個々のスライドに対して生成される読み上げ文が、全体として一貫性を持つように調整される。具体的なプロンプトの例は図 3 に示す。

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

各スライドページに対応する読み上げ文の生成 最後に、各ページのスライド画像を個別に入力し、対応する読み上げ文を生成する。このステップでは、ページごとの内容に即した文が出力されるよう、図 4 に示すようなプロンプトを用いる。

以上の手順により、 i ページ目のスライド画像 I_i に対応する読み上げ文テキスト T_i^{exp} を得ることができる。

3.1.2 読み上げ音声の生成

3.1.1 で生成された読み上げ文テキスト T_i^{exp} を句読点で区切り、その k 番目の文を $T_{i,k}^{\text{exp}}$ とする。この各文に対して、TTS (Text-to-Speech) 技術を用いて音声を合成し、対応する読み上げ音声 $S_{i,k}^{\text{exp}}$ 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 I_i に対して OCR 技術を適用することで、画像中のテキストを検出し、それぞれの位置を示すバウンディングボックスとテキスト内容を抽出する。 j 番目に検出されたバウンディングボックスとテキストをそれぞれ $B_{i,j}^{\text{ocr}}$, $T_{i,j}^{\text{ocr}}$ と表す。

3.2.2 テキスト特徴の抽出

読み上げ文テキスト T_i^{exp} および検出テキスト $T_{i,j}^{\text{ocr}}$ に対して、Sentence-BERT を用いて埋め込み表現への変換処理を行い、それぞれの特徴ベクトル $V_{i,k}^{\text{exp}}$, $V_{i,j}^{\text{ocr}}$ を取得する。

3.2.3 ハイライト対象の選定

各 $V_{i,k}^{\text{exp}}$ に対し、同一スライド内のすべての $V_{i,j}^{\text{ocr}}$ とのコサイン類似度 $C_{i,j,k}$ を計算する。次に、各 $T_{i,k}^{\text{exp}}$ に対応するハイライト文字列の選定を、以下の手順で行う。

(1) 類似度集合 $C_{i,k}^{\text{all}} = C_{i,1,k}, \dots, C_{i,j,k}, \dots$ の中から、あらかじめ定めたいきい値 θ 以上の値をもつものを抽出

し、集合 $C_{i,k}^{\theta}$ を得る。

(2) 類似度 $C_{i,k}^{\text{all}}$ の中から、値が大きい順に n^{sel} 個を選び、集合 $C_{i,k}^{n^{\text{sel}}}$ とする。この操作は、しきい値 θ を超える要素が存在しない場合でも、最低限の数のハイライト対象を確保するために行うものであり、 n^{sel} はその最小選出数を指定するパラメータである。

(3) 最終的な選定集合として、 $C_{i,k}^{\text{sel}} = C_{i,k}^{\theta} \cup C_{i,k}^{n^{\text{sel}}}$ を構成する。これにより、最低限 n^{sel} 個のハイライト対象テキストが選定される。

3.2.4 ハイライトされたスライド画像の作成

$C_{i,k}^{\text{sel}}$ に含まれるすべての類似度要素 $C_{i,j,k}$ に対応するバウンディングボックス $B_{i,j}^{\text{ocr}}$ を、スライド画像 I_i 上に描画することで、該当テキストを視覚的にハイライトする。このようにして、読み上げ文 T_i^{exp} に対応するハイライト情報を付加したスライド画像を $I_{i,k}^{\text{hl}}$ と定義する。

3.3 動画生成

各スライドページにおいて生成された音声 $S_{i,k}^{\text{exp}}$ と、対応するハイライト画像 $I_{i,k}^{\text{hl}}$ を時系列に沿って結合し、講義動画を生成する。

この処理により、読み上げ音声と視覚的なハイライト表示が同期した講義動画が得られ、学修者の理解を支援する教育コンテンツとして利用可能となる。

4. 実験

提案手法が講義動画を生成できることを確認するため、実際の研究発表スライドを用いて実験を行った。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、最低限 1 件のハイライト対象を確保するため、パラメータ n^{sel} は 1 に設定した。

実際に生成された講義動画の 1 ページを図 5 に示す。左の画像では、入力されたスライド画像 I_i から検出されたバウンディングボックスが示されている。それぞれのバウンディングボックスには、説明のために番号を振ってある。右の表では、読み上げ文生成によって生成された読み上げ文 $T_{i,k}^{\text{exp}}$ と、 $T_{i,k}^{\text{exp}}$ に対応するバウンディングボックスが示されている。

5. 考察

図 5 に示す結果から、「モノクロ画像をカラー化し、」や「3D モデルを生成する単純な手法を説明します。」といった比較的長い読み上げ文に対しては、選択されたバウンディングボックスのテキストと読み上げ文との意味的な関連性が高く、適切なテキスト領域が選択されていることが確認できる。

一方、「まず、」や「その後、」のような短い読み上げ文では、選択されたテキストとの意味的関連が乏しく、不適切



生成された読み上げ文 ($T_{i,k}^{exp}$)	選択された バウンディング ボックス
このページでは、	①
モノクロ画像をカラー化し、	④
3Dモデルを生成する単純な手法を説明します。	①
まず、	③
BigColorを用いてモノクロ画像をカラー化します。	②, ④
その後、	③
カラー化された画像をNeRFで処理し、	⑤
最終的に3Dモデルを生成します。	⑤
この方法は工程が簡単で比較的直線的です。	①

図 5 提案手法で生成された動画の 1 ページ

なバウンディングボックスが選ばれていると考えられる。このような誤りが生じる要因として、以下の 2 点が挙げられる：

- (1) 読み上げ文が短すぎるため、どのような文章とも意味的に近くなってしまい、特徴ベクトル $V_{i,k}^{exp}$ が適切な意味を保持できないこと。
- (2) 対応するテキストがスライド上に存在しないにもかかわらず、いずれかのバウンディングボックスが強制的に選択されてしまっていること。

これらの問題は、ハイライト生成処理がテキスト間の意味的類似度を個別に計算しており、文脈を考慮しない単語レベルの比較に依存していることに起因すると考えられる。そのため、文脈情報を加味した類似度計算手法を導入することで、これらの課題が改善される可能性がある。

また、人間の発表者がポインタを通常「上から下」あるいは「左から右」へと自然に移動させるのに対し、本手法では移動順序に一貫性がなく、不自然な動きとなることがある。この点については、ポインタの移動順を全体として最適化する仕組みを導入することで、自然な動作が実現できると期待される。

6. むすび

本研究では、講義スライドを入力とし、OCR や大規模言語モデル (LLM) を用いたコンテンツ解析に基づいて、読み上げ音声の生成およびスライド中のテキストのハイライト表示を行うことで、講義動画を自動生成する手法を提

案した。さらに、実際の研究発表用スライドを用いた予備的な実験を通じて、本手法による講義動画の自動生成が有効であることを確認した。

今後の課題として、ハイライト生成処理の精度向上が挙げられる。現段階ではテキスト情報のみに対応しているが、図表やイラストなどの非テキスト要素に対しても適切なハイライトを行えるように拡張する必要がある。また、個々のスライドに対して独立にハイライト処理を行っている現行手法に対し、スライド全体の構成や読み上げ文の文脈を考慮できるようなモデル構築も今後の重要な方向性である。本研究の成果は、教育現場や企業研修における映像教材の効率的な生成に貢献すると期待される。

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] 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).

[2] Google: Google Cloud Text-to-Speech (online), got at <https://cloud.google.com/text-to-speech>. Accessed 2025-02-16.

[3] Mayer, R. E. and Anderson, R. B.: Animations need narrations: An experimental test of a dual-coding hypothesis, *Journal of Educational Psychology* (1991).

[4] OpenAI: ChatGPT: Optimizing Language Models for Dialogue (online), got at <https://openai.com/blog/chatgpt> (2022). Accessed

2025-02-16.

- [5] 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).
- [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).