2005－SLP－55　(12)
2005／2／5

講演音声とプレゼンテーション資料の対応付けによる講演検索

中澤 聡　　佐藤 研治　　奥村 明俊
ＮＥＣ　メディア情報研究所

テキスト検索と同様に、大量に録画・蓄積されたビデオデータに対しても、利用者の求めるシーンをキーワード検索する機能が求められている。本研究では、講義・講演映像の音声認識結果とプレゼンテーション資料中のテキストという、性質・分量の異なる２種類のテキストを対応付けることで、プレゼンテーション資料を元の講義・講演映像の言語インデックスとする手法を提案する。対応付けはプレゼンテーション資料のスライド単位でとる。これにより、スライド単位で、プレゼンテーション資料のテキストを介した講義・講演映像のキーワード検索が可能となる。正解対応付けとのずれ時間を評価した結果、全スライドの約 87%が誤差１分以内に収まった。

Alignment of lecture speech data and presentation documents for lecture search

Satoshi Nakazawa　Kenji Satoh　Akitoshi Okumura
Media and Information Research Laboratories, NEC Corporation

Keyword search for the certain scene in video data seems to be in great demand as well as text search. We would like to propose a new approach which enables to make a text index without detailed scripts but with presentation slides. We focus on lecture videos, and we will explain how to make a text index by aligning two different materials; speech recognition results and presentation slides. We align them by slide so that keyword search for lecture videos can be done by slide. We evaluated the difference time between the correct alignment and our alignment results. About 87% slides are aligned in 1 minute error or less.

1. はじめに

e-Learning などにおいて、録画・蓄積された講義・講演などのビデオデータから利用者の必要とするシーンを検索する機能が求められている。近年、そうしたビデオデータに音声認識処理を行って、元のビデオデータと同期した時間情報付きのテキストインデックスを作成し、検索やブラウジングに利用する研究が行われている。あらかじめ時間情報付きのテキストインデックスを作成し、テキスト検索に用いることで、ビデオデータに対してもテキストデータと同様に検索・閲覧処理が可能となり、利用者は大量のビデオデータの中から必要なシーンを必要なだけ頭出し再生するといった、柔軟な利用法が可能となる。

こうしたビデオデータから音声認識を用いてテキストインデックスを作成する手法としては次の３つが考えられる。１つ目は、音声認識結果をそのままテキストインデックスとする手法である[1]。この手法ではビデオデータのみからテキストインデックスを自動的に作成できるという利点があるが、一方、音声認識誤りや未知語のため、作成されるテキストインデックスが不正確になるという欠点が生じる。

２つ目は、ビデオデータの書き起こしや台本など別途作成した正確なテキストを用意し、それと音声認識結果との対応をとることで、時間情報付きのテキストインデックスとする手法である。この手法では、書き起こしや台本など正確なテキストを利用するため、得られるテキストインデックスは精度の高いものとなるが、元のビデオデータの音声をそのまま書き起こしたテキストが必要となるため、その書き起こしテキストの作成コストが課題となる。

３つ目は、ビデオデータの関連資料を別途用意し、音声認識結果と対応をとることで、時間情報付きのテキストインデックスとする手法である。２つ目の手法と異なり、関連資料には元のビデオデータの正確な書き起こしとなっていることを要求しないため、入手が容易であるという利点があるが、一方、音声認識結果との対応付けはより難しくなる。関連資料としては、講義の目次[2]や、教科書・参考図書など様々なものが考えられるが、本稿では、講義・講演によく用いられるプレゼンテーション資料(以下、プレゼン資料と省略して記す)に着目し、プレゼン資料と講義・講演ビデオがセットになって与えられたときに、講義・講演の音声認識結果とプレゼン資料との対応付けを行うことで、元のビデオのテキストインデックスとし、検索に利用する手法を提案する。

2. 講演音声とプレゼン資料との対応付け

2.1 プレゼン資料との対応付けの課題

前節で述べたように、講義・講演ビデオとそのプレゼン資料との対応をとり、プレゼン資料のどの部分が、ビデオのどの部分にあたるのかを示すテキストインデックスを作成するのが、本研究の目的である。そのようなテキストインデックスが作成されれば、講義・講演ビデオと同期して自動的にプレゼン資料を切り替え表示させることや、プレゼン資料にテキスト検索を行って、目的とするキーワードを含むプレゼン資料の一部を抽出し、それに対応した箇所だけ元の講義・講演ビデオを再生する、といった操作が可能となる。しかし、プレゼン資料は、ビデオ音声の正確な書き起こしではないため、音声認識結果と対応をとるにあたって、次のような問題点がある。

(1) テキスト分量の不一致
プレゼン資料中のテキスト分量と、ビデオの対応する箇所の発話分量が大きく異なる場合がある。

(2) 発話順、記述順の不一致
プレゼン資料とビデオで大まかな話題の推移は順序が一致するが、１枚のスライド内では必ずしもプレゼン資料に記述された順に単語が発話されるわけではない。

(3) 複数スライドに共通するテーマ・キーワード
講義・講演では、一般に１つのテーマやキーワードを複数のスライドに渡って説明する場合がある。

ビデオの正確な書き起こしがある場合には、その書き起こしと時間情報付き音声認識結果に対して DP マッチングを用いて対応付けを行えば、結果として時間情報付き書き起こしを得ることができる。これはそのままビデオ検索用のテキストインデックスとして利用できる。しかし、プレゼン資料に含まれているテキストとの対応付けを行う場合には、上記のような問題があるため、DP マッチングアルゴリズムを、そのまま利用することができない。そこで、本手法では次のようなアプローチをとる。

(A) スライド単位の対応付け
発話の音声認識結果を文の区切りと推定される箇所で切り、認識文とする。ついで認識文と、プレゼン資料中のスライドを単位とした、文対スライドの、多対１の対応付けを行う。すなわち、プレゼン資料の各スライドが、ビデオデータで発話された何番目から何番目までの文に対応するかを求める。通常の１対１のＤＰマッチングと異なり、１枚のスライドに対して複数の文が対応することになるため、テキスト分量の差を吸収できる。ただし、１枚のスライドに過剰の文が対応してしまうことを防ぐために、各スライドに対応可能な文数の上限の目安に、スライド毎のテキスト長を利用する。

(B) 共通単語のみチェック
各スライド内で、スライドテキストと認識文とを対応付ける際には、音声認識した単語が認識文中に出現する順序を考慮せずに、各認識単語がそのスライド中にも出現しているかどうかのみ調べる。

(C) 談話指標の利用
単語の一致だけでなく、スライド切り替わりの目安となるような談話指標を数パターン予め用意し、対応付けの手がかりとする。

講義・講演ビデオに対する検索では、検索キーワードが発話されているシーンを厳密に頭出しできる必要は少なく、むしろ検索したいキーワードが含まれている説明の区切りから頭出し再生できることが望ましい。そこで、プレゼン資料のスライド単位で対応付けを行うよう設定した。スライドより細かな単位での検索は不可能となるが、通常１枚のスライドは１分程度の長さであり、検索として十分な粒度を持つといえる。また対応付けさえ正確ならば、頭出し再生はスライドの先頭という、説明の区切りから行われる。講演音声側の対応付けの単位も、同様の考え方から、ビデオの音声認識結果を文単位で区切ったものとしている。単語単位でなく、文単位で対応をとることで、発話の途中でスライドが切り替わる、といった望ましくない対応付け結果をなくすことができる。さらに、キーワード以外に文頭・文末で現れる「では次に」のような、スライド切り替わりの手がかりとなる表現を利用できるため、対応付けの精度が向上する。次節でこのアプローチに基づいた対応付け手順について説明する。

2.2 対応付け手順

図 1 では対応付けの大まかな手順を示している。ビデオデータが与えられると、まずビデオのオーディオ・トラックに対して不特定話者の時間情報付き音声認識を行う。

ついで統計的言語モデルとポーズ長、パターンマッチングルールから認識単語を文毎にまとめ、音声認識テキスト側の対応付けの単位とする[3]。さらにまとめた各文の文頭・文末を調べて、「さて」「次に～する」「次の～です」や「思います」「訳です」など予め用意した十数個のスライド切り替わり表現パターンに一致するかどうかチェックし、一致する文はスライド

先頭候補、スライド末尾候補のようにマークする。

これらのスライド切り替わりを示す談話指標は、講義映像とプレゼン資料両方が揃っている講義データ(約３時間分)から、人手で作成した。まず、講義音声の正解書き起こしと、各プレゼン資料のスライドが切り替わる箇所の正解データを作成した。ついで、スライド切り替わり箇所の書き起こし両側１文の文頭・文末に出現する表現をリストアップし、そのリストの中から、スライドの切り替わり箇所以外の書き起こし中に出現する回数が一定値以下のものを選択した。文頭・文末表現のみに限定したのは、講義内容や分野に依存しない談話指標を抽出するためである。また、比較的少量の講義データから抽出する都合上、データ数が少ないときに変動が大きくなる情報量基準を、談話指標の選択に用いなかった。

一方、プレゼン資料からは、スライドごとにそこに含まれているテキストを抽出する。さらにプレゼン資料全体のテキスト長に対する各スライドのテキスト長の割合を求める。

次に、文対スライドのＤＰマッチングを行う。ただし、多対１の対応を許し、マッチングのコスト関数で、１枚のスライドに、閾値を越える数の文が対応しないよう制御する。この閾値はスライド毎に下式のように設定する。

$$\alpha \times (\text{全文数}) \times \frac{\text{スライドのテキスト長}}{\text{全スライドのテキスト 長}}$$

α はスライドに対応する文数の偏りを許すための倍率で、今回は実験的に調整した結果、６で固定した。またマッチングの得点関数(負のコスト関数)として、文とスライドで共通する単語数を使用する。さらに、最初の文切りの処理で、スライド先頭候補、スライド末尾候補としてマークされた文がそれぞれ、各スライドに対応する最初の文、最後の文となるときは、得点関数に一定値を加える。スライドの先頭や末尾に、こうしたマークがついていない文が対応するようなアライメントもとり得るが、この加えられる得点の分だけ、スライド先頭候補、スライド末尾候補としてマークづけられた文がアライメントされやすくなる。マッチングに使用する得点関数のこうしたパラメータ値は、現在実験的に定めている。

最後にＤＰマッチングの結果から、各スライドのテキストと、そのスライドに対応する文の時間情報をまとめて、時間情報付きテキストインデックスとする。このインデックスにより、各スライドがビデオの何分何十秒目から何分何十秒目までに対応しているかが、求められる。よって、プレゼン資料のスライドに含まれるテキストに、キーワード検索をかけることで、そのキーワードについて触れているビデオの箇所を直接頭出し再生可能となる。

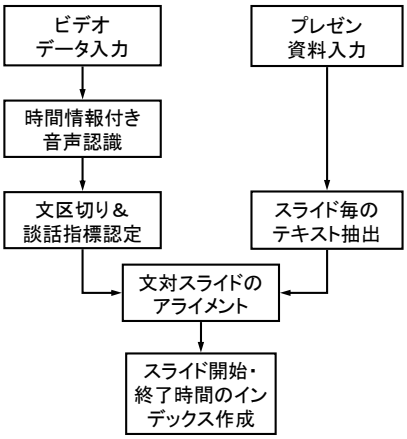

図１　アライメント手順の概要

3. 評価実験

本手法の対応付け精度の評価として、IT 分野の講演２つ、ネットワーク分野の講義２つ、セキュリティ分野の講義１つ、計５つの講義・講演ビデオを対象に、２つの評価実験を行った。各ビデオのサイズは、表１に示してある。これらのビデオは、前述した対応付け手順のいくつかのパラメータ値を定めるために用いていない、オープンコンテンツである。

表１　実験対象コンテンツのサイズ

対象コンテンツ	ビデオの時間	スライド枚数
IT講演1	50分50秒	49
IT講演2	55分13秒	38
ネットワーク講義1	73分41秒	30
ネットワーク講義2	68分50秒	43
セキュリティ講義	63分40秒	87

まず第１の実験として、人手で各ビデオのプレゼン資料スライドに対応する正解時間データを作成し、それと本手法で得られた結果とを比較して、スライド毎のずれ時間を求めた。図２がその結果で、各講義・講演に対して、スライド毎に平均どれだけのずれ時間が生じたかを示している。また図３では、全スライドのうち、何パーセントが誤差 60 秒以下、30 秒以下、3 秒以下に収まったかを示している。

図２，図３で「音声認識１」「音声認識２」「書き起こし」とあるのが本手法による対応付けの結果である。３種類あるのは、ビデオの音声認識精度が対応付け精度にどの程度影響するのか確認するため、複数の音声認識結果を用いて比較したためである。

「音声認識１」は、ディクテーション用音声認識エンジンを本研究に流用した結果である。

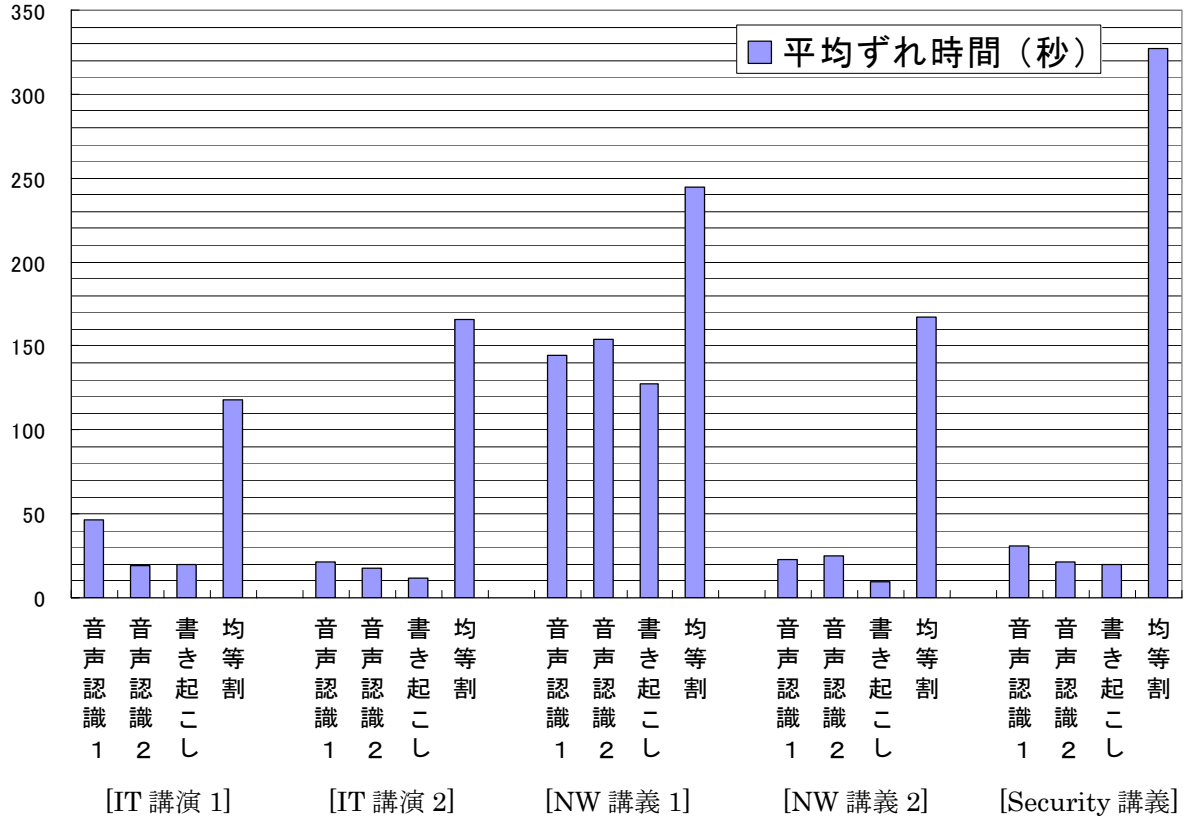

図２　人手による正解を基準としたスライド平均ずれ時間

講義や講演などのコンテンツを音声認識する目的で作成されてはいないため、音声認識率は低くこの５つのコンテンツに対してはパーセントコレクトで約 45%となっている。

「音声認識２」は、音響モデルや言語モデル作成時に、講義・講演映像から集めたテキストを学習コーパスの一部として利用した音声認識エンジンを本研究に流用した結果である。音声認識率は、「音声認識１」のエンジンより良く、この５つのコンテンツに対してはパーセントコレクトで約 65%となっている。

「書き起こし」と記されているのは、人手で作成した時間情報付きの書き起こしを音声認識結果のフォーマットに直して、対応付けに使用した結果である。文区切りも人手で行っている。理想的な音声認識結果が得られた場合の本対応付け手法の上限を示すモデルとして、本実験では比較に取り入れた。

さらにこの実験では、対応付け手法の比較対象として、ビデオの全時間をスライド枚数で頭割りした場合の正解とのずれ時間を「均等割」として挙げている。

つぎに本対応付け手法に対する談話指標の効果をしらべるため、第２の実験を行った。第１の実験と同じ５つのコンテンツに対して、他は同じ条件で、対応付けに談話指標を用いない場合、どれだけずれ時間が増大したのかを示しているのが表２である。表２の中で「談話指標の効果」とあるのが、このずれ時間の増大分(秒)である。この値が大きいほど、対応付けに対する談話指標の効果が大きかったことになる。逆に、この値が負の場合、談話指標を用いない方がずれ時間が小さくなることを示している。

また、表２の「談話指標のＰＣ」は、各音声認識テキストに含まれている談話指標にマッチする箇所のパーセントコレクトを参考情報として示している。

4. 考察

図３によると、本提案手法を用いた対応付け結果で、多くのスライドが誤差 60 秒以下に収まっていることが分かる。これは視聴したい内容のスライドをキーワード検索して対応するビデオを頭出し再生すると、多くのスライドではその前後１分以内に目的とする箇所が見つかるということを示している。

一方、比較対象として行った、ビデオの全時間を単純にスライド枚数で頭割りする均等割手法では、スライド毎に平均２分～６分以上のずれが生じている。これは今回実験対象とした５つの講義・講演において、あるスライドは一瞬でとばされるが、あるスライドは 10 分以上費やして説明されているといった、スライド毎のばらつきが起きている結果である。このようなば

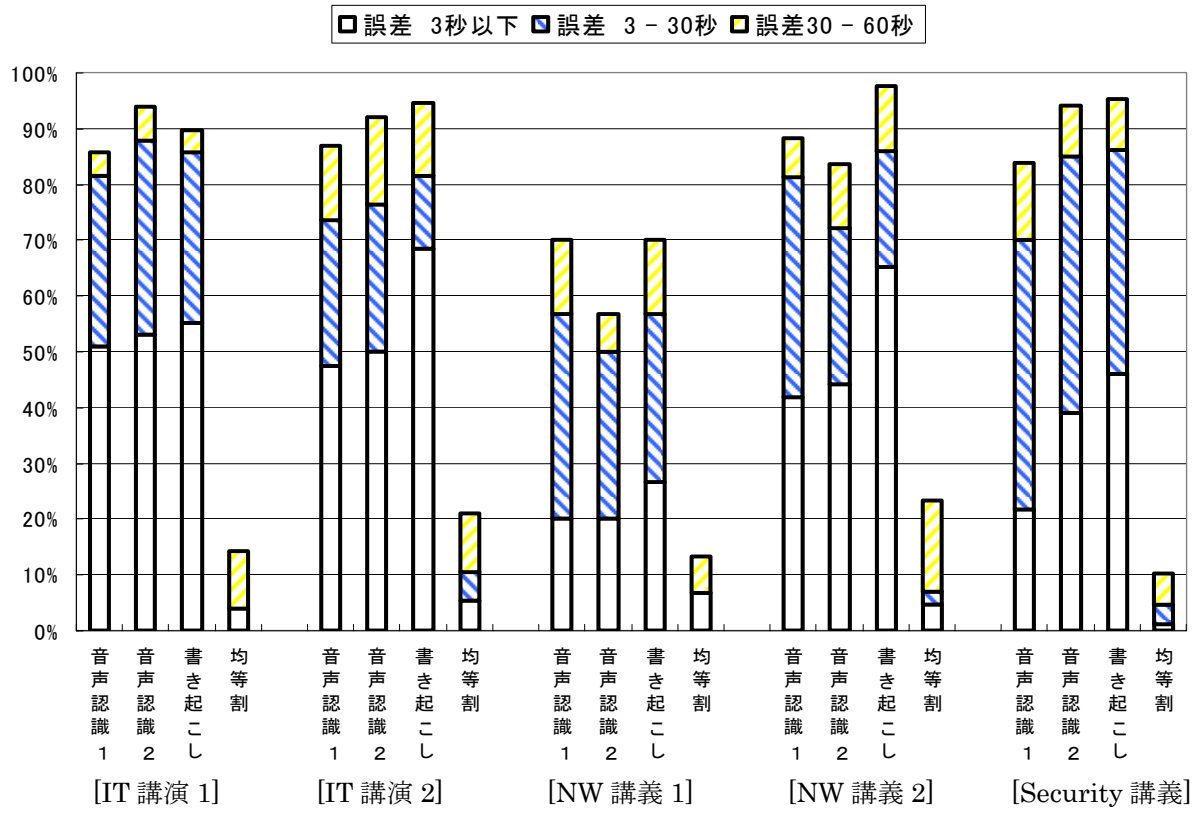

図３　スライド全体に対する誤差一定範囲のスライドの割合

らつきは、今回の実験対象に限らず、講義・講演において一般によく生じることであり、ビデオ検索に利用するには、単純なスライド数による見積もり以上のインデックス作成手法が必要であることが分かる。

本提案手法による対応付け結果にずれが生じる要因としては、音声認識テキストに含まれる認識誤りがまず考えられる。しかし、それが唯一の要因でない事が、図２，図３の結果から判断できる。図２では、音声認識精度が 100%である書き起こしを用いた場合でも、ずれ時間は減少しているものの０にはなっていない。逆に音声認識精度が比較的低い「音声認識１」であっても、スライド平均ずれ時間は「音声認識２」の場合と同程度である。

各コンテンツに対して対応付け結果を詳細に調べたところ、大きなずれが生じているのは、使用した音声認識テキストの種類によらず、次に挙げるような場合・箇所であった。

[対応付け不得意箇所]

- 複数のスライドにわたって共通するキーワードが頻出する箇所
- スライド内容と関係性の薄い挿話などにより、スライドに含まれるキーワードは発話されないが、そのスライドに対応する時間としては長くなる箇所
- プレゼン資料にないスライドやデモが挿入された場合
- スライドの説明がとばされた場合

本提案手法は基本的には、キーワード一致と、テキスト長に基づいて大まかな対応をとり、スライド区切りを示す談話指標でスライドの切り替わり箇所付近の細かな対応付けを修正する手法であるといえる。上記のような「対応付け不得意箇所」では、音声認識精度が向上しても、得られたキーワードが対応付けの手がかりとして有効に働かず、テキスト長も対応付けの指標にはできないため、大きなずれが生じる。実際、ずれの特に大きい「ネットワーク講義１」では、複数の共通のキーワードを 20 枚以上のスライドにわたって説明しているため、音声認識精度によらずその部分で大きなずれが発生している。一方、対応付けの正解箇所ともいえる、誤差３秒以内のスライドの割合は、全てのコンテンツにおいて音声認識精度が向上するとともに増えていることが図３から分かる。

本提案手法に対する談話指標の効果は、表２から判断できる。前述したように、談話指標は大まかな対応をとった後、スライドの切り替わり箇所付近での細かな対応付けを修正するものである。表１の５つのコンテンツに対して、談話指標を用いた場合と、用いない場合の差分を

表２　談話指標の効果と音声認識率

対象コンテンツ	音声認識の種類	談話指標の効果総和(秒)	談話指標の効果スライド平均(秒)	談話指標のPC
IT講演1	音声認識1	89.46	1.83	26.4%
	音声認識2	60.15	1.23	72.5%
IT講演2	音声認識1	424.22	11.16	12.6%
	音声認識2	158.97	4.18	67.8%
NW講義1	音声認識1	0.00	0.00	12.5%
	音声認識2	-1248.67	-41.62	62.5%
NW講義2	音声認識1	0.00	0.00	1.1%
	音声認識2	0.00	0.00	26.7%
Security講義	音声認識1	-1.25	-0.01	8.3%
	音声認識2	3.93	0.05	33.3%

詳細に調べたところ、改善箇所の多くはもともとずれ時間の少ない箇所に見られた。これらの改善は、スライドの頭出し再生を行う場合などにおいて、利用者の印象や利便性の向上に有効である。ただし、表２の「ネットワーク講義１」「音声認識２」の結果のように、大きな悪影響がある場合も存在する。このコンテンツに対して談話指標が影響している箇所を調べたところ、前述した複数の共通のキーワードを 20 枚以上のスライドにわたって説明している場所で、談話指標によるスライドのずれが後続のスライドに受け継がれ、それらのずれ時間が蓄積したためと分かった。今回は全てのコンテンツに対して等しく、談話指標によるマッチング得点関数の調整を行っているが、今後は何らかの対応付け信頼度を導入して、対応付け信頼度が低いコンテンツに対しては、談話指標による調整を行わないという手法も考えられる。対応付け信頼度に利用できる基準としては、例えば、対応付けに用いている共通単語の情報量がある。プレゼン資料中、１枚のスライドにのみ出現する単語は情報量が大きい。一方、全てのスライドに共通する単語は、対応付けには有用でない。実際にどのような対応付け信頼度が有効であるかは、今後の検討課題としたい。

5.　まとめと今後の予定

本稿では、講義・講演につきもののプレゼン資料を利用して、ビデオ検索用のテキストインデックスを作成する手法を提案し、実際のビデオコンテンツにおいて大部分のスライドを誤差１分以内でビデオの適切な箇所に対応させられることを示した。利用者は、任意のスライドに対応するビデオの箇所を頭出し再生することが可能となる。また頭出し再生だけでなく、スライドと同期してビデオ視聴するなど、このインデックスを e-Learning システムなどに応用することが考えられる。インデックス作成後、講義・講演ビデオに対するキーワード検索は、プレゼン資料から抽出したスライドテキストを介して行うため、専門用語など、音声認識エンジンにとっての未知語も検索可能となる。

今後は、本対応付け手法でずれが大きくなる場合について、より詳しく確認実験を行うとともに、他の手がかりを利用した方法についても検討していきたい。また、対応付け結果の利用法についても検討をすすめたい。講義・講演ビデオの検索としても、プレゼン資料のテキストと、音声認識結果の両方を組み合わせてインデックスに利用する手法なども考えられる。また本対応付け手法は、テキスト分量の異なる２種類の類似テキストを大まかに対応させる手法であるといえる。プレゼン資料とビデオの音声認識テキスト以外の組合せに対しても、検討したい。

参考文献

[1]藤井敦, 伊藤克亘, 石川徹也, "音声文書検索の応用によるオンデマンド講演システム", NLP8, Mar.2002.
[2]山本夏夫, 緒方淳, 有木康雄, "トピックセグメンテーションに基づく講義ビデオの構造化の検討", 音声言語情報処理, 42-10.
[3]中澤聡, 佐藤研治, 奥村明俊, "ビデオ音声認識テキストからの文認定", NLP8, Mar.2002.
[4]中澤聡, 佐藤研治, 奥村明俊, "談話指標とテキスト長を利用した講演音声－プレゼン資料アライメント", FIT2003, Sep.2003.
[5]磯健一,磯谷亮輔,畑崎香一郎,服部浩明,奥村明俊,渡辺隆夫,” 大語彙連続音声認識技術と応用”, 日本音響学会 2004 年春季研究発表会,Mar.2004