2005－SLP－55　(11)
2005／2／5

講義の自動アーカイブ化のためのスライドと発話の対応付け

北出 祐　　河原 達也

京都大学 情報学研究科 知能情報学専攻
〒 606-8501 京都市左京区吉田二本松町
e-mail: kitade@ar.media.kyoto-u.ac.jp

あらまし　講義の自動アーカイブ化を目的として，講義音声の書き起こしの各発話と講義に用いられたスライドを自動的に対応付ける手法を提案する．スライドのテキストからキーワードを抽出し，発話との関連度を定義した．その際に単一のスライドへの対応付けが困難である場合もあるので，複数のスライドからなるトピック単位も構成した．また，発話直前のポーズ長や談話標識の出現に着目し，話題の遷移尤度を定義した．これらの尺度を用いてスライドまたはトピックを状態とするマルコフモデルを構築し，スライドやトピック単位に発話の対応付けを行った．実際の大学の講義に対して自動対応付けの実験を行ったところ，F 値が 0.681 であった．

Automatic Alignment of Speech Transcriptions with Viewgraph Slides for Lecture Archiving

Tasuku Kitade　　Tatsuya Kawahara

School of Informatics, Kyoto University, Kyoto 606-8501, Japan
e-mail: kitade@ar.media.kyoto-u.ac.jp

Abstract　Automatic alignment of speech transcriptions with a sequence of viewgraph slides used in a lecture is addressed. We extracted the keywords from the text of the slides and define the similarity with utterances. Here, we introduce a topic unit consisting of multimple slides for more stable and flexible matching. In addition, we use the information of pauses and discourse markers for defining the likelihood of transition between slides or topics. Based on these measures, we set up a Markov model to be matched with utterances. Experimental results using two lectures confirm the effectiveness of the method.

1 はじめに

近年，高速なネットワーク環境が整備され，音声や映像などの大規模なディジタルコンテンツの閲覧が容易になった．教育分野においても，講義をアーカイブとして蓄積して，これらを教材として復習や遠隔学習に利用する環境が整いつつある．このように蓄積された情報から目的とする情報を検索したり，内容を容易に把握できるようにするためには，インデックスなどの二次情報の付与が極めて重要である．その反面，これらの情報の付与には膨大な人的・時間的コストがかかるため，自動的に作成できる技術が望まれている．

これらの二次情報の自動付与を目的としてニュース音声の索引付け [1] やトピックセグメンテーション [2]，討論番組の話者インデキシング [3][4] などの研究が行われてきた．講義を対象とした研究では，トピックセグメンテーション [5] やシーンの自動分割 [6] が行われている．

我々は以前，学会講演を対象として談話標識に基づいてセクションに分割する方法を提案し，この情報が重要文抽出においても有用であることを示した [7][8]．これに対して近年，講演や講義においてスライドが用いられることが一般的になってきたので，このスライドを基にインデックスを作成することが考えられる [9]．本研究でもこのアプローチを採用する．ただし，本研究で対象とする大学の講義では，スライドを用いた説明以外に，ビデオやホワイトボードを用いた説明や学生とのインタラクションなど，スライドと直接対応しない部分もかなりの割合を占めている．また，学生の様子や状況に応じて以前のスライドに戻って説明を繰り返したり，スライドを飛ばしたりすることも多い．そこで本研究では，スライドとトピックの二つの階層を用意して柔軟な対応付けを行うとともに，これらに対応しない部分の検出も試みる．

2 スライド情報を用いた講義のインデキシング

これまで講義を対象とした自動アーカイブ化システムの例として，映像処理を用いるもの [10] や，スライドの切替えなどを記録するシステム [11]，さらにこれらを併用したもの [12] が提案されている．しかし，映像を用いる場合には撮影者を配備する必要があり，また特殊なシステムを用いる場合には限定された環境でのみしか適用できないという問題がある．

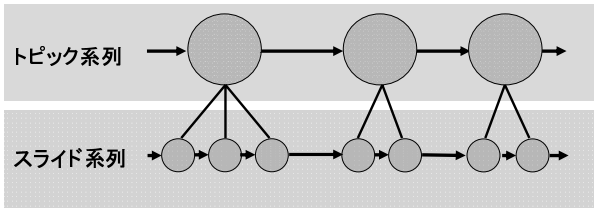

図 1: トピック系列とスライド系列

これに対して本研究で前提とするのは，講義の音声（の書き起こし）とスライドのみ（構成順）で，スライドの提示順序や切替えのタイミングの情報などは一切使用しない．スライドは箇条書きでポイントが記述されており，話題を端的に示すキーワードが得られやすい反面，テキストの量が少なく発話との対応付けが容易でない．特に数式や図のみからなるスライドからはキーワードが取得できない．さらに前述のようにスライドの構成順序通りに講義するのではなく前に戻ったりスキップしたりするため，スライドに直接対応しない発話も多数存在する．

そこで本研究では，複数のスライドから構成されるトピックの単位を用意し，スライドまたはトピックを単位として発話を対応付けることを検討する．また，どのスライド/トピックにも対応付けられない発話は，トピック外の発話として扱うことにする．具体的には，スライドは同一のトピック内でのみ逆戻りしたりスキップしたりすると仮定し，スライドに対応付けられる発話に関してはそのインデックスをふり，対応付けられない発話に関しては，複数のスライドからなるトピックに対してインデックスをふる．

トピックは同一の話題のスライドをグループ化して構成し，スライドのみから構成されるスライド系列とトピックから構成されるトピック系列を用意する（図 1）．そして，スライドからキーワードを抽出し，その抽出されたキーワードを用いて発話との対応付けを行う．

3 マルコフモデルに基づく対応付け

全体の処理の流れを図 2 に示す．

まず構成順に並んだスライドに対して，あらかじ

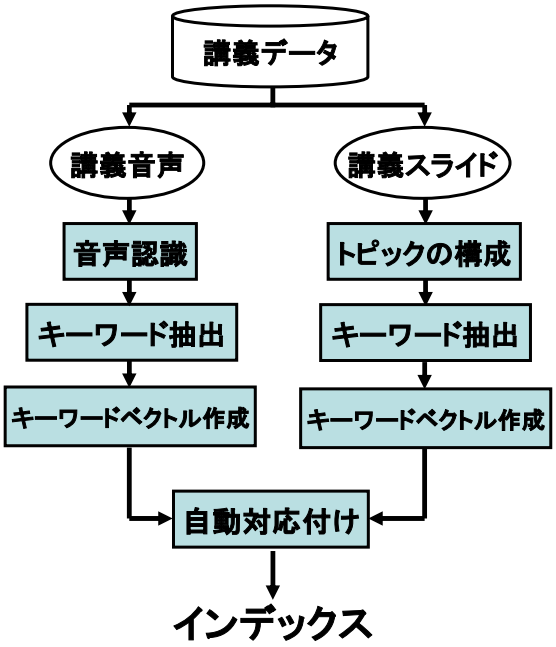

図 2: 講義のインデキシングの処理の流れ

め人手により類似した内容のスライドをまとめあげて，トピックを作成し，トピックの系列も構成する．トピック系列およびスライド系列は構成順に並べた系列である．次にスライドおよび発話からキーワードを抽出する．抽出するキーワードは，品詞が名詞，数詞，記号の単語である．その際，トピック系列においては，そのトピックに含まれるスライド系列のキーワードを用いる．これらのキーワードをもとに，講義の発話をスライドもしくはトピックに対応付ける．

具体的には，各スライドまたはトピックを状態とするマルコフモデルをそれぞれ構築する．その上で，トピックとトピックの境界でスライド系列とトピック系列との間の遷移を許して，ビタビアルゴリズムにより最尤の出力系列を得る（図 3）．すなわち講義の発話をスライドまたはトピックと対応付けた結果を得る．そして，自動対応付けした結果をインデックスとして用いる．

柔軟なマッチングを実現するために、以下のようにマルコフモデルの状態出力尤度および状態遷移尤度を定義する．ここで発話の系列を $U = \{U_1, U_2, U_3, \cdots, U_l\}$，スライド系列を $S = \{S_1, S_2, S_3, \cdots, S_n\}$，トピック系列を $T = \{T_1, T_2, T_3, \cdots, T_m\}$ とする．

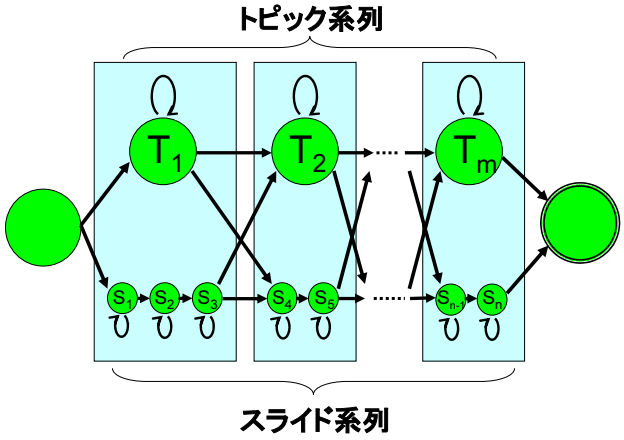

図 3: 対応付けのためのマルコフモデル

3.1 状態出力尤度

講義の大半はスライドに基づいて行われているので，スライドに記されたキーワードを多く含む発話は，そのスライドの内容を説明している可能性が高い．そこで，スライドと各発話との関連度を測る尺度として，コサイン距離（式（1））を用いる．

$$dist(X, Y) = \frac{W(X) \cdot W(Y)}{|W(X)||W(Y)|} \qquad (1)$$

ここで $W(X)$ は，X のキーワードの頻度ベクトルで，キーワード数の次元からなり，各要素に X 中のキーワードの出現頻度が代入されている．

スライドには講義内容を端的に表すキーワードが記されているが，記述されたキーワードは極めて少なく，発話の中に表れるキーワードはさらにその一部であるために，スライドと発話が関連していても全く対応付けられない場合が多数存在する．そこで，以下に述べる 3 パスの方法により状態出力尤度を計算する．

第 1 パス

スライド系列の状態出力尤度は，スライドのテキストと各発話のコサイン距離とする．よって，発話 U_i のスライド S_j に対する尤度 $O(U_i, S_j)$ は以下のようになる．

$$O(U_i, S_j) = dist(U_i, S_j) \qquad (2)$$

トピック系列においては，トピック T_k に含まれるすべてのスライド $S_j (S_j \in T_k)$ と発話 U_i とのコサイン距離（から一定値 λ を引いた値），またはトピック T_k に含まれるすべてのスライドのテキストと発

話 U_i のコサイン距離の最大値を状態出力尤度とする（式（3））.

$$O(U_i, T_k) = \max\{dist(U_i, S_j) - \lambda, dist(U_i, T_k)\} \tag{3}$$

第 2 パス

第 1 パスにおいて，スライドに対応付けられた発話集合も含めたテキストを構成してコサイン距離を求める．具体的には、スライド S_j（またはトピック T_k），および第 1 パスにおいてスライド S_j（またはトピック T_k）に対応付けられた発話 U_i の集合 S'_j（トピックについても同様に T'_k）とのコサイン距離を計算する（式（4））.

スライド系列
$O(U_i, S_j) = \quad dist(U_i, S'_j)$
トピック系列
$O(U_i, T_k) = \quad \max\{dist(U_i, S'_j) - \lambda, dist(U_i, T'_k)\}$ (4)

第 3 パス

第 2 パスで得られた結果を用いて，トピック外の発話を検出する．具体的には，前後 2 発話 $\{U_{i-2}, \cdots, U_{i+2}\}$ の状態出力尤度に基づく平滑化を行い（式（5））, その値 $O'(U_i, S_j)$ がしきい値 δ 以下の発話をトピック外の発話とする．

$$\begin{aligned} O'(U_i, S_j) \quad = \quad & \frac{1}{5}\{O(U_i, S_j) \\ & + \frac{1}{2}(O(U_{i+1}, S_j) + O(U_{i-1}, S_j)) \\ & + \frac{1}{4}(O(U_{i+2}, S_j) + O(U_{i-2}, S_j))\} \end{aligned} \tag{5}$$

3.2 状態遷移尤度

次に，ある発話において状態が遷移する尤度を定義する．スライドやトピックの転換点では，スライドを切り替えたり，少し間をとったりするために，他の箇所よりも比較的長めのポーズが挿入されると考えられる．そこで発話の直前のポーズ長の z-score（式（6））による尺度を定義する．

$$\overline{pause}(U_i) = \frac{pause(U_i) - \mu_{pause}}{2\sigma_{pause}} \tag{6}$$

ここで $pause(U_i)$ は発話 U_i の直前のポーズ長，μ_{pause} は当該講義の平均ポーズ長，σ_p^2 は分散である．極端にポーズが長い場合は，$\overline{pause}(U_i) = 1$ とした．

これに加えて談話標識に基づく統計量も導入した．ここで談話標識は話題の転換点に用いられる特徴的な表現である．談話標識は『日本語話し言葉コーパス』の学会講演 889 講演から以下のように（教師なしで）学習した [7].

（1）学習データからセクション境界候補，具体的には，各講演において平均ポーズ長以上のポーズが挿入された箇所を抽出する．

（2）（1）で得られたセクション境界候補からそれぞれ冒頭の一文を抽出し，セクション冒頭の文集合を得る．

（3）セクション冒頭の文集合に特徴的に出現する単語を談話標識として抽出する．具体的には各単語 m について，以下の式（8）より単語頻度 wf_m と文頻度 sf_m に基づく統計値 DM_m を求める．

$$DM_m = wf_m * \log\left(\frac{N_s}{sf_m}\right) \tag{7}$$

単語頻度 wf_m は，セクション冒頭の文集合において単語 m が出現する回数であり，文頻度 sf_m は，学習データの全ての文（総数 N_s）で単語 m が出現する文の数である．

この談話標識の統計量を利用して，スライドやトピックの切り替えの尺度 $DM(U_i)$ を以下のように定義する．ただし，これもポーズ長と同様に z-score による正規化を行う（$\overline{DM}(U_i)$）.

$$DM(U_i) = \sum_{m \in U_i} DM_m \tag{8}$$

ポーズ長に基づく尺度と談話標識に基づく尺度はともに話題の境界らしさを表す尺度であるので，重み付き和によりスライド S_j（またはトピック T_k）から次のスライド S_{j+1}（またはトピック T_{k+1}）に遷移する尤度 $a_{j,j+1}(U_i)$ を定義する．同様に，同一のスライド /トピックにとどまる尤度 $a_{j,j}(U_i)$ も定義する．

$$\begin{cases} a_{j,j+1}(U_i) \quad = \quad \frac{\beta}{1+\alpha}(\overline{pause}(U_i) + \alpha * \overline{DM}(U_i)) \\ a_{j,j}(U_i) \quad = \quad -\frac{\beta}{1+\alpha}(\overline{pause}(U_i) + \alpha * \overline{DM}(U_i)) \end{cases} \tag{9}$$

表 1: 実験データ

ID	MU1126	PA0707
時間 (min)	78.8	75.3
発話数	437	648
（トピック外の発話）	(22)	(284)
総単語数	14522	14741
スライド数	40	23
トピック数	16	7
キーワード総数	756	572
（異なりキーワード数）	(228)	(175)

3.3 最尤出力系列の導出

以上により定義されたマルコフモデルに基づいて発話系列に対する最尤出力系列をビタビアルゴリズムによりスライドとの対応付けを行う．

冒頭の発話をスライド S_j およびトピック T_k に対応付ける二つの系列を並行に実行して出力系列を得る．そして，最終的に尤度の高い出力系列を最終出力結果とする．

4 自動対応付けの評価実験

4.1 実験データ

予備実験に用いたのは，京都大学で行われた二つの講義である．そのデータの概要を表 1 に示す．人手によってまとめられたトピックの数は，スライドの数のおよそ 30〜40%程度である．

本実験においては，状態出力尤度での式 (3),(4) における λ を 0.1，平滑化の閾値 δ を MU1126 においては 0.04，PA0707 においては 0.08 に，式 (9) におけるパラメータ α を 1，β を 0.001 に，事後的に決定している．

4.2 評価方法

講義の（人手による）書き起こしを対象に発話単位にあらかじめ人手により以下のいずれかにタグ付けしておき，正解として用いる．

- スライド番号/トピック番号
- トピック外の発話

表 2: 自動対応付け結果

		再現率	適合率	F 値
MU1126	スライド系列のみ	0.723	0.711	0.717
	トピック系列のみ	0.773	0.770	0.772
	系列間の遷移あり	0.735	0.724	0.730
PA0707	スライド系列のみ	0.615	0.458	0.525
	トピック系列のみ	0.621	0.457	0.526
	系列間の遷移あり	0.742	0.554	0.635

評価には，その正解数に基づく再現率（recall），適合率（precision），F 値（F-measure）を算出しその平均で評価を行う．ここで，F 値は以下の式 (10) で求められる．

$$F - measure = \frac{2 * recall * precision}{recall + precision} \qquad (10)$$

正解にスライド番号が与えられていたものの，自動対応付けした結果がトピックに割り当てられた場合には，発話に対応する正解のスライドがそのトピックに含まれていた場合に正解と一致したとする．

4.3 実験結果

スライドと発話の自動対応付けの結果を表 2 に示す．スライド系列とトピック系列との間の遷移を許さず，それぞれスライド系列，トピック系列のみで行った結果もあわせて示す．

提案手法により対応付けを行った結果，全発話のおよそ 7 割の再現率が得られ，F 値は 0.681 であった．トピック系列のみの場合は，スライド系列のみの場合よりも対応付けが容易になり，同等もしくは高い精度が得られた．また，スライド/トピック系列間の遷移を認めた方が，スライド系列のみで行うよりも精度が高くなったが，PA0707 においては，トピック系列のみで行った場合よりも高くなった．これは，トピック外の発話が多く，トピック系列のみではかえってその検出に失敗することが多くなったためと考えられる．

提案手法によりスライドに対応付けられた発話の正解精度を表 3 に示す．正解のスライドとの一致精度（適合率）は，およそ 55%であった．

最後にトピック外の発話の検出精度を表 4 に示す．トピック外の発話の割合が，MU1126 では全体の 5%程度であるのに対して PA0707 は約 44%とかなりの差がある．正解のトピック外の発話数と自動判別さ

表 3: スライドに対応付けられた発話数

講義 ID	適合率 (正解一致数/スライド判別数)
MU1126	0.714 (287/402)
PA0707	0.417 (149/357)

表 4: トピック外の発話の検出結果

講義 ID	再現率	適合率	F 値
MU1126	0.455	0.625	0.526
PA0707	0.401	0.708	0.512

れた発話数とは大きな隔たりがあるため，MU1126 は適合率が，PA0707 は再現率が大きく低下している．トピック外の発話の多い講義（PA0707）においては，その検出精度がスライド/トピックへの対応付けの精度（適合率）に大きな影響を及ぼすことがわかる（表 2）．

5 結論

本研究では，講義の自動アーカイブ化のためにスライドと発話の対応付けの方法を提案した．スライドのキーワードをもとに，発話とのコサイン距離を計算し，両者の関連度を定義した．また，状態遷移尤度としてポーズ長と談話標識に基づく尺度も導入した．これらの尺度に基づいたマルコフモデルによって発話との対応付けを行った．より柔軟な対応付けを行うために，複数のスライドから構成されるトピックを用意し，またトピック外の発話の検出も試みた．今後は，実験データを増やし，音声認識結果に対しても評価を行う予定である．

参考文献

[1] 櫻井光康, 有木康雄. キーワードスポッティングによるニュース音声の索引付けと分類. 電子情報通信学会技術研究報告, SP96-66, 1996.

[2] 鷹尾誠一, 緒方淳, 有木康雄. ニュース音声に対するトピックスセグメンテーションと分類. 電子情報通信学会技術研究報告, SP98-103, 1998.

[3] 西田昌史, 秋田祐哉, 河原達也. 統計的話者モデル選択に基づく討論音声の教師なし話者インデキシング. 電子情報通信学会技術研究報告, SP2002-157, 2002.

[4] 秋田祐哉, 河原達也. 多数話者モデルを用いた討論音声の教師なし話者インデキシング. 電子情報通信学会論文誌, Vol. J87-DII, No. 2, pp. 495–503, 2004.

[5] 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. 情処学研報, 2002-SLP-42-10, 2002.

[6] N. Kanedera, A. Sumida, T. Ikehata, and T. Funada. Subtopic Segmentation in the Lecture Speech. In *Proc. ICSLP*, Vol. III, pp. 1821–1824, 2004.

[7] 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. 情処学論, Vol. 43, No. 7, pp. 2222–2229, 2002.

[8] T. Kawahara, M. Hasegawa, K. Shitaoka, T. Kitade, and H. Nanjo. Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers. *IEEE Trans. Speech &Audio Process.*, Vol. 12, No. 4, pp. 409–419, 2004.

[9] 河原達也, 石塚健太郎, 堂下修司. 発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化. 情報処理学会論文誌, Vol. 40, No. 4, pp. 1491–1498, 1999.

[10] 金出武雄, 佐藤真一. Informedia: CMU ディジタルビデオライブラリプロジェクト. Vol. 37, No. 9, pp. 841–847, 1996.

[11] 片山薫, 香川修見, 神谷康宏, 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. Vol. 39, No. 10, pp. 2837–2845, 1998.

[12] 丸谷宜史, 西口敏司, 角所考, 美濃導彦. 教材指示情報を付与した講義コンテンツの作成. 画像の認識・理解シンポジウム (MIRU2004) 論文集 II, pp. 323–328, 2004.