音声言語情報処理 42−10
（２００２．７．１３）

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

山本 夏夫　　緒方 淳　　有木 康雄

龍谷大学 理工学部
〒 520-2194 大津市瀬田大江町横谷 1-5
Tel: 077-543-7427
E-mail: {ymmt, ogata}@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

あらまし 本研究では，講義ビデオの構造化を目的として，連続した一連の講義音声を個々のトピックに自動的に分割する (トピックセグメンテーション) システムの検討を行った．講義は，話題の移り変わりが少なく，話題の境界を判断することが難しい．そこで，講義音声を目次をベースに講義テキストと対応づけることにより，平均 93.7%と高精度にトピックセグメンテーションを行うことができた．これを基にして，目次から対応する講義の部分を見ることができ，また索引語からその語に関する部分の講義を見ることができる学習支援システムを構築した．

キーワード ：講義音声，トピックセグメンテーション，構造化

A Study on Lecture Video Structuring by Topic Segmentation

Natsuo Yamamoto　Jun Ogata　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University
1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan
Tel: +81-77-543-7427
E-mail: {ymmt, ogata}@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

Abstract In this paper, we study on a segmentation method of continuous lecture speech into topics. A lecture has a few changes of subject and it is difficult to judge their boundaries. To solve this problem, we matched a lecture speech with the lecture text based on the table of contents, and obtained the high performance of the topic segmentation with an average of 93.7%. Incorporating this method, we constructed a system where we can see a part of lecture concerning a table of contents, by specifying the chapters or sections, as well as index words by specifying them.

Key words : lecture speech, topic segmentation, structuring

1 はじめに

現在，社会の情報化が進み，様々な情報が大量に蓄積されるようになってきた．しかし，蓄積される情報が増えるにつれ，必要な情報を効率よく取り出すことが難しくなる．これに対し，索引付け [1]-[2]などの研究が進められている．教育の分野においても，蓄積された情報からユーザが必要とする情報を効率よく提供するための研究が進められている [3]-[5].

本研究では，講義音声を用いて，講義ビデオを適切に構造化する手法を検討する．構造化とは，一定の内容を持つ区間に分割する（トピックセグメンテーション）ことをさす．

講義音声データは切れ目なく連続的に構成されているため，大量の講義音声データに対し人手で構造化を行うことは不可能に近い．そこで本研究では，この構造化を自動化することを目的としている． 講義が構造化されていれば，受講者は効率よく受講することができる．例えば，ユーザにとって興味がある部分だけを受講したり，復習の際に理解できていない部分のみを再び受講するといったことが可能となる．

講義音声においては構造化にあたりいくつかの問題がある．まず音声認識の問題である．講義音声は，ニュース音声，講演音声に比べ，自由発話的な要素が多く，音声認識が困難である．次に，話題の移り変わりが少ないことが挙げられる．ニュース音声においては，ある記事と別の記事境界をその話題から容易に判断することができる．また講演音声においても，講演の内容に定型があり，目的，結果などの標識となる単語を抽出することにより境界を判断することができる [2]. しかし講義音声においてはもともと一つの大きな話題で話されており，そのためその中の小さな話題の境界を判断することが難しい．

これらの問題に対し，本報告では以下に挙げる方法を用いた．まず，音声認識の問題として，音響的モデルについては教師なし適応を用いた．また言語モデルについては，講義に附属する講義テキストの索引キーワードを未知語カテゴリに追加した．索引キーワードは，講義テキストの著者により，重要であると判断された単語である．したがって，構造化を目的とする音声認識においては，この索引キーワードが正しく認識されていることが重要である．

次に，話題の移り変わりが少ない問題に関しては，講義テキストの章立てを利用した．講義は，この講義テキストにしたがって進められており，また講義テキストにおいては，講義内の小項目について分割して記述されている．これにより，あらかじめ分割されている講義テキストのトピックに合わせた講義音声のトピックセグメンテーションができ，教科書を併用した学習を支援することができる．

本報告では，まず2章で講義音声の音声認識について述べる．次に3章では索引付けにおける基礎技術であるTF-IDFとベクトル空間モデルについて，4章では，講義テキストとの対応付けによるトピックセグメンテーションについて，5章ではこれらの技術を基にした学習支援システムについて述べる．

2 講義の音声認識

本研究で対象とする講義音声は，自由発話の要素を多く含んでいる．ここでは，上記の音響的な問題に関しては，教師なし適応によって対処する．また，言語モデルに関しては，講義音声の特徴に比較的近いと思われる，学会講演音声から学習されたモデルを用いることにする [6].

ただし，本研究で扱う音声認識システムは，講義データの構造化を目的としており，講義のトランスクリプションを目指すものではない．自動索引付けは，音声認識された単語列のうちの名詞とキーワードを用いて行うため，通常の単語正解率より名詞やキーワードの正解率，特にトピック内容を直接表現するキーワードの正解率を重視している．

2.1 認識システムの構成

本報告では，音声認識システムとして，ワードグラフを中間結果とする2-pass デコーダーを用いている [7]. まず，1st-passで単語 bigram を用いた lexical tree search を行ない，認識結果をもとにワードグラフを作成する．このとき，最もスコアの高い単語にのみ back-off 接続を行う最尤単語 back-off接続を用いることにより，認識精度を落すことなく処理時間を大幅に削減している [7]. 2nd-passでは，ワードグラフに登録された 1st-passの音響尤度と trigram を用いてリスコアリングを行なう．

2.2 キーワードの設定

自動索引付けにおいては，その話題の内容を表したキーワード群の抽出が重要になると考えられる．ここでは，そのテキスト中から，講義の内容を端的に表したキーワード群を抽出することを考える．そこで，講義テキストの索引キーワードリストを，事前に用意したモデルの未知語カテゴリに登録することにした．

講義テキストの索引キーワードリストは，そのテキスト中の特に重要な単語が集められたものであり，自動索引付けを行う際にも，話題内容を表す重要な役割をもっていると考えることができる．

2.3 実験条件

音響モデルとしては，前後の音素環境を考慮したtriphone HMMを用いた．音響モデルの学習には，日本音響学会新聞記事読み上げコーパス(JNAS)のうち，男性話者137名分の21,782発話を用いた．ただし実際の認識では，音響的なミスマッチに対処するために，評価用データに対し、MLLRによる教師なし適応を行った。

言語モデルには，融合研究コーパスを用いて学習された，講演音声認識用言語モデル[6]を用いた．学習講演数は186で， 学習テキストサイズは549,612語，語彙サイズは10,350語である．

また，評価用データとしては，それぞれ約40分間の2つの講義の音声を用いた．構造化のための音声認識では，ある部分の話題の特徴を得るために，入力音声データはある程度の長さの発話に区切れていた方が良い．ここでは，それぞれの講義音声を，ある一定の無音区間を基準に自動的にセグメンテーションした．したがって，セグメンテーションされた一つの発話（以下，発話区間と呼ぶ）は，日本語文として意味的，文法的な区切りには必ずしもなっていない．

2.4 実験結果

以上の条件のもとで認識実験を行った．実験結果を表1，表2に示す．表1は通常の単語認識率を，表2は索引キーワードリストの正解率をそれぞれ表している．lec2はlec1に対して認識率が低くなっているが，これは，lec2の講義音声について，語尾の声量が小さくなってしまうなどはっきりと発生されていない部分が多かったためと考えられる．また，トピックセグメンテーションの際に重要と考えられるキーワードに関しては，5%程度の差があるものの，平均正解率91.12%で共に高い正解率を得ることができた．

表 1: 認識率 (%)

	Correct	Accuracy
lec1	61.7	58.23
lec2	50.5	42.03
average	56.1	50.1

表 2: キーワードの正解率 (%)

	Correct
lec1	93.64
lec2	88.73
average	91.2

3 構造化の概要

講義音声の認識結果は，文章の羅列である．これを一定の意味を持つ区間に区切ることが構造化の目的である．従って，あらかじめある程度細かい区間に機械的に分割しておく必要があり，これを分析区間と呼ぶ．また，各分析区間ごとに単語を基底とする分析区間ベクトルを作成し，これらをベクトル空間上に配置する．

本研究においては，以下のようにして分析区間ベクトルを作成した．まず，形態素解析により名詞を抽出する．これは，名詞が話題をよく表しているからである．なお，形態素解析には茶筅[8]を用いた．

2.3節で示したように，講義音声認識結果は，発話区間毎に出力されている．本報告では，予備実験より10発話区間を分析窓とした．この分析窓で，図1のように，分析区間をずらしながら，分析区間毎に単語の頻度分布を求める．講義音声を分析区間ベクトルの系列で表現し空間上に配置することにより，分析区間ベクトルの余弦によって類似度を求めることができる．

また，音声認識の結果はテキストで出力されるが，その中に出現する単語には，音声認識の誤りや，全体に渡って頻繁に出現する単語など，索引付けにおいて不要な単語も含まれる．従って，単語の重要性を索引付けの重みとするためにTF-IDFを用いた．

以下，これらについて具体的に述べる．

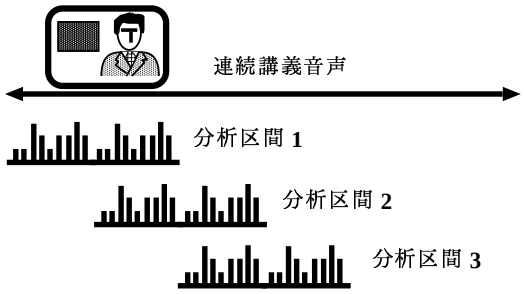

図 1: 分析区間ベクトルの作成過程

3.1 TF-IDF

本研究では，音声認識の結果を基に講義テキストとの対応付けやトピックセグメンテーションを行なう．従って，音声認識結果に含まれる誤認識単語，湧き出し単語などや，形態素解析結果により誤って抽出された単語により，索引付けの精度が低下するという問題点が存在する．また，全分析区間で平均的に出現する単語より，一部の区間で頻繁に出現する単語のほうが，索引付けにおいて有用である．このように，湧き出し単語の除去や，一部の区間に頻出する単語，即ち重要語に重み付けをするため，本研究ではTF-IDFを用いた．

TF-IDFは式(1)で表され，索引語w_iが文書t_kに現れる回数が高いほどTF(Term Frequency)が高くなり，索引語w_iが現れる文書数が少ないほど，IDF(Inverse Document Frequency)が高くなる．したがって，TFは頻度の高い索引語という性質を表し，IDFはその文書に偏って現れる単語という性質を表しているので，TF-IDFの値が大きければ，索引語の特徴をよく表す単語であるといえる．

$$
\begin{aligned}
TF \cdot IDF &= TF(w_i, t_k) \cdot IDF(w_i) \qquad (1)\\
TF(w_i, t_k) &= \text{索引語 } w_i \text{が文書 } t_k \text{に現れる回数}\\
IDF(w_i) &= \log \frac{\text{全文書数}}{\text{索引語 } w_i \text{が現れる文書数}}
\end{aligned}
$$

3.2 ベクトル空間モデル

情報検索の分野では，文書集合から検索質問に適合する文書を検索する際，ベクトル空間モデルが頻繁に用いられる．ベクトル空間モデルについて以下に述べる．

表 3: 単語 × 文書マトリクス A

	d_1	d_2	…	d_d
t_1	a_{11}	a_{12}	…	a_{1n}
t_2	a_{21}	a_{22}	…	a_{2n}
⋮	⋮	⋮	⋱	⋮
t_t	a_{m1}	a_{m2}	…	a_{mn}

表3のような，単語 × 文書の例で，要素が索引語の出現頻度をあらわしている行列 $A_{t\times d}$ を作成する．$A_{t\times d}$ の各列を列ベクトルとみなすと，これらの列ベクトルは文書に対応するベクトルとみなすことができる．ベクトルの次元は全文書に出現する単語の異なり数と等しい．また，頻度を TF-IDF のような重みに置き換えると，文書を索引語の重みベクトル $\boldsymbol{d}_j$ で表すことができる．同様に，検索質問も索引語の重みベクトル $\boldsymbol{q}$ で表現することができる．これらのベクトルの類似度を求めることにより，文書と検索質問の類似度を求めることができる．また，これらのベクトルは，図2のように，ベクトル空間上に配置することができる．よって，類似度 $sim(\boldsymbol{d}_j, \boldsymbol{q})$ は式 (2) で表される余弦により，式 (3) のように表すことができ，これを用いると優れた検索性能が得られることが知られている [9].

$$
\cos\theta = \frac{\boldsymbol{x}^t \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \qquad (2)
$$

$$
\begin{aligned}
sim(\boldsymbol{d}_j, \boldsymbol{q}) &= \frac{{\boldsymbol{d}_j}^t \boldsymbol{q}}{\|\boldsymbol{d}_j\|\|\boldsymbol{q}\|}\\
&= \frac{\sum_{i=1}^{m}(a_{ij}q_i)}{\sqrt{\sum_{i=1}^{m}(a_{ij})^2}\sqrt{\sum_{i=1}^{m}(q_i)^2}} \qquad (3)
\end{aligned}
$$

4 講義テキストとの対応付けによるトピックセグメンテーション

本研究で用いた講義には，それぞれ講義テキストが附属しており，1回の講義に該当する部分が，数節に分けられて書かれている．また，講義は講義テキストにそって進められているので，講義テキストの各節の話題にあわせて分割することができる．また，受講者はこのテキストを用いて学習を進めるので，講義と講義テキストとの対応がとれていれば便利である．例えば，講義テキストで不明な部分については，対応した講義を音声映像で受講することができる．以下に，講義に附属している講義テキストとの対応付けの実験結果について報告する．

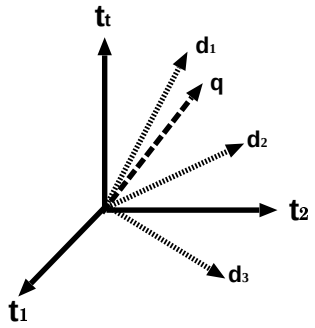

図 2: ベクトル空間モデル

4.1 講義テキストとの対応付けの手順

図3に概略を示す．講義テキストより，各節ごとに講義テキスト分析区間ベクトル（以後これをトピックベクトルと呼ぶ）を作成した．また，本節においては，講義音声より作成された分析区間を講義音声分析区間と呼ぶ．次に，各講義音声分析区間ベクトルごとに，講義テキストにおけるトピックベクトルとの類似度をそれぞれ求め，もっとも類似度が高い講義テキストのトピックベクトルを対応する節とした．しかし，この時点では，講義音声分析区間ベクトルがもつ時系列順

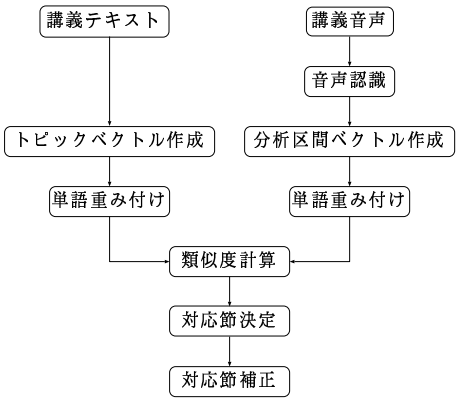

図 3: 講義テキストとの対応付け

序が無視されている．従って，次のような誤った対応付けが生じてしまう．

1. 最初の講義音声分析区間が講義テキストの第2節に対応付けられている．
2. 第2節に対応付けられている講義音声分析区間の直後に，第4節に対応付けられている講義音声分析区間がある．
3. 第3節に対応付けられている講義音声分析区間の直後に，第2節に対応付けられている講義音声分析区間がある．

つまり，第i講義音声分析区間ベクトルが第s_i節に対応付けられていた場合，その次の講義音声分析区間ベクトルは第s_i節または第s_i+1節であるという条件を満たす必要がある．このため，次の法則により$i=1,2,\cdots$と順に対応節の補正処理を行なった．

(1) $i=1$ならば $s_i=1$.

(2) $s_{i+1}=s_i$と$s_{i+1}=s_i+1$のどちらでもないならば，$s_{i+1}=s_i$.（ノイズとみなす）

(3) $s_{i+1}=s_i+1$のとき，s_i+1節がm区間続き，その次にs_i節がn区間続く場合，$m<n$ならば該当するs_i+1節の区間はs_i節.

また，その補正処理の例を図4に示す．

(1) 分析区間 →
4 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3

第1分析区間が講義テキストの第1節以外に対応付けられることはないので，

分析区間 →
1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3

(2) 分析区間 →
1 1 1 1 2 2 2 4 2 2 2 1 3 3 3 3 3

第2節の直後に第4節があったり，第1節があったりすることはあり得ないので，

分析区間 →
1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3

(3) 分析区間 →
1 1 1 1 2 2 2 3 3 2 2 2 3 3 3 3 3

3節が2区間続いた後，2節が3区間続いている．3節の区間より2節の区間の方が長いので，

分析区間 →
1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3

図 4: 補正処理例

4.2 実験条件

2.4節の音声認識結果に対し実験を行なった．形態素解析を行い名詞を抽出した結果を表4に示す．表中の異なり単語数とは，単語数から重複を取り除いたものを指す．また，講義音声に対応するそれぞれの講義テキストは，表5に示す構成である．

表 4: 講義音声中の名詞単語数

	単語数	異なり単語数
lec1	2423	697
lec2	2104	839

表 5: 講義テキスト

	単語数	異なり単語数	章数	節数
lec1	1179	459	4	13
lec2	1180	626	4	10

評価は式(4)の正解率を用い，4.1節で述べた対応節補正前後それぞれで評価した．また，対応する節の正解は講義テキストの節にあわせ，人手で作成した．

$$\text{正解率} = \frac{\text{正しく対応付けられた分析区間の数}}{\text{全分析区間の数}} \quad (4)$$

4.3 実験結果

実験の結果を，表6-表7に示す．表6は，講義テキストの章に対して対応付けを行った場合，表7は，節に対して対応付けを行った場合である．また，表6(b)，表7(b)は，補正処理を行なった結果である．

表 6: 章対応付け結果

(a) 補正なし

	正解率 (%)
lec1	80.2
lec2	88.0
平均	84.1

(b) 補正あり

	正解率 (%)
lec1	98.6
lec2	97.9
平均	98.3

表 7: 節対応付け結果

(a) 補正なし

	正解率 (%)
lec1	60.4
lec2	74.9
平均	67.7

(b) 補正あり

	正解率 (%)
lec1	91.1
lec2	86.9
平均	89.0

以上のように，高い正解率を得ることができた．特に，章への対応付けでは，平均98.3%と高い正解率を

得ることができた．節への対応付けに関して，正解率の平均が 89.0% と下がっているが，1章の単語数より1節の単語数が少ないため，話題の特徴を得るのが難しかったためと考えられる．

5　学習支援システム

以上を基に，図5に示す学習支援のためのシステムを構築した．本システムは，

- ビデオ表示部
- 目次ブラウズ部
- 検索部

で構成される．ビデオ表示部は，指定された講義ビデオの部分が表示される部分である．目次ブラウズ部，検索部については以下に示す．

5.1　目次ブラウズ部

目次ブラウズ部は，講義テキストにおける目次が階層化されて表示される．講義回数，章を選択するとそれに属する下位の部分が表示される．例えば，第1回目の講義を選択すると，章の表示部に第1回目の講義に属する章が表示され，更にその中から，第2章を選択すると，節の表示部に第2章に属する節が表示される．

ユーザは，これらの中から指定することにより，対応するビデオの部分を見ることができる．

5.2　検索部

検索部には，講義テキストの索引語がリストアップされている．また，これらの索引語は，講義テキストの情報を基に対応する節が内部的にリンクされている．したがって，ユーザは任意の索引語を選択することにより，その語が説明されている部分のビデオを見ることができる．

6　おわりに

本報告では，講義ビデオと講義テキストを対応づけることにより，講義ビデオのトピックセグメンテーションを行い，対応節補正後の章，節の平均が 93.7%と高い精度を得ることができた．また，この結果を基に，学習支援のためのシステムを構築した．

今後は，講義ビデオを増やすと共に，講義テキストによらないトピックセグメンテーションを行い，複数の講義間での検索を行えるシステムを構築する予定である．

図 5: システム画面

参考文献

[1] 鷹尾誠一, 緒方淳, 有木康雄: “ニュース音声に対する教師なしトピックセグメンテーションにおける単語重要度決定方法の比較”, 日本音響学会平成11年度春季研究発表会,3-Q-27,pp.171-172,1999.

[2] 長谷川将宏, 秋田祐哉, 河原達也, “談話標識の抽出に基づいた講演音声の自動インデキシング” 音声言語情報処理,SLP36-6,pp.75-82,2001.

[3] 緒方 淳, 山本夏夫，鷹尾誠一，有木康雄: “講義データを対象とした音声認識と構造化の検討”, 情報処理学会研究報告, SLP37-14,pp.79-84,2001.

[4] 伊藤克亘, 藤井敦, 石川徹也 “音声文書検索を用いたオンデマンド講義システム” 音声言語情報処理,SLP39-28,pp.165-170,2001.

[5] 片山薫，香川修見，神谷泰宏，對馬英樹，吉廣卓哉，上林彌彦: “遠隔教育のための柔軟な講義検索手法”, 情報処理学研究会論文誌，Vol.39 No.10, pp.2837-2845,1998.

[6] 加藤, 南条, 河原達也: “講演音声認識のための音響・言語モデルの検討”, 信学技報, SP2000-97, pp.13-18,2000.

[7] 緒方淳, 有木康雄: “back-off接続を考慮した大語彙連続音声認識の高速化”, 音講論集, pp.43-44,2000.

[8] 茶筅:http://chasen.aistnara.ac.jp/index.html.ja

[9] G.Salton and C.Buckley: “Term Weighting Approaches in Automatic Text Retrieval”, Infomation Processing & Management, 24(5),513-523,1988.