様式 C-19

科学研究費補助金研究成果報告書

平成２２年６月１日現在

研究種目：基盤研究（C）
研究期間：2007～2009
課題番号：１９５００８４５
研究課題名（和文）　自学自習を支援するビデオ教材検索システムの開発

研究課題名（英文）　A lecture video retrieval system supporting self learning

研究代表者
金寺　登（KANEDERA　NOBORU）
石川工業高等専門学校・電子情報工学科・教授
研究者番号：５０１９４９３１

研究成果の概要（和文）：ビデオ教材を最初に閲覧することによって、初学者にとって、その分野の概要を学習しやすい利点がある。ビデオ中の音声情報を利用し、どのような形式の講義ビデオに対しても、高速に検索できるシステムを開発した。提案するビデオ検索システムでは、講義ビデオ音声より抽出したサブトピック情報を利用するため、目印やスライド中のキーワード、キャプションなどを必要としない。これにより、あらゆるビデオコンテンツに対応することができる。

研究成果の概要（英文）：The method of retrieving a subtopic of a lecture video is examined. When the subtopic of the lecture video is retrieved, we can use slide information etc. However, only speech information is used in this study, because the lecture using blackboards that doesn't use the slide is targeted.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	900,000	270,000	1,170,000
２００８年度	1,200,000	360,000	1,560,000
２００９年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：音声情報処理
科研費の分科・細目：科学教育・教育工学
キーワード：教材情報システム、ビデオ教材、サブトピックセグメンテーション、音声認識、ビデオ検索

１．研究開始当初の背景

近年、ブロードバンドネットワーク環境が拡大し、インターネット上に様々なビデオ教材が公開されるようになってきた。また、地上波ディジタル放送が開始し、ハードディスクビデオレコーダが普及するとともに、一般ユーザでも大量の高画質映像データを蓄積できる環境が整った。蓄積された大量のデータの中から閲覧したい映像を探し出すのは、困難である。そこで、ビデオ自体を解析して検索を行うことができる映像情報検索技術のニーズが高まっている。

ニュースやスポーツ中継に関しては、画像のカラーヒストグラムなどを利用し、検索を行う方法が提案されている。また、キーとなるフレームを並べて表示することで、検索を

容易にする方法もよく用いられる(IEEE Signal Processing Magazine, Vol.23, No.2,pp.18-123, 2006.3)。ニュース等の場合には話題（トピック）や画像が明確に切り替わるため、話題を分割し、検索することは比較的容易である。一方、講義ビデオの場合には、画像が明確に切り替わることは少ない。また、１回の講義では、１つか２つの話題について取り上げることが多い。すなわち、講義ビデオにおける検索は、サブトッピック検索に相当し、難しいタスクである。

最初にビデオ教材を閲覧することによって、初学者にとって、その分野の概要を学習しやすい利点がある。しかし、再度、ビデオ教材を見返すためには、必要な部分を見つける必要があり、時間を要する。必要な部分を高速に検索することができれば、自学自習に有効である。講義ビデオ中から必要な部分を検索する方法として、目印をつけることができる付箋機能などが一般的である。またスライドを用いる講義の場合には、スライド中のキーワードを用いることができる（情報処理学会研究報告 2006-SLP-62 , p57, 2006.7)。しかし、目印をつけてない場合やスライドを用いない講義の場合には再度閲覧したいところを高速に検索することは困難である。

２．研究の目的

本研究では、ビデオ中の音声情報を利用し、どのような形式の講義ビデオに対しても、高速に検索できるシステムを開発することを目的とする。提案するビデオ検索システムでは、講義ビデオ音声より抽出したサブトッピク情報を利用するため、目印やスライド中のキーワード、キャプションなどを必要としない。これにより、あらゆるビデオコンテンツに対応することができる。

３．研究の方法

(1) サブトピック分割方法

本研究ではトピックよりも小さいサブトピックをビデオ素材の分割単位とし、サブトピックのビデオ区間をシーンと呼ぶことにする。講義ビデオシーンを自動分割できれば、ユーザは必要なシーン候補(segment) を選択もしくは不要なシーンを削除するだけで容易にビデオ教材を作成することができる。

① 隣接シーン間の類似度を用いる方法

まずビデオを仮にいくつかのシーンに分割する。次にシーンごとに指標に変換し隣接するシーンごとに似ているかどうかを調べる。隣接するシーンが似ているかどうか調べるには指標の余弦を用いる。余弦が小さい程シーンは似ておらず、大きい程シーンは似ていると考えられる。つまり、指標の余弦の総和が最小であれば全てのシーン間が似ていないことになり、ビデオを適切にシーン分割できると考えられる。そこで、シーン分割位置推定を余弦の総和が最小となるようなシーンの組合せを探す問題とみなし、動的計画法(Dynamic Programing;DP) を用い解くことができる。

②統計手法

①では、隣接するシーン間が類似していないと仮定し、シーン分割(サブトピック分割)を行った。本節では統計的手法を用いた場合のシーン分割方法を示す。

テキストセグメンテーションにおいて、以下のような統計的手法が提案されている。$W = w_1 w_2 \cdots w_n$ を n 単語からなるテキストとし、$S = S_1 S_2 \cdots S_m$ を m セグメントからなるセグメンテーションとする。このとき、セグメンテーション S の確率は次のように定義される。

$$\Pr(S \mid W) = \frac{\Pr(W \mid S)\Pr(S)}{\Pr(W)} \qquad (1)$$

よって、セグメンテーションの推定値 $\hat{S}$ は次のように求めることができる。

$$\hat{S} = \arg\max_{S} \Pr(W \mid S)\Pr(S) \qquad (2)$$

i 番目のセグメント S_i における単語数を n_i、i 番目のセグメント S_i における j 番目の単語を w_j^i、$W_i = w_1^i w_2^i \cdots w_{n_i}^i$ とする。W_i が独立で S_i のみに依存し、w_j^i も独立と仮定したとき、$\Pr(W \mid S)$ は次式のように求められる。

$$\begin{aligned} \Pr(W \mid S) &= \Pr(W_1 W_2 \cdots W_m \mid S) \\ &= \prod_{i=1}^{m} \Pr(W_i \mid S) \\ &= \prod_{i=1}^{m} \Pr(W_i \mid S_i) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n_i} \Pr(w_j^i \mid S_i) \end{aligned} \qquad (3)$$

W_i において、w_j^i と同じ単語の出現数を $f_i(w_j^i)$、W における異なり単語数を k としたとき、$\Pr(w_j^i \mid S_i)$ は次式のように近似される。

$$\Pr(w_j^i \mid S_i) \equiv \frac{f_i(w_j^i)+1}{n_i + k} \qquad (4)$$

(4) 式は、ラプラス法として知られている。

また、S に関する事前知識を用いず、$\Pr(S)$ を次式のように定義する。

$$\Pr(S) \equiv n^{-m} \qquad (5)$$

以上より、式(2) を対数化し、動的計画法を用いれば、最適なシーン境界を求めることができる。

(2) サブトピック検索方法

①検索キーワードの補完

テキスト検索では、検索キーワードを補完する方法が利用されることがある。本研究では、講義音声を書き起こしたテキストや音声認識したテキストに対して、検索キーワードの補完が どの程度有効であるかを調査する。

検索キーワードを補完する方法を情報源で分類すると、辞書を用いる方法、web 等の情報を用いる方法、併用する方法などが考えられる。辞書を用いる方法は一般用語を多く含むが専門用語を含んでいないことが多い。逆に、web 等の情報を用いる方法は、一般用語を含まないこともあるが、専門用語を含む可能性が高い。

概念で分類すると、上位概念、下位概念、同義語などが考えられる。これらの概念を検索することができる日本語 Wordnet を用いて調査することとした。

講義ビデオのサブトピック検索方法を図 1 に示す。講義ビデオの各サブトピックの指標として TF-IDF 値を事前に求めておく。検索時には、指定されたキーワードを、辞書や Wordnet などを用いて、上位概念、下位概念、同時出現語などの連想キーワードにキーワード拡張する。次にキーワード及び連想キーワードを用いて検索ベクトルを作成する。検索ベクトルのキーワードに対応する要素には α をセットし、連想キーワードに対応する要素には $(1-\alpha)/(\text{連想キーワード数})$ をセットする。α には実験的に 0.25 を用いた。この検索ベクトルと、音声認識テキストよりサブトピック毎に予め作成しておいた TF-IDF ベクトルとの余弦を計算する。余弦値の大きいサブトピックの順に検索結果とする。

②サブワード検索

検索キーワード拡張は指定されたキーワードが直接サブトピック中で発話されなくとも検索できる。一方、サブワード検索では、指定されたキーワードが直接サブトピック中で発話されているが、正しく音声認識されていない場合や、未知語の場合に有効である。よって、キーワード拡張とサブワード検索を併用することが望ましい。

サブワードを用いたサブトピック検索方法を図２に示す。サブワードとしてtri-phoneを用い、音声認識時に使用する音響モデルを用いて、tri-phoneモデル間のconfusion matrixを作成する。分布間の距離には、Bhattachayya distanceを用いる。与えられた検索キーワードはtri-phone系列に変換され、各サブトピックのtri-phone系列と連続DPによって照合される。照合の結果、スポッティング距離の小さい N (本実験では500個)を抽出する。N 個を距離の小さい順に順位づけし、対応するサブトピックに1/(順位) をスコアとして加える。スコアの大きいサブトピックを検索結果とする。

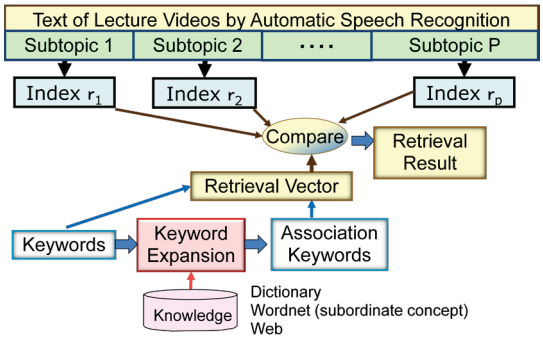

図1　キーワード拡張によるサブトピック検索

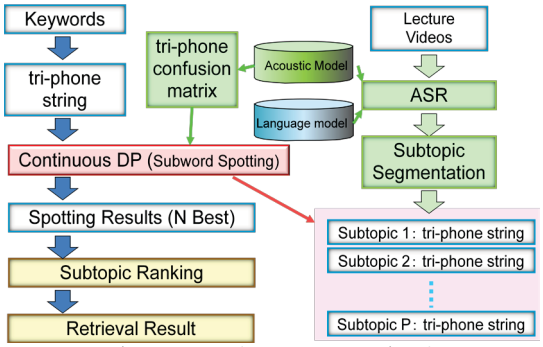

図2　サブワードモデルによるサブトピック検索

４．研究成果

(1) 講義ビデオ音声に対する認識性能

実験対象として、5名の男性教員による約90分の講義5回分のビデオ素材を用意した。収録には接話型ヘッドセットを用いたため、雑音等の影響は少ない。対象となるビデオ素材から音声情報のみを抽出し、16 kHzにダウンサンプリングを行った。次に音声区間ごとにJulius-4.0を用いて音声認識を行った。音声認識結果（単語正解率 ・単語正解精度）を表１に示す。言語モデルは講演の書き起しテキスト(CSJ)により学習された言語モデルを用いて認識を行った。音響モデルは、953学会講演(CSJ)により学習された3000状態16混合の状態共有tri-phoneモデルを用いた。本研究で評価に用いた講義内容は、新聞や話し言葉の講演内容と異なることからパープレキシティが大きく難しいタスクである。

(2) サブトピック自動分割性能評価

ビデオ教材作成支援のために講義ビデオ素材を自動的にサブトピック分割する方法として、テキストセグメンテーションに用いられている統計的シーン分割手法を講義音声に応用した。図３のように統計的手法は、隣接するシーンの語分布が類似していないと仮定する類似法に比べ、サブトピック分割において優れていることが確認された。

また、本報告の方法は、音声認識を一種の符号化器として利用しており、符号化された単語が同一の単語であるか異なるのかでシーン分割しているため、音声認識誤りに強い。特に、置換誤り率が大きくなってもシーン分割結果には、ほとんど影響がないことがシミュレーション等によって確認された。

(3) サブトピック検索評価

講義ビデオのサブトピック検索結果を図４に示す。図中の MRR は、平均逆数順位（Mean reciprocal rank）である。「Ideal」は、各サブトピック毎に TF-IDF 値の大きい理想的な検索キーワードが指定された場合の結果である。「Real」は、５名の評価者が実際に検索に用いた検索キーワードによる結果である。理想的な検索キーワード「Ideal」に比較して、実際のキーワードでは検索性能が悪くなっていることが確認できる。例えば、書き起こしテキストを用い、３個のキーワードを用いたとき、理想的な検索キーワードでは MRR が 1.0(音声認識テキストでは 0.802)であるのに対し、実際のキーワードでは MRR が 0.677 に低下した。さらに音声認識テキストを用いた場合には MRR が 0.511 になった。この結果からも、未知語や抽象表現への対処が必要と考えられる。

講義音声を音声認識したテキストに対して、検索キーワードの補完が どの程度有効であるかを調査した結果を図５に示す。検索キーワードには実際のキーワードを用いた。図中の「dic」は、検索キーワードを辞書で補完し、検索した結果である。具体的には、まず辞書を検索し、検索キーワードに対応する説明文を取得する。取得した説明文より、自立語のみを抽出し、検索ベクトルを生成する。この検索ベクトルと各サブトピックの TF-IDF ベクトルを比較することで検索を行う。図中の「Wordnet (subordinate)」は、日本語 Wordnet により、検索キーワードの下位概念を連想単語として検索した結果である。これらの結果よりキーワード拡張が有効であることが確認された。

Wordnet において、上位概念、下位概念、同時出現語による違いを音声認識したテキストに対して調査した結果を図６に示す。図中の「All」は、日本語 Wordnet により、検索キーワードの上位概念、下位概念、同時出現語のすべてを連想単語として検索した結果である。図中の「Subordinate & Consequent」

表１ 講義ビデオの音声認識結果

Lecture video	Word correct rate [%]	Word accuracy [%]	Out of vocabulary [%]
1	59.5	46.3	3.9
2	55.7	44.6	2.2
3	53.9	42.3	3.7
4	41.3	30.3	3.6
5	54.5	41.7	3.8
mean	53.0	41.0	3.4

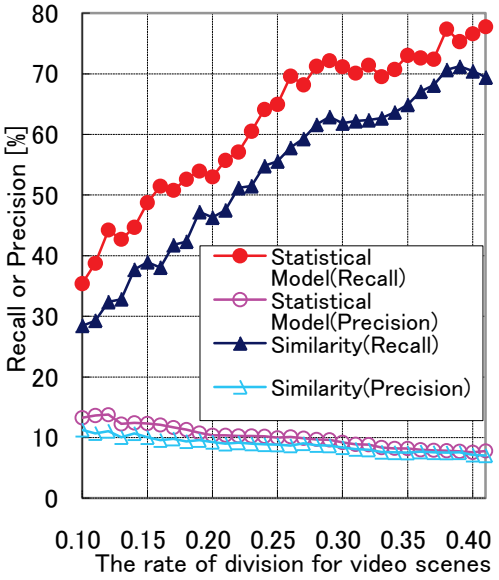

図３ サブトピック分割結果

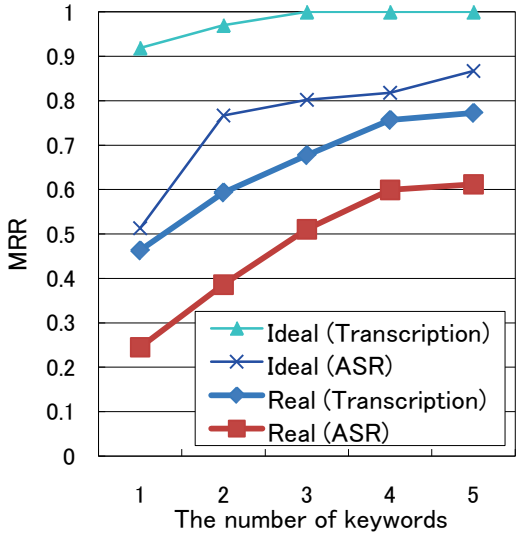

図４ 講義サブトピック検索結果

は、日本語 Wordnet により、検索キーワードの下位概念、同時出現語を連想単語として検索した結果である。図中の「Consequent」は、日本語 Wordnet により、検索キーワードの同時出現語を連想単語として検索した結果である。この結果より、下位概念と同時出現語を併用して補完する方法と、下位概念のみを用いて補完する方法は、図では重なって見えないが、同等であることがわかる。また、上位概念、下位概念、同時出現語すべてを併用して補完する方法は、下位概念のみを用いて補完する方法よりも、検索性能が劣化するが、同時出現語のみで補完するよりも優れていることがわかる。以上より、下位概念のみを用いる方法が最も良いことがわかった。

講義ビデオを、5 名の評価者に閲覧してもらい、各サブトピックキーワードを抽出してもらった。抽出されたキーワードの内、未知語(OOV)及び発話内容に含まれないキーワード数を表 2 に示す。表中の抽象化(Abstraction)は、「演習，実習，説明，結果，解説，表し方，書き方，概要，本題，導出，雑談，昔話，種類，導入，回答，図解，話の流れ，例題，経緯，動作，　質問，質疑応答，まとめ」などのように、直接発話していないが、発話内容を総括したり、イベントを表現したキーワードである。表より、未知語ばかりでなく抽象化表現への対処も重要であることが確認された。図 6 において、下位概念のみを用いる方法が最も良かったのは、直接発話されていないキーワードを下位概念にキーワード拡張することによって発話されたキーワードに変換できたためと考えられる。また、音声認識されたテキストにおいて、37.2%のキーワードが音声認識テキスト中に出現しなかったが、Wordnet(下位概念)を用いることで 26.9%に、辞書を用いることで 14.3%に改善された。

辞書によるキーワード拡張とサブワードを併用した結果を図 7 に示す。「Dic+Subword(all)」は図 1 のように辞書を用いたキーワード拡張による検索方法と、キーワード拡張された全ての連想単語に対して図 2 のサブワードに基づく検索方法を併用した結果である。「Dic+Subword」は、キーワード拡張による検索方法と、直接指定されたキーワードに対してサブワードに基づく検索方法を併用した結果である。「Subword」は、キーワード拡張しない図 1 の検索方法と、直接指定されたキーワードが未知語の場合のみサブワードに基づく検索方法を併用した結果である。「Subword only」は、直接指定されたキーワードに対してサブワードに基

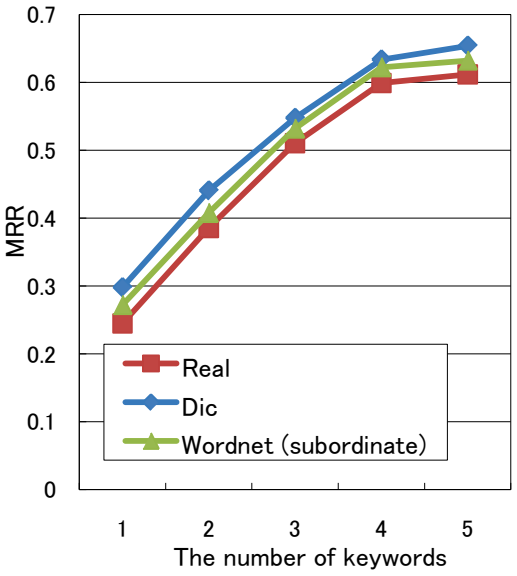

図5　音声認識テキストに対してキーワード拡張を用いた講義サブトピック検索結果

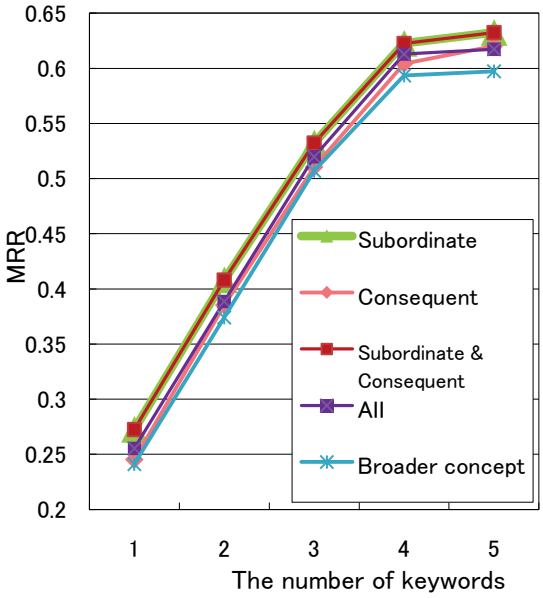

図6　Wordnet によるキーワード拡張を用いた講義サブトピック検索結果

表2　評価者によるキーワード抽出

Lecture video	OOV	Abstraction	The number of keywords
1	12	9	90
2	13	10	152
3	1	8	45
4	14	17	115
5	23	8	84
Total	63 13.0%	51 10.7%	486 100%

づく検索方法のみを用いた結果である。キーワード数が１のとき、「Subword only」はキーワード拡張「Dic」よりも優れているが、キーワード数が２以上の場合、「Dic」の方が優れている。キーワード拡張とサブワードに基づく方法併用した場合には、全てのキーワード数で優れていることが確認された。このことは、Wordnet を用いた場合にも確認された。

以上のように、以下の知見が得られた。

① ２種類のシーン分割方法を比較し、統計的なシーン分割方法がよいことが明らかになった。
② 音声認識性能は講義ビデオシーン分割への影響が小さい。
③ ２つの教育機関でのビデオ素材に対して、同様の講義ビデオシーン分割性能が得られた。
④ 音声認識性能は話題検索性能への影響が大きい。
⑤ 発話されていないキーワードに対応するため、キーワードを辞書等の知識によって補完することにより、シーン検索性能が向上する。
⑥ 未知語や音声認識誤りに対応するためサブワード単位での検索を併用すると検索性能が向上する。

今後は、高専間教育素材共有システム(http://ctm.ishikawa-nct.ac.jp)に本研究の講義ビデオ検索機能を組み込む予定である。

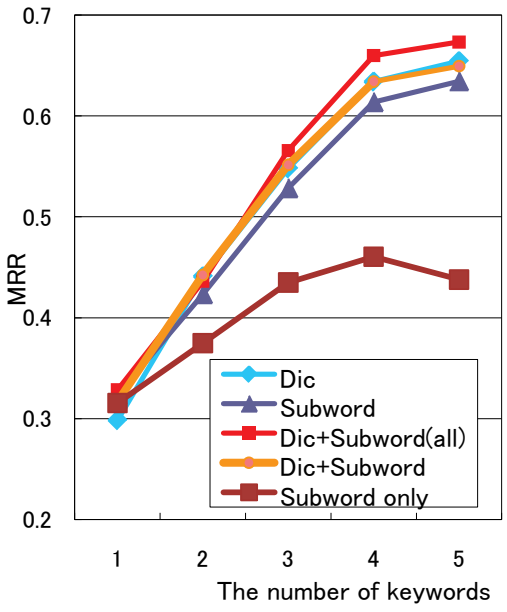

図7 キーワード拡張とサブワードモデルを併用した講義サブトピック検索結果

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
① 増田 哲也、金寺 登、講義音声認識における音声認識特徴量の検討、石川高専紀要、査読有、No.42、 2010、37－42

〔学会発表〕（計７件）
①金寺 登, 上江まり子, 船田 哲男, 中川 聖一、統計的な手法による動画シーン分割性能の改善、第２回音声ドキュメント処理ワークショップ、2008.3.1(豊橋)

②金寺 登, 本多 由依, 増田 哲也, 船田 哲男, 中川 聖一、第３回音声ドキュメント処理ワークショップ、 2009.2.27(豊橋)

③金寺 登, 本多 由依, 船田 哲男, 中川 聖一、 講義音声を利用したサブトピック分割、日本音響学会 2009 年春季研究発表会、2009.3.17（東京）

④ペク キムホーチ, 荒井 隆行, 金寺 登、変調フィルタリングによる自動音声区間検出とその多言語における比較、日本音響学会 2009 年春季研究発表会、2009.3.17（東京）

⑤Y. Yamada, N. Kanedera, R. Komura and S. Okano、Improvement of ECE Student Skills Through Achievement Test and Project-Based Learning、ACE2009 （8th IFAC Symposium on Advances in Control Education）、2009.10.23 (Kumamoto)

⑥金寺 登, 船田 哲男, 中川 聖一、検索キーワード補完による講義サブトピック検索、第４回音声ドキュメント処理ワークショップ、2010.2.26 (豊橋)

⑦金寺 登, 船田 哲男, 中川 聖一、検索キーワード補完による講義サブトピック検索、日本音響学会 2010 年春季研究発表会、2010.3.8（東京）

〔その他〕
ホームページ等
http://sail.i.ishikawa-nct.ac.jp

６．研究組織
(1)研究代表者
金寺 登（KANEDERA NOBORU）
石川工業高等専門学校・電子情報工学科・教授
研究者番号：５０１９４９３１