音声言語情報処理 37−14
（２００１．７．１４）

講義データを対象とした音声認識と構造化の検討

緒方 淳　　山本 夏夫　　鷹尾 誠一　　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: {ogata,ymmt,tail}@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

あらまし 本研究では，講義データ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．トピックセグメンテーションを精度良く行うためには，トピックの内容を表すキーワードを抽出することが重要となる．講義テキストの索引キーワードリストを，事前に用意した言語モデルの未知語カテゴリに登録することで，高精度にトピック内容を表すキーワードを抽出することができた．実際に，ある講義に対して，トピックセグメンテーションを行ったところ，F-measureは最大で49.7であった．

キーワード：大語彙連続音声認識，講義音声，トピックセグメンテーション，単語重要度

A Study on Speech Recognition and Structuring for Lectures

Jun Ogata　　Natsuo Yamamoto　　Seiichi Takao　　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: {ogata,ymmt,tail}@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 the topics. In the topic segmentation, extraction of topic words (keywords) is important. We selected the keywords from indices of lecture text and added them as a category of unknown-word in language model. As a result, the keywords were recognized accuratelly and we achived a F-measure of 49.7% in the topic segmentation experiments.

Key words : LVCSR, lecture speech, topic segmentation, word importance

1 はじめに

我々はこれまでに，放送ニュースデータの構造化，データベース化を目的として，トピックセグメンテーション[1][2]やトピック検索[3]，パッセージ検索[4]，クロスメディア検索[5]，クロスメディア・パッセージ検索[6]，ブラウジング検索[7]などの検討を行ってきた．その中でも，トピックセグメンテーションは，構造化やデータベース化の基本となる技術であり，ニュースや座談会など話題が連続的に移り変わっていく状況の中で話題の移り変わりを自動検出して，映像や音声を分割する技術である．従来のトピックセグメンテーション[8]-[10]では，各話題に関する情報を得るために，予め大量の学習データが必要であった．時々刻々と変化していくニュースデータに対してトピックセグメンテーションを行うためには，その度それに適した学習データでモデルを構築しなければならない．しかし，これは現実的には不可能で，テストデータと学習データの時期差が生じてしまう．この問題に対して，我々は，事前の学習を必要としない教師なしトピックセグメンテーションを提案し，放送ニュースデータに対してその有効性を示した[1][2].

本研究では，放送ニュースでの上記の技術をもとにして，講義データの構造化について検討している．講義などのビデオ教材は，放送ニュース同様，内容的な区切れ目がなく連続的に構成されているため，人手でキーワードやインデックス，要約を打ち込むことは不可能に近い．本研究では，この講義ビデオ教材を対象として，目次やインデックス，更に内容を表現する記述や要約を，自動的に作成することを目的としている．このように講義ビデオ教材をハイパーメディア化することによって，コンテンツをブラウジングしたり，分類や検索を行うことが可能となる．

本稿では，講義データとして，比較的安易に，かつ大量に入手できることから，放送大学の講義音声を用いることにし，上記システムを作成するうえで最も基本的な技術となる，トピックセグメンテーションについて検討する．

2 講義音声データの構造化

我々はこれまで，放送ニュースデータの構造化を目的として，記事単位への自動分割(トピックセグメンテーション)や記事検索について検討を行ってきた．

本研究では，対象を講義データとし，その構造化について検討している．具体的には，一方向的で退屈しがちな一連の講義ビデオ教材をディジタル化し，内容記述とハイパーメディア化を行い，ユーザー(学生)が関心のある事例に対して，個人個人のレベルに合わせて検索できるような，マルチメディア教材(システム)の構築を目指している．本稿では，構造化の1手段としてトピックセグメンテーションについて検討する．一連の連続音声を内容毎に自動的に分割しておくことは，上記システムを構築するうえで，最も基本となる技術と考えられる．

講義データとしては，放送大学の講義を用いている．放送大学の講義は，1講義あたり約40分で構成され，ほとんどの場合，一つの講義は一人の講師によって進められている．我々がこれまで報告してきた放送ニュースに比べると，見た目の状況は良く似ているが，喋り方の訓練がなされたアナウンサーに比べると，講師はかなり自由で自発的な発声をしている．そのためトピックセグメンテーションのための音声認識が非常に困難になる．

3 講義の音声認識

日常会話などにおける純粋な自由発話は，発話スタイルとして非常にくだけており，その認識はもちろん，学習データの収集や作成が非常に困難である．それに対して，講演音声は，自由な発話スタイルであるが，ある程度丁寧に発声されており，書き言葉と話し言葉の中間的な位置付けであると考えられる．しかし，読み上げ音声と比べると，その発話スタイルにはかなりのギャップがあり，読み上げ音声ベースの音響モデルと発話スタイルのほぼ一致した音響モデルとでは，その認識精度の差は歴然としている[11][12].

本研究で対象とする放送大学講義音声も，学会講演と同様の特徴を有しており，自由発話の要素を多く含んでいる[13]．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，話し言葉の言い回しを含めており，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[11].

ただし，本稿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．後述するトピックセグメンテーションは，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．

3.1 認識システムの構成

本研究では，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[14]．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[14]. 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.

3.2 キーワードの設定

後述するトピックセグメンテーションにおいては，そのトピック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．放送大学の講義に関しては，個々の一連の講義科目に対して，その講義内容に合致したテキストが用意されている．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．言語モデルの適応法として，その講義テキストからモデルを作成し，事前に用意した講演音声認識用の言語モデルと混合することも考えられるが，本稿では簡便さを考慮し，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，トピックセグメンテーションを行う際にも，トピック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.

3.3 実験条件

3.3.1 音響モデル

音響モデルとしては，前後の音素環境を考慮したtri-phone HMMを用いた．音響モデルの学習には，まずA-TR連続音声データベースa～jセットから6名分のデータとその視察ラベルを用いて初期モデルを作成した．次に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21782発話を用いて連結学習を行なった.

3.3.2 言語モデル

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[11]を用いた．学習講演数は186で， 学習テキストサイズは549612語，語彙サイズは10350語である.

3.3.3 評価用データ

評価用データとしては，ある一人の教師の一つの講義40分間の音声を用いた．音声認識する際は，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，それぞれの講演データを，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.

また，3.2節で述べたように，実験で用いる評価用データに対する講義テキストから索引キーワードリストを全て抜き出し，言語モデルの未知語カテゴリに登録した．本実験で用いた評価用データに対するテキストからは，259種類のキーワードが得られた．259種類のキーワードのうち，評価用データ中に実際に発話されたキーワードは26種類であり，繰り返しを含めるとこの26 種類のキーワードが220回出現していた.

3.4 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1-3に示す．表1は講義テキストの索引キーワードリストの登録を行わかった場合の結果を，表2は索引キーワードリストの登録を行った場合の結果を，表3は表2の結果を用いて音響モデルの教師なし適応を行った場合の結果をそれぞれ示している．また，それぞれの表中の” 単語” は通常の単語認識率を，” キーワード” 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.

実験結果より，索引キーワードを登録することによって，約4%の精度向上が見られる．また，音響モデルの教師なし適応を行うことによって，単語正解精度で約6%，キーワード正解精度で約15%向上していることから，MLLRによる適応が，発話スタイル，話者性の違いに非常に有効に働いているといえる．最終的には，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，正解率93%で高精度に認識できていた.

表 1: 認識率 (キーワード登録なし，音響モデル適応なし)

	Correct	Accuracy
単語	53.65	48.65
キーワード	-	-

表 2: 認識率 (キーワード登録あり，音響モデル適応なし)

	Correct	Accuracy
単語	56.29	52.80
キーワード	87.73	50.91

4 教師なしトピックセグメンテーション

文献[8]-[9]のトピックセグメンテーションは，学習データをN個の異なるトピックに分類し，単語とトピックの

表 3: 認識率 (キーワード登録あり，音響モデル適応あり)

	Correct	Accuracy
単語	61.72	58.23
キーワード	93.64	66.36

関連度をユニグラム確率で表現することにより，モデルを構築している．このため，テストデータに存在しているトピックが学習データに存在しなかった場合，単語とトピックの関連度であるユニグラム確率を正確に求めることができない．この問題点を解決するために，我々は学習データを持つことなしに，トピックセグメンテーションする手法を提案している [1]-[2]．この手法は，連続ニュース音声において，分析区間をずらしながら，隣接する分析区間との類似度を求め，大きく類似度が下がった地点をトピックの境界とする手法である．トピックが何であるかを明示的に学習しないことから，この方法を教師なしトピックセグメンテーションと呼ぶ．

4.1 教師なしトピックセグメンテーションの手順

教師なしトピックセグメンテーションの手順を，以下に述べる．

1. 講義データを音声認識し，名詞とキーワードを抽出する．
2. 分析区間をずらしながら，分析区間毎に重要単語を決定し，トピックベクトル (作成方法は 4.2参照) を作成する．
3. 隣接する分析区間のトピックベクトル同士の類似度を求める．
4. 隣接した分析区間で求めた類似度が閾値以下となった部分でトピックが変わっているとし，そこをトピックの境界とする．

4.2 トピックベクトルの作成方法

トピックベクトルを作成する方法について述べる．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．次に，単語の頻度分布をもとに単語重要度を決定する．単語重要度の計算方法としては，我々がこれまでに提案した，TF-IDFを考慮した相互情報量を用いている [3]．これは式 (1) のように表され，音声認識の湧き出し誤り単語などに対して頑健な方法であることがわかっている [2]．

$$
\begin{aligned}
& i(t_k; w_i) \times TF - IDF \\
&= (i(t_k) - i(t_k|w_i)) \cdot TF(w_i, t_k) \cdot IDF(w_i) \\
&= \left(\log \frac{P(t_k, w_i)}{P(t_k)P(w_i)} \right) \cdot TF(w_i, t_k) \cdot IDF(w_i) \quad (1)
\end{aligned}
$$

最後に，重要度を閾値処理して，閾値より高い重要度を持つ単語だけでベクトルを作る．このベクトルをトピックベクトルと呼ぶ．

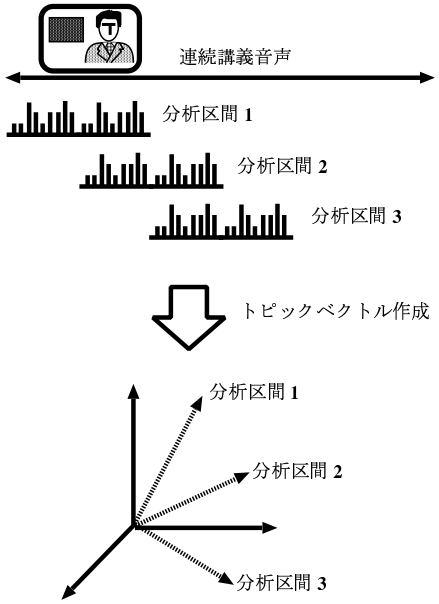

図 1: トピックベクトルの作成過程

4.3 トピックベクトル間の類似度

それぞれの分析区間は，前節で抽出された重要な単語を成分とするトピックベクトルで表されている．従って，トピックベクトル同士の内積は，トピックベクトル同士の単語のオーバーラップ率を表し，トピックベクトル間の類似度を表している．

今，分析区間 t_k における正規化されたトピックベクトルを次のように表す．

$$
\begin{aligned}
X_k &= (x_{1k}, x_{2k}, \cdots, x_{nk})^T \\
&= (x_{1ko}, x_{2ko}, \cdots, x_{nko}, x_{1kc}, x_{2kc}, \cdots, x_{nkc})^T \\
x_{nko} &: \text{分析区間 } t_k \text{ のみに出現した単語の重要度} \\
x_{nkc} &: \text{分析区間 } t_k \text{ と分析区間 } t_l \text{ に共通に出現した単語の重要度}
\end{aligned}
$$

分析区間 t_l における正規化されたトピックベクトルを次のように表す．

$$
X_l = (x_{1l}, x_{2l}, \cdots, x_{nl})^T
$$

$$= (x_{1lo}, x_{2lo}, \cdots, x_{nlo}, x_{1kc}, x_{2kc}, \cdots, x_{nkc})^T$$

x_{nlo} : 分析区間t_lのみに出現した単語の重要度

x_{nkc} : 分析区間t_kと分析区間t_lに共通に出現した単語の重要度

トピックベクトルX_kとX_lの類似度は次のように表される．

$$\begin{aligned}
\cos\theta &= (X_k, X_l) \\
&= (x_{1k}, x_{2k}, \cdots, x_{nk})(x_{1l}, x_{2l}, \cdots, x_{nl})^T \\
&= (x_{1ko}, x_{2ko}, \cdots, x_{nko}, x_{1kc}, x_{2kc}, \cdots, x_{nkc}) \\
&\quad (x_{1lo}, x_{2lo}, \cdots, x_{nlo}, x_{1kc}, x_{2kc}, \cdots, x_{nkc})^T \\
&= (x_{1kc}, x_{2kc}, \cdots, x_{nkc})(x_{1lc}, x_{2lc}, \cdots, x_{nlc})^T \\
&= \sum_i x_{ikc} \cdot x_{ilc} \qquad (2)
\end{aligned}$$

$\cos\theta$が1に近ければ近いほど，トピックベクトル間の類似度は高くなる．

4.4 隣接した分析区間における類似度

式(2)を使って，隣接した分析区間において類似度を求めると，図2のような関数(以下，類似度関数と呼ぶ)として表される．トピックの境界は，この類似度関数の値がある閾値以下になった時点であり，図2ではトピックの境界として表されている．

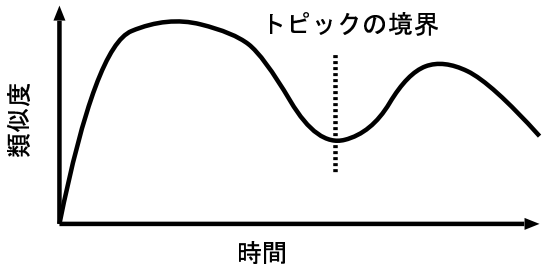

図 2: 類似度関数

5 トピックセグメンテーション実験

5.1 実験条件

表3に示した音声認識結果に対して，トピックセグメンテーションの実験を行った．評価方法には，式(3)に示される再現率と式(4)に示される適合率を統合して評価することができるF-measureを用いた．F-measureは，式(5)に示されるように，再現率と適合率の調和平均を表す．また，正解のトピックの境界の決定は人手で行い，トピックがある程度変化したと思われる発話終端を検出した．今回用いた講義データに対するトピック境界は全部で16個存在した．

$$\text{再現率} = \frac{\text{正しく検出されたトピックの境界の数}}{\text{人手で検出されたトピックの境界の数}} \qquad (3)$$

$$\text{適合率} = \frac{\text{正しく検出されたトピックの境界の数}}{\text{コンピューターが検出したトピックの境界の数}} \qquad (4)$$

$$F-measure = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \qquad (5)$$

5.2 実験結果

実験結果を表4に示す．表中，”Text”は評価用講義音声データを人手により書き起こしたものに対する実験結果を，”LVCSR”は音声認識結果に対する実験結果を表している．表中の値は，予備実験においてF-measureが最大になるように設定した閾値(分析区間長，単語重要度，単純類似度の$\cos\theta$)を用いたときの結果である．

表 4: トピックセグメンテーション結果(境界完全一致)

	Recall(%)	Precision(%)	F-measure
Text	31.3	14.3	19.6
LVCSR	50.0	11.1	18.2

表4の結果より，文献[2]にあるような放送ニュースに対するトピックセグメンテーションと比較して，F-measureの値は非常に低い結果となった．この原因としては，まず，実際のデータのトピック境界の曖昧性が挙げられる．放送ニュースの場合，一つのトピックは一つの”記事”という大まかな内容の移り変わりが存在するのに対して，今回用いた講義に関しては，トピックの移り変わりは内容的に微妙な変化であり，いわゆるサブトピックのセグメンテーションを行っていることになっている．

また，3.3.3節で述べたように，本研究では発話単位へのセグメンテーションを一定の無音を基準に行っており，そのため，1，2 単語で区切れている発話も多くあった．すなわち，一つの発話が日本語としての文章の区切りになっていないため，トピックセグメンテーションを行った際に，必要以上の過分割を引き起こしていた．そこで，トピックの境界が完全一致する条件ではなく，ある程度幅を持たした条件(正解の境界から前後1発話以内，または前後2発話以内は正解とする)での結果を求めた．表5，6にそれぞれ，正解の境界から前後1発話以内を正解とした場合，2発話以内を正解とした場合の結果を示す．

表5，6 の結果より，トピック境界が完全一致の条件(表4)より，F-measureが大きく向上していることがわかる．実際の応用として，セグメンテーションされたト

ピックの検索を考えた場合，検索対象のトピックの前後に余計な短い発話が 1つ，2 つ程あったとしても，利用者にとって大きな問題にはならないと考えられる．

また，全体的な結果として，書き起こしテキストを用いた場合と音声認識結果を用いた場合とでは，それほど精度の差はなかった．これは，音声認識を行った際の全体の単語認識率は低い値であったが，トピックセグメンテーションの際に重要な要素となるキーワードに関しては，比較的高い割合で認識できていたためと考えられる．

表 5: トピックセグメンテーション結果(正解の境界から前後 1 発話以内)

	Recall(%)	Precision(%)	F-measure
Text	43.8	40.0	41.8
LVCSR	56.3	30.6	39.6

表 6: トピックセグメンテーション結果(正解の境界から前後 2 発話以内)

	Recall(%)	Precision(%)	F-measure
Text	43.8	57.1	49.6
LVCSR	56.3	44.4	49.7

6 おわりに

本研究では，講義データの構造化を目指して，音声認識を用いたトピックセグメンテーションに関する検討を行った．従来，行われてきた放送ニュースにおけるトピックセグメンテーションに比べると，トピック境界の内容的な曖昧性の問題，発話単位へのセグメンテーションの問題等により，著しくセグメンテーション精度が劣化することがわかった．

今後は，以上の問題の解決を目指すとともに，単語間の類似度も考慮したトピックセグメンテーション法 [2] の適用なども行う予定である．また，今回は少量のデータでの結果であるため，今後は大量の講義データに対して実験を行う必要がある．

謝辞

本研究では，音声認識用の言語モデルとして,「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクトにおいて作成されている話し言葉コーパス試用版中のモデルを使用させて頂いた．開発された方々に感謝致します．

参考文献

[1] S.Takao, J.Ogata and Y.Ariki: “Topic Segmentation of News Speech Using Word Similarity”, *Proc. ACM'00*, pp.442-444(2000).

[2] 鷹尾誠一,有木康雄,緒方淳: “テロップ出現区間と音声認識の統合によるトピックセグメンテーション”, 人工知能学会研究会資, **SIG-CII**-00, pp.41-46, 2000-11.

[3] S.Takao, J.Ogata, and Y.Ariki: “Study on New Term Weighting Method and New Vector Space Model Based on Word Space in Spoken Document Retrieval”, RIAO00, Volume I, pp.116-131 (2000.4.12).

[4] 鷹尾誠一, 緒方淳, 有木康雄: “ニュース音声に対するパッセージ検索法の比較”, 日本音響学会, 平成 12 年度秋季研究発表会発表会, 2-Q-5, pp.139-140(2000).

[5] S.Takao, J.Ogata and Y.Ariki: “Expanded Vector Space Model based on Word Space in Cross Media Retrieval of News Speech Data”, *Proc. ICSLP'00*, Volume II, pp.1085-1088(2000).

[6] 鷹尾誠一, 有木康雄, 緒方淳: “テロップやフリップ文字を検索質問とした発話文書に対する検索方式”, 第 6 回知能情報メディアシンポジウム, pp.87-88(2000).

[7] 鷹尾誠一, 緒方淳, 有木康雄: “ニュース音声記事データベースにおける観点の自動抽出と構造化”, 信学技報, DE00-12, pp.89-96(2000).

[8] J.P.Yamron, I.Carp, L.Gillick, S.Lowe, and P.van Mulbregt: “A Hidden Markov Model Approach to Text Segmentation and Event Tracking”, ICASSP98, Volume I, pp.333-336, 1998.

[9] P.van Mulbregt, I.Carp, L.Gillick, S.Lowe and J.Yamron: “Text Segmentation and Topic Tracking on Broadcast News Via A Hidden Markov Model Approach”, ICSLP98, Volume VI, pp.2519-2522, 1998.

[10] Jay M.Ponte and W.Bruce Croft: “Text Segmentation by Topic”, First European Conference on Research and Advanced Technology for Digital Libraries, 1997.

[11] 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 (2000-12).

[12] 篠崎, 斎藤, 堀, 古井 : “話し言葉音声の認識を目指して”, 信学技報, **SP2000**-96, pp.7-12 (2000-12).

[13] 西村雅史,伊東伸泰 : “講義コーパスを用いた自由発話の大語彙連続音声認識”, 信学論, D-II Vol. J83-D-II No.11 pp.2473-2480 (2000-11).

[14] 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 (2000-03).