2006－SLP－62（11）
2006／7／8

講義音声の認識・要約・インデックス化の検討

富樫 慎吾　　山口 優　　北岡 教英　　中川 聖一

豊橋技術科学大学情報工学系 〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1 の 1
E-mail :　{togashi, yamaguchi, kitaoka, nakagawa }@slp.ics.tut.ac.jp

概要

本稿では、大学院における講義音声を対象として行った音声自動書き起こし、要約、およびセグメンテーション／インデキシングといった一連の自動処理について報告する。これらの処理は視聴者が、動画や音声といったいわゆるマルチメディア情報をより効率よく利用できるようにするために不可欠な技術である。講義音声においては、ビデオ映像とあわせて要約やインデックスを付加することによって、利用者の学習効率の向上が見込まれる。我々はまず、講義を録音する際に問題となる収録方法が認識性能に及ぼす影響を調べた。その結果、収録方法の違いによって、単語正解精度に最大 23.9% の差が現れることがわかった。また、音声認識結果を利用した要約実験では、κ 値で平均 0.389 という結果が得られたが、人間による要約の平均値 0.484 とはまだ大きな隔たりがあった。箇条書き形式のスライドの各項目と文とを対応付けるセグメンテーション実験においては、ベースラインとして設定した時間均等や文均等分割と同程度の結果であるなど、課題が残る結果となった。スライド中のキーワードや発話内容からインデキシングする方法は、講義音声の視聴に有効であることがわかった。

キーワード　講義音声、音声認識、音声要約、インデキシング

An Investigation of Recongnition, Summarization and Indexing of Lecture Speech

Shingo TOGASHI Masaru YAMAGUCHI Norihide KITAOKA Seiichi NAKAGAWA
Department of Information and Computer Sciences, Toyohashi University of Technology
1-1, Hibarigaoka, Tempaku-cho, Toyohashi 441-8580, Japan
E-mail :　{togashi, yamaguchi, kitaoka, nakagawa}@slp.ics.tut.ac.jp

Abstract

In this paper, we described the procedure of automatic speech recognition, sentence extraction and segmentation/indexing for classroom lecture data of our university's graduated cource. This process is necessary to improve the usability of broadcasting sound or video data. In the case of lecture, summarized and indexed lecture speech or video enables to students to more effective leaning. Our goal is to construct a framework of such structured lecture contents. To achieve this goal, first, we investigated influence of the recording methods on the recognition performance. It turned out that there was 23.9% difference on the accuracy. Second, we tried automatic summarization by extracting important sentences, and we obtained 0.399 κ value , but still far from the score by human doing 0.484. We also tried to segment the speech according to the items in a slide, but our method was not so better in comparison with the baseline and found that the automatic segmentation is very crucial tasks. Finally, automatic indexing functions gave us the accessibility of video contents.

key words　Classroom lecture speech, Speech recognition, Speech summarization, Indexing

1　はじめに

ネットワークの容量増加に伴い、講義や講演をビデオ録画し、自宅からでも容易に学習や復習が可能なシステムが実用化されている [1]。また、講義中に使用しているスライドの切り替わりとビデオを自動対応付けするツールもある[1]。しかし、このようなシステムやツールには動画の再生、早送り、巻き戻しや、スライドバーによる任意位置からの再生など、標準的な閲覧環境しか実装されていない。e-Lerning の長所はいつでも好きなときに学習、復習が可能なことであるが、そのビデオは基本的に通常速度で再生されるため、利用には収録時間と等しい時間が必要となる。そこで近年、講義内容をより効率的に把握できるようにするため、音声要約や、自動インデキシング、セグメンテーションの研究が注目を集めている [1][2][3][4][5]。本研究では、講義の際に収録する音声や動画といったメディアファイルに対して、書き起こし [5]、要約、セグメンテーションやインデキシングを自動的に行い、効率的な学習を支援するシステムを構築することを最終的な目的とする。2 節では、講義音声の収録方法について、3 節では、2 節で収録した音声に対する認識実験について記す。4 節では、分割した各音声に対して文の重要度を推定し、自動的に抽出

[1]日立 EZ Presentator ／ Microsoft Producer

する音声要約を試みた。5 節では、スライドの文、および単語と音声再生位置を自動的に対応付けるセグメンテーション／インデキシングについて実験を行った。

2 講義音声収録

収録対象とする講義は、我々の大学で実施されている大学院向けの音声言語情報処理に関するものである。講義は一回 150 分で，前後半 75 分ずつとなっている。３名の講義を合計４回分収録した。その際、録音機材が認識結果に与える影響を調査するため、各講義について、表 1 に示す 3 種類の装置を同時に使用して収録を行った。

収録した講義は、1，2 回目の話者が同一で、それ以外は異なる講義者が担当している。それぞれ話者 SN、NK、TN とする。以下、これらの音声データを便宜上表 2 に示すように記述する。1-1 と 1-2 の講義内容は、「音声言語処理の概要」、2-1 と 2-2 は「DP マッチングと連続音声認識」、3-1 と 3-2 は、「マルチモーダルインタラクション」、4-1 と 4-2 は、「音声認識とパターン認識」である。各講義について、録音機材の違いによって表 1 のソース A, B, C が存在する。以降は、たとえば 1-1 のソース A なら、1-1A と表記する。なお、表 2 の”PP” は後述する CSJ 言語モデル (語彙サイズ 17K) に対するテストセットパープレキシティを示す。文は、便宜上おおよそ 200ms の無音区間のところで切り出した単位としている。

3 講義音声認識

2 節に示した各講義音声について、音声認識実験を行った。認識条件などを以下に記す。

3.1 認識条件

認識率の評価には最初の 10 分程度だけを使用した。これは使用した単語辞書にあわせて、単語分割を人手作業により行う必要があったためである。

この音声試料を、Julius と SPOJUS[6] の 2 種類のデコーダを用いて音声認識実験を行い，結果を比較した．音響モデル及び言語モデルの学習用データは CSJ(日本語話し言葉コーパス 2004 年度版) に収録されている音響学会講演と模擬講演であり，音声認識および対話といったトピックが主であるため，今回の認識対象である講義の内容とドメインは近く、表 2 に示すように未知語

表 1: データ収録条件

収録データ	マイクロフォン	録音装置
ソース A	SONY C-355 ハンドマイク 周波数特性：20Hz～20,000Hz 指向特性：単一指向性	DAT
ソース B	SONY ECM-C10 ピンマイク 周波数特性：50Hz～15,000Hz 指向特性：全指向性	DAT
ソース C	TOA WM-1300 ワイヤレスピンマイク 指向特性：全指向性	PC (WMV 圧縮)

表 2: 各講義音声の概要

表記	話者	総時間	総文数	PP	未知語率
1-1	SN	1:07:56	742	186.4	0.37%
1-2		0:56:59	709	443.7	2.15%
2-1		1:06:11	831	159.8	0.70%
2-2		1:15:28	798	305.5	1.65%
3-1	NK	1:05:49	680	177.7	1.88%
3-2		1:11:14	1099	180.8	3.14%
4-1	TN	1:10:16	582	285.6	1.94%
4-2		1:18:30	648	239.5	2.11%

率 (語彙サイズ 17K) は比較的小さい。音声認識システムは以下に示すものを用意した。

(a)SPOJUS

SPOJUS は，16kHz でサンプリングされた音声より導出された MFCC(12)、Δ MFCC(12)、Δ POWER(1) の計 25 次元の特徴ベクトルを使用する。2 パスデコーダであり、1 パス目はコンテキスト独立の 133 音節からなる音響モデル [10] と bigram 言語モデルを用い、得られた N-best 候補を trigram でリスコアする [6]。

(b)Julius3.5

Julius では特徴パラメータは、上記したものと同じ MFCC，Δ MFCC，Δ POW の 25 次元を用いる．なお，Julius には高速版と高精度版が存在する．高精度版はビームサーチのアルゴリズムが異なり，2-pass 目の単語間 triphone をより厳密に計算するようになる．高速版と高精度版の比較実験を行ったところ、高精度版の性能が高かったので，本実験では高精度版を使用した．

表 3: 認識システム

表記	デコーダ	音響モデル	言語モデル
システム i	SPOJUS	音節 (133 音節)*1)	CSJ*2)
システム ii	Julius	音節 (133 音節)*1)	CSJ*2)
システム iii	Julius	音素 (triphone*3))	CSJ*4)

*1 CSJ 最終版 797 講演 (男性話者) より学習、コンテキスト独立

2 CSJ 最終版で学習、trigram 言語モデル (17635 語彙)

3 CSJ 最終版 DVD に収録。2496 講演 (男性+女性話者) で学習

4 CSJ 最終版 DVD に収録。trigram 言語モデル (25300 語彙)

3.2 認識結果

認識精度 (accuracy) の評価結果を図 1 に、正解率 (correct) の評価結果を図 2 に示す。ここに示すスコアは各講義の前半と後半の平均である (例：1-1A と 1-2A の平均=1A)。なお、集計時にフィラー (全発話単語の 5%前後) や言い淀みなどを取り除く操作は特に行っておらず、それらも認識精度を左右する要因となっている。

マイクによる差という観点からソース A,B,C を比較すると、各講義において，A ＞ B ＞ C という順に認識性能が劣化していることがわかる．人間の耳では A と B の音声はいずれも聞き取りやすい音質でほとんど差はなかったが、特に話者 SN の認識精度には最大 13.4%もの差が表れている。ソース C は，マイクの性能としてはソース B のものと同程度であるが，録音時に WMV による圧縮をかけるために，かなり機械的な音声に変

質しており，ノイズも増幅されているような印象を受けた．結果として，ソース B と C の間には話者 TN では最大 16.8%の認識精度の差が表れている．さらにソース A と C では，話者 TN で 23.9%もの認識精度の差が表れた。

デコーダの比較という観点からシステム i と ii を比較すると，さほど差は顕著ではないが，正解率 (correct) はシステム ii の方がやや高く，正解精度 (accuracy) はシステム i の方がやや高いことが言える．また，話者による認識性能の差という観点では、NK の講義音声が全体的に認識性能が高い．これは、話者 NK が音響モデル作成話者集合に近い年齢であったためと発声が比較的明瞭であったためと思われ、システムにとって認識が容易／困難な声の特性というものが存在することを示している。

音響モデルの比較という観点からシステム ii と iii を比較すると，学習データが若干異なるので厳密な比較はできないが、トライフォンモデルが音節モデルを上回っている場合の方が多かった。音節モデルのコンテキスト依存化が必要と思われる [9]。

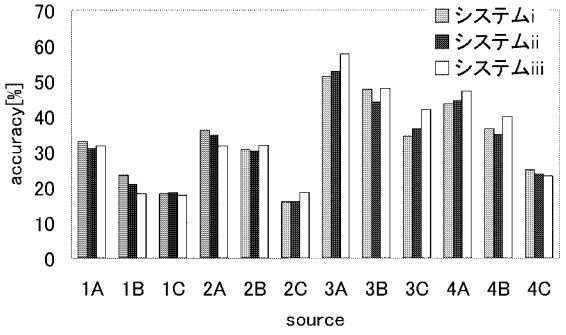

図 1: 認識結果 (accuracy)

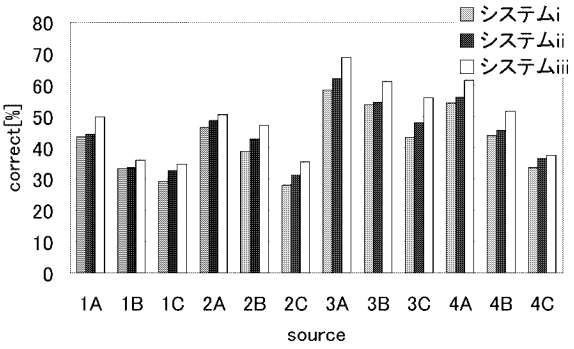

図 2: 認識結果 (correct)

4 要約実験

本節では、音声のポーズで区切られた各区間 (以降「文」と呼ぶ) のうちの重要区間を抽出することによる自動要約を検討する。対象データは 1-1, 1-2 および 3-1, 3-2 であり、設定要約率は文単位換算で 25%である。

4.1 人間による要約

対象講義の分野に造詣が深い 6 人の被験者に対し、重要と思われる文を設定要約率を目安に抽出してもらい、被験者同士の抽出文集合の一致度、および被験者による抽出文集合と、その被験者を除いた 5 人のうち 3 人が重要であると判定した文より構成される文集合 (man3/5) との一致度を計算した。ここで一致度の評価尺度としては、κ 統計量 (κ 値) と F 値を用いた。

被験者同士の一致率の平均が話者 SN の講義 (1-1,1-2) で κ 値 0.378(F 値 0.539)、話者 NK の講義 (3-1,3-2) で κ 値 0.420(F 値 0.543) となり、被験者と man3/5 との一致率の平均が、話者 SN の講義で κ 値 0.477(F 値 0.604)、話者 NK の講義で κ 値 0.490(F 値 0.593) となった。これらを自動要約による重要文抽出の目標値とし、6 人中 3 人が重要であると判定した文集合 (man3/6) を、自動要約に対する正解文とした。なお、CSJ の学会講演の人間による要約の κ 値は 0.407(F 値 0.563)[11] であり、これよりも講義音声の方が要約しやすいということが言える。

4.2 自動要約手法

特徴量には韻律情報、表層的言語情報の両方を用い、またそれらを組み合わせた組合せ実験も行った。ここで用いた特徴量の説明を以下に記す [11]。

tf 各文中の名詞の tf を計算し、スコアの高い文から 25%を抽出する。

頻出単語 出現頻度の高い方から、その語を 2 つ以上含んでいる文の数が全体の 25%になるように語を選び、文を抽出する。

slide-title スライドのタイトルに含まれる名詞が出現する文を重要文として抽出する。

slide-tf スライド中に三回以上現れる名詞を頻出単語とし、その頻出単語が一回以上含まれる文を重要文として抽出する。

F0 基本周波数の高い文の順に 25%を抽出。(スライド情報不使用時、slide-tf の代替)

パワー パワーの強い順に 25%を抽出。(スライド情報不使用時、slide-title の代替)

発話時間長 発話時間の長い文から 25%を抽出

組合せ実験 上記の特徴量に加え、以下の話速の遅い文、発話時間長の短い文といった棄却特徴 (非重要文である特徴) も組み合わせる。**話速の遅い文**は、話速の遅い順から 15%を抽出し、**発話時間長の短い文**は、発話時間の短い文から 15%を抽出する。設定要約率 25%に対する 15%という数字には根拠はなく、単に経験則として定めた。この重要文の指

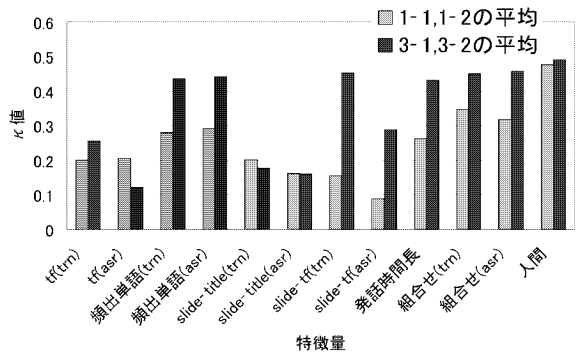

図 3: 要約結果の評価（スライド使用）

表 4: 各特徴量と寄与度

特徴量	スライド情報使用		スライド情報不使用	
	テキスト	音声入力	テキスト	音声入力
tf	0.2	0.4	0.2	0.6
発話時間長	0.4	0.6	0.2	0.2
slide-tf/f0	0.6	0.6	0.0	0.2
slide-title/パワー	0.6	0.6	0.2	0.6
頻出単語	0.6	0.6	0.6	0.6
発話長の短い文	-∞	-∞	-∞	-∞

標特徴 F_k による i 番目の文 S_i に対する重要度の判定結果を、式 (1) のように定義する。

$$Score_{F_k}(S_i) = \begin{cases} 1 & \text{if important} \\ 0 & otherwise \end{cases} \quad (1)$$

棄却可能な特徴 D_p による文 S_i に対する棄却スコア $Score_{D_p}$ も式 (1) と同様に、1 と 0 の 2 値をとる。最終的な文スコアは、式 (2) のように示される。

$$TotalScore(S_i) = \sum_k \alpha_k Score_{F_k}(S_i) + \sum_p \beta_p Score_{D_p}(S_i) \quad (2)$$

ここで、α_k、β_p はそれぞれ特徴 F_k、特徴 D_p の寄与度である。実際の実験では、α_k を 0 から 0.6 まで 0.2 刻み、β_p は 0 または $-\infty$ で組み合わせた。寄与度の推定は 1-1 の要約実験で最も κ 値の高かった組合せを採用することで行い、1-2,3-1,3-2 についても同一の特徴量・寄与度を使用した。

4.3 要約結果

各特徴量による要約結果、ならびに特徴量の組合せによる話者 SN と NK の講義の要約結果を、図 3 に示す。また、使用した特徴量および寄与度を表 4 に示す。図表中の「trn(テキスト)」は人手による書き起こしを、「asr(音声入力)」は音声認識結果をもとにして言語情報の抽出を行ったことを示す。

単独特徴量の中では頻出単語、および発話時間長によるものが性能がよかった。発話時間の長い文が抽出されているので、時間的な要約率は話者 SN で 44%、話者 NK で 50%程度である。頻出単語は本質的には tf と変わらないが、tf が文の長さに影響されやすいのに対し、頻出単語では文の長さにかかわらず、設定した単語が 2 回以上出現する文をすべて同位として抽出している点が効果があったものと思われる。

ただし、表層的言語情報を用いた特徴量の精度は書き起こしの音声認識精度に依存するため (単語頻度などの情報を多用するので、認識結果の誤りがそのまま特徴量の誤差につながる)、検討する必要がある。特徴量の組合せで、話者 NK の講義の書き起こしデータ (テキスト) を用いた要約では、κ 値 0.451(F 値 0.583) が得られ、音声認識結果を用いた結果でも κ 値 0.458(F 値 0.588) と、同等の結果が得られた。これは人間の要約結果の κ 値 0.490(F 値 0.593) と大差ない結果である。話者 NK の講義では、講義内容の余談部分であるかないかが明らかであったためと思われる。一方、話者 SN の講義については、テキストと音声入力の κ 値の差はやや大きく、κ 値で 0.348 − 0.319(F 値で 0.518 − 0.497)、かつ、人間による要約の κ 値 0.477(F 値 0.539) とでは大きな差があった。

スライド情報を使用していない場合の要約実験では、音声入力の場合話者 SN で κ 値 0.273(F 値 0.463)、NK で κ 値 0.425(F 値 0.563) となり、若干精度が低下し、スライド情報が有効であることがわかった。

5 スライド、項目、キーワードと映像のリンク

5.1 スライド項目と映像の対応付け

本講義コンテンツは日立製の EZ プレゼンテーターで作成されており、スライド切り替え時刻と講義音声との対応情報は記録されている。そこで、 本節ではスライド中の箇条書き項目と、それに対応する発話文との対応付けについて述べる. 箇条書き形式となっているを対象とし、そのスライドを用いている時間内の発話文集合を項目に対応するようセグメンテーションする。

5.1.1 セグメンテーション手法

Hearst 法

Hearst 法 [7][8] では、隣接する一定数の単語列から構成されているブロックを設定し、そのブロックとブロックの間の類似度の変化を利用して、類似度が極小値となるブロックを分割位置とする。

ローカル尺度／グローバル尺度に基づく最適分割法 (提案手法)

ある文 j に対して、文 j〜i までのセグメント内の類似度を測り、またスライド項目 k との類似度を考慮した尺度を $s(j:i,k,r)$ とする。この尺度については後述する。$s(j:i,k,r)$ の文ごとの総和 $D(j:i,k)$ を

$$D(j:i,k) = \sum_{r=j}^{i} s(j:i,k,r) \quad (3)$$

のように求める。式 (4) を動的計画法で解くことで、最

適なセグメント境界 j を得ることができる。

$$D_k(i) = \min_j \{D_{k-1}(j-1) + D(j:i,k)\} \qquad (4)$$

ここで，$D_k(i)$ はスライドの第 k 項目までと 1〜i 個の発話文を k 個に分割して対応付けられた最適結果である．k を無視すれば，発話文集合のみを用いてセグメンテーションする方法になる．

$s(j:i,k,r)$ の尺度としては、参照する文と隣接する文、参照する文とスライド中の項目文との余弦類似度の総和をとるローカル尺度と、参照する文、およびスライド項目文の出現単語を要素とした文ベクトルをとり、セグメント (一定数の文を便宜上ひとかたまりにしたもの) の平均ベクトルとの距離を求めるグローバル尺度の二種類 を用いた。グローバルな尺度を用いる場合には、式 (4) の $\min_j$ が $\max_j$ となることに注意されたい。

5.1.2 セグメンテーション結果

前節に示した手法を評価した。評価は、人手で作成したセグメンテーション結果 (正解境界) との誤差を時間 (秒) で測定することで行った。なお、被験者間の境界推定の差はほとんど見られなかった。ベースラインとして、発話文集合を文数で等分割した文均等法と、時間で等分割した時間均等法および，Hearst 法を用いて，提案手法による結果と比較した。

ローカルな尺度とグローバルな尺度いずれを用いた場合にも Hearst 法に及ばず、Hearst 法もベースラインより悪い結果となった。これは，一枚のスライドに関して同じ話題を発話している文集合をセグメンテーションすることの難しさを示している．

5.2 インデキシング

本節では、スライド中のキーワードと発話単語をマッチングし、キーワード単位での対応付けを行う。最終的に、画面上のスライドにある対応付けられたキーワードをクリックすることで、講義ビデオをその単語が発話された位置へジャンプさせることができるようなシステムを構築した。図 4 にシステムのインターフェース画面を示す。画面右側のスライド画面中にある下線を引かれた単語は、インデキシングによって音声認識結果と対応付けられており、クリックすると対応する位置へビデオがジャンプする。また画面右下のキーワード一覧には、音声認識結果に現れたキーワードを時系列順に並べ、スライド中のリンクと同様に対応する位置へビデオをジャンプさせることができる。前者の方法は，認識誤りによりスライド中でキーワードとならない単語が存在する．一方，後者の方法は，音声認識の誤りに頑健であり，必ず頭出しできる利点がある．

5.2.1 インキシング手法

スライド中のキーワードは tf·idf によるスコアが平均値以上の単語とした。idf には、CSJ に含まれる講演データ (テーマ : 音声処理 · 聴覚、男性話者 264 人の講演) を用い、マッチング対象の書き起こしテキストは名詞のみを用いた。本実験の対象とするスライドは、スライド中のキーワードの出現順序が時系列順なもの、すなわち文章や箇条書き文で構成されている 4 枚のスライドを選択した。対応付けには DP マッチングの手法を用いた。

5.2.2 インデキシングの評価結果

インデキシング機能を備えた講義教材のスライド、数分間の講義の視聴を被験者 9 人に 15 分程度実際に体験してもらい、以下の 2 点について 5 段階評価した。また、被験者は当大学の情報工学系に所属する学部 4 年と修士 1 年の学生である。

質問 1 インデックス機能を持った講義教材を便利だと感じたか
1. とても不便である
2. 不便である
3. どちらともいえない
4. 便利である
5. とても便利である

質問 2 スライド中に表示されるリンクによるインデックスと、音声認識結果からの時系列によるインデックス、どちらが便利だと感じたか
1. スライド中のキーワードによるインデックスの方が断然よい
2. どちらかというと、スライド中のキーワードによるインデックスの方がよい
3. どちらとも言えない
4. どちらかと言うと、音声認識結果からのインデックスの方がよい
5. 音声認識結果からのインデックスの方が断然よい

各質問に対し、5 段階評価された結果を表 5 に示す。質問 1 の結果より、インデックス機能が講義教材を利用する際に便利であるという意見が大多数を占め、本システムによる効率的な学習が可能であると考えられる。しかし、抽出されたキーワードが不自然 (「隠れマルコフモデル」というキーワードの際に、「隠れ」と「マルコフ」に分割してインデックスされていたり、さらにキーワードとしては「隠れ」しか抽出できていないなど」) という意見があった。

また、質問 2 の結果より、「スライド中のキーワードによるインデックスの方がよい」という意見が多数を占めた。これは、音声認識結果から抽出されたキーワードと、キーワード一覧の表示の仕方に問題があったためであると考えられる (図 4 の右下のフレーム参照)。また、「同一キーワードがいくつもあり選択に迷った」や、「キーワード一覧が時系列に並んでいることを明確にしてほしい」などという意見があった。tf・idf に代わるキーワード抽出手法の検討と、キーワード一覧の表示の仕方をわかりやすくするなど、ユーザビリティーの面でも改善を考慮すべき事柄は多い。また、音声認識結果を利用して発話箇所を特定する手法に、現在は DP マッチングを使用しているが、認識誤りに脆弱であるという問題があった。システム面でも、より頑健なマッチング手法の検討が必要であるなどの課題が明らかとなった。

図 4: 教材動作画面例

表 5: インデックス機能の評価結果 (人)

質問／回答	1	2	3	4	5
質問 1	0	0	1	5	3
質問 2	2	3	3	1	0

6 結論

本研究では、収録した講義ビデオ・音声をコンテンツ化するための一連の処理—音声認識・要約・セグメンテーション・インデキシングの手法と評価結果を示した。音声認識においては、指向性ハンドマイクとワイヤレスピンマイク (WMV 圧縮音声) で、話者によっては最大 29.1%正解精度に差が現れることがわかった。また、音響モデルや話者による差も無視できないと言える。音声要約では、人間による要約結果の κ 値が平均 0.484 であったのに対し、組み合わせ実験で最もよい結果の平均が 0.389 と、人間による要約に近づくにはまだかなりの課題が残っていることが明らかとなった。セグメンテーション実験では、Hearest 法と提案手法のいずれも、ベースラインより悪い結果となった。インデキシング実験では、講義コンテンツの頭出しに有効なことと，スライド中のキーワードによるインデキシングの使い勝手が良いことが分かった.

謝辞

本実験にご協力いただいた豊橋技術科学大学、音声言語処理研究室の藤井康寿君に感謝の意を表します。

参考文献

[1] 奥村学、久光徹、増山繁、灘波英嗣、福島孝博、中川祐志、渡部聡彦、江原暉将、和田裕二：テキスト自動要約 知的活動支援の基本技術として、情報処理学会誌、Vol.43、No.12、pp1286-1316、2002

[2] 青柳滋己、佐藤孝治、高田敏弘、菅原俊治、尾内理紀夫：映像短縮再生システムの教育映像への適用評価、情報処理学会論文誌、Vol.46、No.5、pp.1297-1305、2005

[3] 堀智織、古井貞熙：音声要約技術の現状とこれから、電子情報通信学会、音声技報、SP2003-174、pp.21-26、2004

[4] C.Chelba, A.Acero : Indexing uncertainty for spoken document search, Eurospeech, pp61-64, 2004

[5] J.Glass, et.al:Analysis and processing of lecture audio data; Preliminary investigations : in HLT-NAACL 2004 Workshop : Interdisciplinary Approaches to Speech Indexing and Retrieval, pp.9-12, 2004

[6] 北岡教英，高橋伸寿，中川聖一，N-best 線形辞書検索と 1-best 近似木構造辞書探索の併用による大語彙連続音声認識，電子情報通信学会論文誌，Vol.87-DII，No.3，pp.799-807，2004

[7] M.Hearst. Texttiling: Segmenting text into multiparagraph subtopic passages: Association for Computational Linguistics, pp. 33-64, 1997

[8] 金寺登、隅田飛鳥、池端孝夫、船田哲男：ビデオ教材作成支援を目的とした講義音声によるシーン分割：電子情報通信学会論文誌、Vol.J88-D1、No.5、pp.977-984、2005

[9] 北岡教英、梁頴、中川聖一：Trigram・4-gram と文脈依存音響モデルを用いた 1 パス大語彙連続認識アルゴリズムとその高精度化： 電子情報通信学会、音声技報、SP2006-16、2006.

[10] 池田太郎, 山本一公, 松本弘, 西谷正信, 宮澤康永：音節連鎖モデルによる大語彙連続音声認識： 第 5 回音声言語シンポジウム、情報処理学会研究報告、2003-SLP-49-26, pp.151-156, 2003.

[11] 小林聡、山口優、中川聖一：表層的言語情報と韻律的情報を用いた講義音声の重要文抽出：自然言語処理、Vol.12, No.5, pp.43-68, 2005