Vol.2011-CLE-4 No.6
2011/5/13

OCW のための音声情報を用いた講義スライドの説明箇所の推定

松田恵理菜[†,††]　堀内靖雄[††]　黒岩眞吾[††]

OCW で配信される講義映像の形式の一つに教師映像と講義スライドを同期させる形式がある．この形式の問題点は教師が講義スライド中のどこを説明しているかが分かりにくいことである．本論文では教師の音声情報を用いて，講義スライド中の教師の説明箇所を推定する手法を提案する．8 講義に対する実験の結果，認識率は 49%であったが，人手による書き起こしを用いた場合は 76%となった．結果の考察から，認識率の改善について検討した．

Estimation of Explanation Spot in Lecture Slide Using Speech Information for OCW

Erina Matsuda[†,††]　Yasuo Horiuchi[††]
and Shingo Kuroiwa[††]

One of the styles of lecture video delivered in OCW is synchronizing the teacher image with the lecture slide. However, in this style, it is difficult to understand where the teacher is explaining in the lecture slide. In this paper, we propose the method of estimating explanation spot in lecture slide using teacher's speech information. We performed the experiment for 8 lectures in Chiba University. As a result, the estimation rate was 49% with only speech information. And using transcription of lecture, the rate was 76%. We discuss about improvement of the method by analysis of the results.

1. はじめに

近年，国内外の教育機関でインターネットを利用した e-learning が広く実施されている．中でも 2003 年に米マサチューセッツ工科大学が開始した OCW（OpenCourseWare）は大学や大学院など正規な教育機関で行われた講義を学ぶことができ，かつ当時有償のものが多かった e-learning としては珍しく無償であったことから注目され，現在も多く利用されている．

OCW とは大学の講義情報をインターネット上で配信する取り組みのことである．OCW で提供している講義情報には，シラバス，カレンダー，講義ノート，講義課題，定期試験と解答，講義映像，講義スライドなどがあり講義を受講している学生が復習・予習に使用する他，OCW 配信機関に所属していない人でも自由に OCW の配信情報を閲覧することが出来る．

図 1　そのまま配信する形式[a]

図 2　教師映像とスライドを並べる形式[b]

本研究では OCW の配信している講義情報のうち講義映像と講義スライドに注目する．OCW で配信する講義映像の形式は大学によって異なり，さらに同じ大学であっても期間によって異なる形式を採用している場合があるが，大別すると二つに分けられる．一つ目は講義をビデオカメラで録画した映像をそのまま編集せずに公開するものである（図 1）．この形式は教師の動きや授業全体の雰囲気はわかりやすいが，アングルが頻繁に変わるものは自分の見たい箇所が見られず，また教室全体を映すものは解像度の問題から黒板やスライド上の文字が見づらいという問題がある．二つ目に教師をアップで追尾した映像と教師の作成した講義スライドを同期させ並べて公開するものがある（図 1）．この形式は講義スライド映像が切り替わるタイミング以外では静止画であるために，受講者に臨場感を持たせにくく集中力を持続させることが難しい．また教師の映像と講義スライドが別に表示されていることから両者のつながりが見え

† 東京工業大学　Tokyo Institute of Technology（2011 年 4 月以降）
†† 千葉大学　Chiba University

a) http://academicearth.org/courses/justice-whats-the-right-thing-to-do, Justice,2011 年 2 月 14 日
b) http://videolectures.net/bootcamp2010_murray_iml/ , Introduction to Machine Learning, 2011 年 2 月 14 日

にくく,教師がスライド中のどこを説明しているかがわかりづらいという問題がある．各教育機関では OCW の形式の前者から後者への移行が多いことから，本研究では後者を対象とし，教師の説明と講義スライドの関連性を明確にするため，講義の進行に従って，講義スライド上に教師の説明箇所を明示することを検討している．そこで本研究では，教師の音声情報を利用して講義スライド中の教師の説明箇所を推定する手法を提案する．

講義スライド中の教師の説明箇所を推定する手法として，教師の使用するレーザーポインタや指示棒（以降，指示デバイスとする）の先端または光の位置や動きを画像処理によって検出し，それらに一番近い箇所を教師の説明箇所とする研究が多く行われている[1][2][3][4][5]．しかし指示デバイスは指示位置が不明確であったり，無意味な動きも多い．さらに指示デバイスを用いない講義には対応することができない．

また講義音声と講義スライドに関する先行研究としては，講義中の教師の発話区間，教師の発話のうち講義スライドに含まれる単語を多く含むものを重要度が高いものとして講義の要約を作成する研究[6][7]がある．他にも講義スライドまたは教師の発話中に出現する単語から視聴者が選択した単語が含まれる説明部分を講義スライドと同期させ再生する研究[8]や講義スライドの情報を用いたインデックスの研究[9]，講義中に教師が指示デバイスを使用した箇所をインデックスのタグとして用いる研究[10]などが行われている．音声に関しては講演音声の認識に関する研究の延長に講義音声の研究もされており，一般的に難しいとされる講義音声の認識率に関しては講義スライドから言語モデルを作成し学習させることで認識率を上げるという研究が行われている[11]．また教師映像とスライド映像を同期させる研究としてスライド映像に画像処理を行いスライドの切り替え情報を取得する研究などが行われている[12]．

2. 提案手法

本手法では，まず，講義スライドを形態素解析し，重要な形態素のみを抽出した形態素リストを作成しておく．次に，講義中の教師の録音音声を音声認識（Julius[13][c]）した出力結果と形態素リストを比較することにより，講義スライド中の教師の説明箇所を推定する．

2.1 形態素解析

形態素解析とは，文章を言語で意味を持つ最小単位（形態素）の列に分割してそれぞれの品詞を判別する手法のことである．日本語形態素解析ツールには Kakashi，Mecab，Chasen などがあるが，本システムでは Chasen[14]を使用する．

c) Julius はフリーの汎用大語彙連続音声認識ソフトウェアであり，認識結果をテキストで出力する．また今回は無音区間で分割して認識を行うオプションを用いている．これは Julius がデフォルトでは 20 秒を超える発話を認識できないことによる．

表 1　日本語形態素解析結果例

	読み	原形	品詞の種類	活用の種類
一	イチ	一	名詞	数
週間	シュウカン	週間	名詞-接尾-助数詞	
ばかり	バカリ	ばかり	助詞-副助詞	
ニューヨーク	ニューヨーク	ニューヨーク	名詞-固有名詞-地域-一般	
を	ヲ	を	助詞-格助詞-一般	
取材	シュザイ	取材	名詞-サ変接続	
し	シ	する	動詞-自立	サ変・スル
た	タ	た	助動詞	特殊・タ

2.2 説明箇所の定義

本研究では教師の説明箇所を講義スライド上に矢印などの記号で指示するシステムを目指している．そのため，本手法で推定する説明箇所は，実際にシステムが矢印などで指示する場所であることが望ましい．そこで，説明箇所を以下のように定義する．

(1) 新しいスライドへ移動直後はスライド全体の説明やスライドのタイトルの説明などをすることが多い．スライド全体の説明をしている場合，スライド中で指示すべき適切な場所は存在しない．また，タイトルを説明している状況においても，タイトルはスライド全体を表しているため,タイトル部分を指し示す必要はないと考えられる．そこで新しいスライドへ移動直後（以下の定義(2)で説明箇所が決定されるまで）はスライド中のどこも説明箇所としない「該当なし」と定義する．「該当なし」の場合は矢印などによる説明箇所の指示は行わない．

(2) 教師がタイトルを除くスライド中のある箇所を説明しているときはその箇所を説明箇所とする．

(3) 説明の途中でスライドの内容と直接的に関係のないたとえ話や雑談（以下，スライド外発話とする）が発生した場合，その時点でスライド中のどこまでを説明し終えたかを表すため，直前に説明した箇所を説明箇所とする．

(4) スライド中のある箇所を説明した後，そこよりも上の箇所をわずかに説明して，またもとの場所に戻ってくるような場合は，上に書かれていることを単に参照しているだけであると考えられるので，説明箇所を上へ移動させず，現在の説明箇所を保持する方が望ましい．そこである箇所を説明中に上の箇所を一発話以内で説明した後，元の箇所に戻ってくるか，下へ移動する場合， 上への移動は説明箇所の移動を行わな

い．ただし上の箇所となりうるのは参照となりやすい一段目の箇条書きのみとする．

2.3 講義スライドの形態素リストの作成

本節では講義スライドの形態素解析結果から重要な形態素を抽出した形態素リストを作成するまでの過程を述べる．はじめに各講義スライドを小さい単位（以下セグメントとする）に分割する．ただし，本研究では教師の授業を理解するための補助的手段として説明箇所を明示することを目的としている．しかし，あまり細かい単位でセグメント分割を行うと説明箇所が頻繁に移動してしまうという問題が起きる．そこで本研究では以下の定義に従ってセグメントを分割する．セグメント分割ではスライドが木構造となっていると仮定する．

(1) 箇条書きの行頭文字や改行が行われたところで内容が切れていると仮定し，セグメントの分割点の初期候補とする．

(2) 一段目の箇条書きの内側に二段目の箇条書きが一つのみ存在する場合，一段目の内容は二段目を包含していると考え，二段目の箇条書きを一段目の箇条書きと同じセグメントに含める（図 3）．

(3) 一段目の箇条書きの内側に二段目の箇条書きが複数存在する場合は，二段目の箇条書きは互いに内容が異なるとしてそれぞれを異なるセグメントとする（図 3, 図 4）．

(4) 箇条書きが三段階以上の入れ子構造になる場合は，三段目の内容は非常に細かい二段目の内容補足であるとして過剰に細かい分割を防ぐため三段目以上の深い箇条書きは単数か複数かに関わらず自身を含む二段目の箇条書きと同じセグメントに含める（図 4）

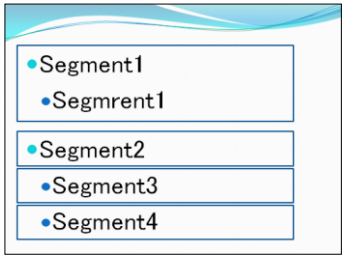

図 3　セグメント分割の条件(2)(3)

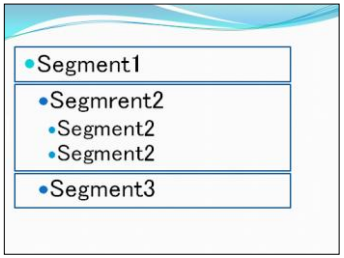

図 4　セグメント分割の条件(3)(4)

続いて形態素リストの作成方法を記述する．各セグメントを形態素解析により分析し出力された品詞のうち名詞，副詞および未知語のみを検出し，それら以外を全て除去することで形態素リストを作成する．この理由としては名詞・副詞・未知語がセグメントの特徴を表しやすいと考えられること，一つのセグメント内に多くの形態素を残したいことが挙げられる．名詞および未知語を採用する理由は講義で使用される専門用語である可能性が高いためであり，副詞を採用する理由は「同時に」「順に」など通常の発話に含まれにくくセグメントの特徴を表す形態素が多いことである．その後，以下の規則に従って，最終的な形態素リストを作成する．

(1) 形態素解析により名詞・副詞・未知語として判断されたもののうちスライドの内容に関わらず出現しやすい形態素は説明箇所の推定に悪影響を及ぼす可能性があるため，形態素リストから削除する．削除した単語の例は「もの」「こと」「の」「場合」「とき」「例」「部分」「的」「化」などである．

(2) 同じ形態素が同一セグメントに複数含まれる場合一つのみ形態素リストに残す．

(3) スライドに「321」「5.2」など 2 桁以上の数値が含まれるときは，それ自体を形態素リストに加えると同時に「3」「1」「2」とそれぞれの数字も単独で形態素リストに追加する．これは例えば教師が「321.2」に対して，「約 321 ですね」というようにその一部のみを発話した場合，「321.2」という形態素と対応づけることができないが，「3」「1」「2」と各数字ごとに形態素リストに登録しておけば，このような対応も可能となる．また現段階では高い音声認識率が期待できないため，「321」という数値を完全に認識することができなかった場合に対する処理でもある．

(4) Julius の音声認識出力結果は数値が全て漢数字で表示される．よってスコア計算時に形態素リストと音声認識結果を対応させるため，スライドに含まれる数値は全て漢数字に変換して形態素リストに加える．

2.4 講義音声の音声認識

講義音声から教師の音声認識結果を出力する過程を述べる．講義中の教師の録音音声は一つの長い音声ファイルとなっている．よって本研究では音声は無音区間が発話の切れ目であると仮定し，これを検出して音声を分割するプログラムを作成した．これにより録音音声を無音区間ごとに短い音声区間に分割し，それぞれに対して Julius を使用して音声認識の結果を出力する．作成プログラムによる無音区間検出方法は音声波形の振幅の二乗値であるパワー（単位：dB）が閾値以下になっている区間の時間長がある閾値以上となった区間を無音区間とすることで行う．

2.5 講義音声と形態素リストのマッチング

説明箇所を推定するためのマッチング処理では音声認識結果と形態素リストを比較して，そのスコアを計算することで行う．計算方法は，あるセグメントの形態素リストの各形態素が音声認識結果に含まれる度に該当するセグメントにポイントを 1 ポイント加算していく．この工程を繰り返し，1 音声区間分マッチングをとった後に最大ポイントを持つセグメントをこの音声区間での教師の説明箇所とする．ただし，全てのセグメントが 0 ポイントであるときはスライド外発話である可能性が高いとして一つ前に決定された説明箇所を引き継ぎ，新しいスライドの場合は「該当なし」とする．最大ポイントを持つセグメントが複数存在する場合は以下の規則により決定する．

(1) 直前に説明箇所としたセグメントが最大ポイントを持つ場合にはそれを優先する．また，スライドは上から下に進行するという仮定を用いて，直前に説明箇所としたセグメントよりも下にあるセグメントは直前のセグメントの上にあるセグメントよりも

優先する．最大ポイントを持つセグメントがともに直前のセグメントの上，あるいは下にある場合には直前のセグメントに近いセグメントほど優先する．
(2) 新しいスライドに移動した場合には一番上に近いセグメントほど優先する．以前説明したスライドに移動した場合は，あるスライドの途中で別のスライドに移動し，その後，もとのスライドに戻ってきた場合と考え，以前説明していた際に最後に説明箇所と推定されたセグメントを直前のセグメントとして(1)の規則を適用する．ただし，以前説明したスライドに戻ってきた場合であっても，一番下のセグメントが直前の説明箇所となっている場合は，すでにスライド全体の説明が終わっていると仮定して一番上に近いセグメントほど優先する．

上記の処理でスライドのタイトルが説明箇所と推定された場合は，2.2 節の説明箇所の定義(1)に従い，「該当なし」とする．また，定義(4)を実現するため，以下の規則により推定箇所を決定する．
(1) 説明箇所の推定結果が連続した三音声区間で，セグメント A⇒B⇒A と移動したときは A⇒A⇒A に変更する．
(2) ただし説明箇所の推定結果が連続して 4 音声区間で，A⇒B⇒A⇒B となっている場合は A⇒A⇒A と変更できる可能性も B⇒B⇒B と変更できる可能性もあるため(1)の処理は行わない．
(3) また説明箇所の推定結果が連続して 3 音声区間で，セグメント A⇒B⇒A と移動している場合でも B に変わる直前の A が 0 ポイントで，過去の説明箇所を引き継いだ場合も(1)の処理は行わない．

2.6 説明箇所の提示

説明箇所を提示する際には，各音声区間に対して，上述の手法で推定された説明箇所のセグメントの横に教師の音声に合わせて説明箇所を指示する矢印を表示する．矢印を採用する理由は画像の目的が指示することであると一見してわかるからである．また，講義スライドの上に重なっていることがわかるようにするという観点から矢印は透過表示とする．

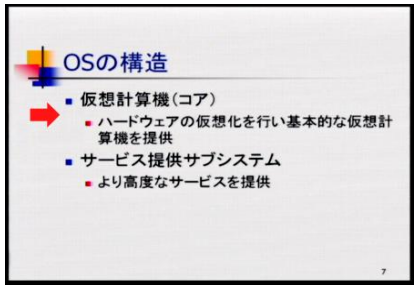

図 5 説明箇所の提示例

2.7 書き起こし情報の検討

OCW で配信するコンテンツを作成するにあたり二つの考え方が出来る．一つは教師の音声認識結果をシステムに入力して推定させる方法，もう一つは教師の音声を人手により書き起こした結果をシステムに入力して推定させる方法である．本研究で想定しているのは前者の音声認識結果を入力とした推定法である．この方法は配信の責任を考慮すると配信前の段階で，エラー確認・修正などを人手で行う必要があるが，録音音声を音声認識システムで認識させるため自動的に OCW のコンテンツを作成できる．しかしその反面，音声認識の認識率は現在 80%程度であるため教師の発話を完全に再現することは困難であり，説明箇所の認識率が書き起こしを用いたときと比較して低くなることが予想される．一方，書き起こしの作成方法としては録音した音声を聞きながらキー入力するものや，講義を聴きながらもう一度音声認識システムに向かって話すリスピークという方法などが考えられるが，これらは人手を必要とするため人件費や時間がかかるという欠点がある．しかしながら，海外では情報保障に対する考え方から講義の書き起こしを作成し字幕を付与することが多く行われている．よって将来的に字幕を付与することや耳が不自由な人でも使用できるコンテンツを作成する状況を想定した場合，書き起こしの利用性は高く，コンテンツ作成時に書き起こしも同時に作成することの利点も考えられる．よって本研究ではこのような書き起こしを用いた手法も考慮に入れ，二つの方法で音声情報をテキスト化し，それぞれの認識率を調査する．また，書き起こし情報を利用できるのであれば，無音区間によって音声を自動分割せずに，自然言語解析により適切な場所で分割することも可能である．そこで，教師の音声情報の句点に相当する場所で音声区間を分割した実験も行う．

3. 評価実験

本研究で提案した手法を用いて講義スライド中の教師の説明箇所の認識率を検証する実験を行う．実験では，人手による書き起こしを入力とする手法と音声認識結果を入力とする手法の両方の認識率を検証する．対象となる講義は本大学で行われた 5 名の教員による講義（計 8 講義）である．

3.1 実験資料

認識実験に用いた講義は本学科の教員が 2010 年度後期に行ったもので，以下の通りである．
・教師 A：講義 A-1（第一回：2010/10/1），講義 A-2（第二回：2010/10/15）
・教師 B：講義 B-1（第一回：2010/10/5），講義 B-2（第二回：2010/10/12）
・教師 C：講義 C-1（第三回：201010/25），講義 C-2（第七回：2010/11/11）
・教師 D：講義 D（第二回：2010/10/20）
・教師 E：講義 E（第三回：2010/10/18）

教師映像はハンディカメラを三脚に固定してフルハイビジョンの画質で撮影し，講義スライド映像は教師の操作するパソコンからプロジェクタに映像情報を送信する途中でスキャンコンバータによって画面情報を抽出し，標準画質で DV テープに録画した．また教師音声は胸元に付けたハンディマイクにより録音した．講義を行う教員には事前に撮影の許可はとってあり，通常通りの授業を行ってもらった．また教師 D を除くすべての教師は録音用のマイクの他に講義室に設置されている拡声機を使用した．

3.2 実験手順

形態素リストと教師の音声情報を用いて各講義における教師の説明箇所の認識率を算出する実験を行う．実験は講義内容が落ち着いたところで開始するため，講義開始から講義時間の 10 分の 1 である 9 分後にスクリーンに投影されていたスライドを含む，それ以降のスライドを実験の対象とする．基本的に連続したスライドを使用するが，以下のスライドは実験の対象外とする．

・写真や画像のみで構成されるスライド
・スケジュール，参考書の連絡など授業関連の連絡用のスライド
・小テストなど演習問題用のスライド

3.3 実験結果

説明箇所の正解は 2.2 節の説明箇所の定義に従い決定した．なお正解位置は教師が講義中に用いた指示デバイスの位置と文脈から判断したが，判断する人間による個人差はほとんどないと考える．最終的に音声認識結果を用いて出力した説明箇所と，人手による音声書き起こしを用いて出力した説明箇所の認識率を比較した．

実験では講義音声の分割方法の違いと入力したテキスト情報の違いにより，4 通りの実験条件が存在する．具体的には講義音声の分割位置を無音区間をもとに自動決定した場合を「自動分割」，講義音声の書き起こしテキストの句点の位置で分割した場合を「句点分割」とする．また，システムの入力に書き起こしテキストを用いたものを「書き起こし」，音声認識の結果を用いたものを「音声認識」とする．これらの組み合わせにより，「自動分割－書き起こし」「自動分割－音声認識」「句点分割－書き起こし」「句点分割－音声認識」の 4 通りの実験条件となる．正解の位置および用いたアルゴリズムは全て同じである．ただし自動分割において一音声区間に複数の発話を含む場合のみ複数正解を認める．

表 2　各講義の説明箇所平均認識率

	自動分割-書き起こし	自動分割-音声	句点分割-書き起こし	句点分割-音声
A-1	95%	51%	88%	60%
A-2	71%	64%	69%	60%
B-1	56%	30%	66%	31%
B-2	69%	59%	71%	68%
C-1	60%	38%	55%	36%
C-2	62%	58%	61%	69%
D	86%	56%	85%	54%
E	88%	14%	75%	17%
平均	76%	48%	71%	49%

結果より，最も高い認識率は無音区間を利用して音声区間を分割する自動分割を行い音声情報として書き起こしを与えた「自動分割-書き起こし」の 76%となった．また自動分割で音声情報として音声認識結果を用いた場合の説明箇所の平均認識率は 48%と，書き起こしを用いた場合と比較して 28%低下した．なお，後述するように音声認識率は平均 23%となっているが，本提案手法は音声認識の失敗に対してロバストなアルゴリズムとなっているため，説明箇所の認識率はそれほど悪い結果にはならなかったと考えられる．

4. 考察

4.1 書き起こしでの推定に関する考察

書き起こしを与えた場合の認識誤り結果を分析した結果，以下の原因が明らかとなった．

(1) 指示語の影響

指示デバイスを用いた講義では説明箇所に含まれる形態素を言葉で説明しなくても指示デバイスを用いてセグメントや画像を指し示すことが可能なため「これ」「ここ」などの指示語が用いられることが多い．指示語が使用された場合，スライド中の形態素が発声されないとマッチングのポイントが与えられないため，形態素リストと音声のマッチングが失敗する事例がみられた．この問題点を解決するためには先行研究で行われている指示デバイスの位置を画像処理によって検出する方法などを併用することで改善できると考えられる．そこで，今回の自動分割＋書き起こしによる実験結果に対し，指示語による誤認識を指示デバイスの情報を利用したと仮定して認識率を推定したところ，76%から 81%まで向上できることが分かった．

(2) 略語の使用による影響

実験で使用した講義は専門科目であるため専門用語が使用されやすい．またスライド上では文章を簡潔に表記することが望まれるため，講義で用いられる専門用語はスライド上に正式名称ではなく略語で表記されることが多い．たとえば「エントロピー」を「H」,「リードソロモン記号」を「RS 記号」と略記される事例がみられた．このとき教師がスライド中の略語に対して正式名称を読み上げて説明すると，教師の音声情報中に説明箇所の形態素が含まれず，マッチングに失敗してしまう．そこで本研究では実験時に形態素リストに略語と正式名称を併記することとし,どちらを発話されても対処できるようにした．しかしながら，将来，自動システムを想定した場合には略語から正式名称を推測し，それら両方を形態素リストに加える必要がある．

(3) 形態素リストに登録した品詞による影響

現在は形態素リストに加える形態素の品詞として名詞･未知語･副詞を採用している．しかし教師の音声情報中に形態素リストと一致する形態素が一つも含まれずセグメントの切り替わりが検出出来ない，または各セグメントで同一の形態素しかヒットせず各セグメントのポイントに差がつかない音声区間が多く見られ，これにより認識率が低下した．よって本研究で採用している形態素のみではスコア計算に用いるセグメントの特徴量として不十分である可能性が考えられる．これを解決するためには現在形態素リストに採用している品詞の他にも，単語マッチングなどに用いられやすい形容詞や形容動詞，動詞の採用を検討する必要がある．

(4) スライド外発話の影響

講義音声の中にはスライドの内容を直接説明する発話だけではなく，たとえ話や雑談などのスライド外発話も含まれる．しかしスライド外発話はスライド内容とは直接関連性のない内容であっても,間接的にスライドと関連する内容であることが多いため，講義スライド中の形態素が含まれてしまうことがある．このような形態素が発声されることにより認識率が低下してしまった．これに対してはある閾値以下のポイントの場合にスライド外発話と認定するようなアルゴリズムとすればよい．また，その際，各形態素に対して重みづけをすることによりスライド外発話の検出が改善されると考えられる．

(5) スライド内の相互関連性の影響

並列に並べられた箇条書きが複数存在する場合，それらには関連性があるため，同じ形態素が含まれることが多くなるが，そのことによりマッチングに失敗する事例が見られた. たとえば, 図 6, 7 において, 下二つのセグメントでは 3 単語が重複している．もし,「仮想化というのは実際に物理的にある，資源の量」という発話が発声された場合,「量」という単語もマッチしてしまうため，一番下のセグメントが出力結果となるが，実際には真ん中のセグメントが正解である．また，箇条書きのうち上に位置するセグメントが下に位置するセグメントを補足する構造を持つ場合がある．たとえば図 8 において，上のセグメントでは「光の特性」と「視覚の特性」と書かれているが，下のセグメントでは「両者」と書かれている．そのとき，「(網膜像の処理には) この光の特性と視覚の特性両方が含まれているんですね」と発話されると，上のセグメントとマッチングしてしまうが，実際には下のセグメントを説明している．

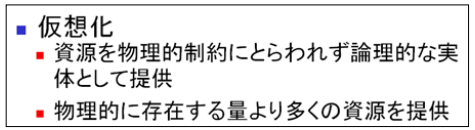

図 6　類似構造を持つスライド例

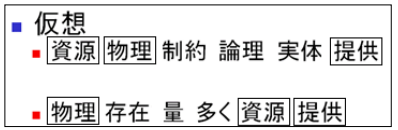

図 7　図 6 の形態素リスト

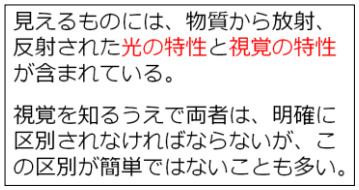

図 8　補足構造を持つスライド例

4.2 音声認識に関する考察

表 3 に各講義の単語音声認識率を示す．なお単語音声認識率とは書き起こしに含まれるスコア計算に使用された形態素の数に対する音声認識結果に含まれるスコア計算に使用された形態素数の比率とする．

表 3　各講義の単語音声認識率

講義	単語認識率
A−1	18%
A−2	26%
B−1	26%
B−2	24%
C−1	19%
C−2	28%
D	36%
E	4%
平均	23%

一般に講義音声は日常会話とは異なり講義に依存した専門用語を多く含む．また原稿がなく, その場で考えながら話し言葉で行うため, 言い淀みや無意味語 (フィラー) などが含まれ音声認識率は低いと言われている．講義音声に関する先行研究[11]でも，もともとの音声認識率は 58.61%となっており，言語モデルの改良により 60.97%まで

改善している．その結果を踏まえると，本実験での単語認識率は平均 23%となっており，著しく低い．この理由としては以下のことが考えられる．

(1) 録音環境の影響

教師 D を除く全ての教師が録音用のマイクの他に拡声機を使用したことが大きな要因として考えられる．これにより録音用のマイクに教師による直接の音声と拡声機による音声が二重に録音され，非常に悪い録音環境となってしまっていた．通常音声認識に用いる音声はヘッドセットなどを装着し口元で録音するのが一般的であるが，本研究では通常の講義に支障を与えないように収録するためピンマイクを使用したことで教師の口からマイクまでの距離が長くなったことも原因の一つと考えられる．

(2) 専門用語の使用

今回使用した音声認識システム Julius では日常会話に基づく言語モデルを使用しているため講義に使用される専門用語は検出しにくくなっている．また今回実験に用いた講義は教養科目ではなく専門科目であるため専門用語も多用された．たとえば,「カーネル」「バースト」「グリフィス」「シャノン」「仮数」などの用語が用いられるが，これらは一般的な音声認識辞書には登録されていない．これらは先行研究[11]などと同様，スライドから専門用語を抽出し，言語モデルを学習すれば改善できると考えられる．

(3) スライド外発話

各音声区間の音声情報と単語音声認識結果を比較するとスライドの説明をしているときよりもスライド外発話を行っているときに音声認識率が低くなることが多かった．これは教師がスライド上の箇所を説明するときや専門用語を説明する際には丁寧に発声を行おうと気を配るのに対して，講義内容に関するものであっても，スライド外発話の中で単語を発音する場合には，さほど発声に気を配らないことが要因として考えられる. このような発話スタイルの違いにより, 音声認識率が低下したと考えられる．

4.3 句点分割と自動分割に関する考察

システムに入力する音声情報について書き起こしを用いた方が音声認識結果を用いた時と比較して説明箇所の認識率が高いものが多かった．また句点分割は人手で確認して発話を綺麗に分割するため認識率が高いと予想したが，平均認識率を比較すると自動分割の方が認識率が高かった．これは句点分割の平均音声区間数が 32，自動分割の平均音声区間数が 19 であることから，意味情報により音声区間を区切ると，一つ一つの音声区間が長くなってしまうためスコア計算に使用される形態素の数が増え，スコア計算に失敗してしまうことが原因として考えられる．このことから，自動分割を用いることが良いと考えられるが，意味情報の観点から見ると本来，発話が切れるべきところで分割されずに，一つにまとまってしまう例が数多く見られた．説明箇所の提示を行う際，分割点が不適切であると矢印の描画のタイミングがずれてしまうという問題が起きてしまう．また自動分割の場合には，書き起こしの場合とは逆に一つの音声区間に含まれる形態素が少なすぎてスコア計算に失敗する例も見られた．これらの結果から，自動分割と句点分割の利点を活かした分割点を推定する手法を検討することが必要であると考えられる．

5. おわりに

本研究では OCW の配信コンテンツを作成することを目的として教師のスライド中の説明箇所を推定し提示するシステムを作成した．また書き起こしから説明箇所を推定するシステムと音声認識から説明箇所を推定するシステムを想定して各システムの認識率を調べた．その結果書き起こしを用いた場合 76%と比較的高い認識率が得られた．ただし音声認識を用いたシステムという点では録音環境の問題もあり認識率が著しく低下してしまったため改善の必要性が考えられる．

今後は音声認識率の改善を検討すると同時に，形態素リストに追加する形態素の検討，スコア計算・形態素リスト作成時に適切な重み付けを行うことなどにより説明箇所の認識率の向上を図りたい．

謝辞 本研究の実験に際して講義の撮影を快く引き受けてくださった千葉大学の先生方 5 名に心から感謝いたします．

参考文献

1) 勝山裕，小澤憲秋，武部浩明，直井聡，横田治夫”講義ビデオ中のレーザーポインタ抽出の一検討”，電子情報通信学会技術研究報告，103(656)，pp.37-42，2004-02-12
2) 高見澤大輔,”レーザーポインターを用いたプレゼンテーションシステムの提案”，明治大学理工学部情報科学科計算理論研究室卒業論文 (参考文献として良いのか？)
3) 高澤剛，福間慎治，大泉好史，奥田篤士，桜井哲真,” 効果的な遠隔講義のためのポインティングシステムとその評価”， 電子情報通信学会総合大会講演論文集 2007 年_情報・システム(1)，168，2007-03-07
4) 丸谷宜文，西口敏司，各所孝，美濃導彦,”講義における教材中の指示対象の抽出”，電子情報通信学会論文誌，J90-D(5)，pp.1238-1248，2007-05-01
5) 丸谷宜文，西口敏司，各所孝，美濃導彦,”講義における教材中の指示対象の抽出”，電子情報通信学会論文誌，J90-D(5)，pp.1238-1248，2007-05-01
6) 横井隆雄，桐井孝嘉，藤吉弘亘,”講義イベント検出に基づく短縮講義ビデオの自動作成”，第 12 回画像センシングシンポジウム予稿集，pp.535-540，2006-06
7) 藤井康寿，山本一公，北岡教英，中川聖一,”重要文抽出に基づく講義音声の自動要約”，情報処理学会論文誌，51(3)，pp.1094-1106，2010-03
8) 富樫慎吾，山口優，北岡教英，中川聖一,”講義音声の認識・要約・インデックス化の検討”，情報処理学会研究報告，SLP-62(11)，pp.57-62，2006-07-08
9) 井上宗徳，下川俊彦,”講義スライドのフッターを用いたラベルづけによる講義映像のインデックス作成に関する研究”，電子情報通信学会技術研究報告，ET-107(391)，pp.1-6，2007-12-08
10) 仲野亘，越智悠太，小林隆志，勝山裕，直井聡,”統合プレゼンテーションコンテンツ検索

におけるレーザーポインタ情報の利用”，第 16 回電子情報通信学会データ工学ワークショップ (DEWS2005) 論文集，2B-o1，2005
11) 根本雄介，河原達也，秋田裕哉，”スライド情報を用いた言語モデル適応による講義の音声認識と字幕付与”，情報処理学会研究報告，NL-179(16)，pp.91-96，2007-05-25
12) 柳沼良知，”スライドとの同期による講義映像のデータベース化”，メディア教育研究第 5 巻第 1 号， pp.109-114，2008
13) 汎用大語彙連続音声認識エンジン Julius　http://julius.sourceforge.jp/
14) 日本語自然言語処理システム chasen
15) 松本裕治,北内啓,山下達雄,平野善隆,松田寛,浅原正幸,”日本語形態素解析システム『茶筅』Version2.0 使用説明書第二版”1999-12