音声文書検索の応用によるオンデマンド講演システム

藤井 敦[†,†††] 伊藤克亘[††,†††] 石川徹也[†]
[†] 図書館情報大学
[††] 産業技術総合研究所
[†††] 科学技術振興事業団 CREST
fujii@ulis.ac.jp

1 はじめに

近年，マルチメディア情報の普及や情報通信のブロードバンド化に伴って，多種多様な情報を誰もが容易にオンラインで受信や発信できるようになった．このような現状では，大量な情報の中から必要な情報だけをいつでもどこでも手軽に活用できる基盤技術が重要である．

現在普及しているマルチメディア情報として，テキスト，音声，画像がある．これらが有機的に混在し，しかも繰返し利用したくなるコンテンツとして「講演」が挙げられる．講演は，予稿やスライドなどの資料を併用しながら，対面で話すのが一般的である．また，出版された教科書に基づいた講義を放映するテレビ番組もある．

我々は，講演のビデオデータを対象にして，要求に応じた内容を視聴するオンデマンドシステムの研究開発を行っている [4, 8]．本システムを用いると，教科書や予稿などの講演資料テキストを閲覧しながら，関心があるビデオ内容（音声と画像）を選択的に視聴することが可能になる．

2 資料と講演の相違点

一つの内容を伝達するために，資料（書き言葉）と発話による講演（話し言葉）という2つの異なる手段が存在する．もしも，どちらか一つの手段で講演が十分に成立するならば，本研究で提案するシステムの意義は希薄である．しかし，資料と講演は，一方が存在すれば他方は必要がないというような排他的な関係ではない．そこで，両者の相違点について考察することは，本研究の意義を明確にするために有効である．

資料は，冊子体や電子版などの形態に依らず，章立てのような文書構造や文字種などの表層情報を手掛かりにして「斜めに読む」ことができる．また，何度も繰り返し読み返すことができる．そこで，全体の概要を把握したり，関心のある箇所を高速に探索することが容易である．すなわち，ランダムアクセスに適している．

それに対して，講演は逐次アクセスが原則であり，ランダムアクセスには適さない．実際の講演では，内容が資料のページやスライドの単位で構造化されていても，資料のように戻ったり飛ばしたりして聞くことはできない．録画された講演ビデオの場合は，意味内容に基づいて索引付けを行わない限り，構造を手掛かりに早送りや巻戻しによって必要な箇所を特定することは困難である．音声認識（ディクテーション）によって発話内容をテキストに変換しても，書き起こしされた話し言葉を読むことは，資料を読む場合に比べて負担が大きい．

他方において，情報の量は講演の方が相対的に多い．資料は講演内容に対してページ数が制限されることがある．これは主に印刷コスト等の理由によるものなので電子版には該当しない．しかし，現状では冊子体と電子版の資料が等価な形で存在することが多いので，必然的に電子版も字数の制限を受ける．それに比べると，講演は発表内容に適した長さであることが多い．

また，資料を読むよりも講演を聞く方が分かりやすいことがある．発話内容が適度に冗長であったり，会話的な表現も使用される．熟練した講演者ならば，聴衆の反応に応じて難易度や説明の仕方を動的に調整することもある．講演時には，資料を執筆した時よりも新しくかつ正確な情報が補足されることもある．

以上の考察から，講演ビデオデータを高度に“視聴”するためのある一つのモデルが成立する．まず，資料を読んで内容の概要を把握し，興味のある箇所を素早く特定する．そして，特定した箇所に関する内容だけを講演ビデオから選択的に視聴し，理解を深める．その結果，講演ビデオを全て見なくても，最小限のコストで必要な情報を取得することが可能になる．

3 システム構成

3.1 概要

本研究で提案するオンデマンド講演システム（Lecture-On-DEMand system: LODEM）の構成を図1に示す．システムで利用するコンテンツは，同一の講演に関する資料テキストとビデオデータである．ユーザが指定した章や節などに基づいて資料テキストから対象範囲を抽出し，それに関連するビデオトラックの頭出しを行い，再生する．また，ユーザが思いついたキーワード，フレーズ，文など資料テキストに書かれていない任意のテキスト情報も入力することができる．すなわち，システムの

オンライン処理は，資料テキストから抽出した一定の範囲やユーザが思いついたキーワードを検索質問（query）として用いて，講演ビデオデータから関連するビデオ内容を検索する処理に相当する．

現在は，検索処理だけをサーバで行い，クライアントPC上のウェブブラウザで資料の閲覧，検索質問の入力，ビデオの視聴を行うことができる．

そのためには，通常の情報検索と同様に，講演ビデオデータに対する索引付けが必要である．具体的には，講演ビデオから音声データを抽出し，発話単位に基づいて講演ビデオを意味のある一定のまとまりに分割する．さらに音声認識によって音声データの書き起こしを生成して講演データベースを作成する．ここでは，分割されたそれぞれの部分を「パッセージ」と呼ぶことにする．

当該データベースは，テキスト情報をキーにして，関連するビデオパッセージを検索することができるように編成されている．現在，索引付けは全て自動で行われる．

本システムの検索処理は，音声文書検索（Spoken Document Retrieval: SDR）[1]の一種である．英語を対象にしたSDRの研究では，音声認識の単語誤り率が30%程度あっても，テキスト検索と同等の検索精度を達成できることが知られている[2]．音声認識の精度は対象とする音声データの品質や講演内容によって変化するものの，現在の音声認識技術で本システムを実用化できる可能性は十分にある．

また，本研究は講演の特徴に着目し，通常の検索システムにはない機能を実現した．講演資料は簡潔に書かれるため，資料中では一つにまとまっている内容が，実際の講演では離散した箇所で説明されることがある．そこで，単一の検索質問に対して，関連する説明を尤度が高い順に複数出力する方が網羅性の点では好ましい．

しかし，同じような説明を何度も出力することは冗長であり，非効率的である．そこで，以前出力した内容との重複を避けながらビデオパッセージを検索する機能を実装した（3.4節参照）．この機能によって，最小限の情報でユーザの情報要求を満足することが可能になる．

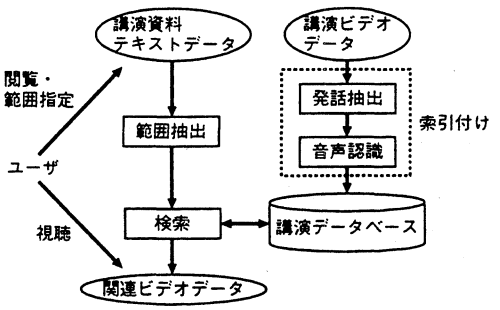

図 1: オンデマンド講演システム LODEM の構成

3.2 講演ビデオデータの索引付け

講演ビデオデータの索引付けは以下の手順からなる．バッチ処理で動作するツール群によって，全ての処理は自動化されている．

1. ビデオデータから音声データを抽出する．
2. 音声データをパッセージに分割する．
3. 音声認識によって各パッセージに対する書き起こしを生成する．
4. 書き起こしに対して，通常のテキスト検索と同様の索引付けを行う．

手順1は，対象とするビデオデータの規格（フォーマット）が分かっていれば技術的な問題はない．

手順2では，書き言葉の段落に相当するような意味的・論理的なまとまりに音声データを分割することが理想的である．しかし，話し言葉に対しては，文を認定することすら難しく，また人間が書き起こしを作成しても，意味のある単位に分割することは難しいのが現状である．

本システムでは音声認識システムを用いて自動的に書き起こしを作成しているため，話題の転換に用いられる接続表現が正しく認識されない可能性がある．また，音声認識用の統計的言語モデルを書き言葉の学習データから作成しているため，話し言葉特有の表現に対する認識率が低い．

以上の問題点を考慮して，現在はパワーなどの物理的な尺度で音声を分解し，無音区間を区切りとして発話を抽出する．さらに，複数の発話をまとめて一つのパッセージを構成する．

手順3では，連続音声認識コンソーシアムのディクテーションソフトウェア[7]で提供されている音声認識エンジン（デコーダ）と音響モデルを利用する．また，新聞記事や論文抄録などから学習した言語モデルを独自に作成して[5]，対象に応じて適宜使い分けている．

手順4では形態素解析システム「茶筌」[1]を用いて書き起こしを単語に分割し，品詞情報に基づいて名詞を索引語として抽出する．また，カタカナ語や新語は未知語と認定されることが多いので，未知語も索引語として抽出する．抽出した索引語を用いて転置ファイルを編成し，テキストによる検索質問を用いた検索を可能にする．

3.3 関連説明の検索

3.2節で説明した処理によって，講演ビデオデータ（書き起こし）を複数のパッセージに分割することができる．そこで，各パッセージを異なる文書と見なせば，情報検

[1] http://chasen.aist-nara.ac.jp/

索の分野で提案された各種の手法を用いて，検索質問に関連するパッセージを効率的に特定できる．

関連度の計算には確率型の手法 [3] を用いた．これは，近年の情報検索手法の中でも比較的高い検索精度を実現することで知られている．具体的には，文書 d の関連度スコアを式 (1) によって計算し，スコアが高い順番に文書を出力する．

$$\sum_t \left(\frac{TF_{t,d}}{\frac{DL_d}{avglen} + TF_{t,d}} \cdot \log \frac{N}{DF_t} \right) \quad (1)$$

ここで，$TF_{t,d}$ は索引語 t が文書 d に出現する頻度である．DF_t は t を含む文書数であり，N は総文書数である．DL_d は文書 d の長さ（バイト数）であり，$avglen$ は平均文書長である．

検索質問からは，書き起こしパッセージの場合と同じ方法（3.2 節の手順 4）で索引語を抽出する．

3.4 効用最大化に基づく再帰的検索

資料中では一つの章や節としてまとまっている内容が，実際の講演では複数の箇所で分散して説明されることがある．そこで，単一の説明だけを出力するのではなく，3.3 節の式 (1) で計算されるスコアが高い順に複数のパッセージを出力する必要がある．

しかし，ユーザが十分理解したにも拘わらず，同じようなビデオ内容を再生することは効果的ではない．一度検索（再生）した内容は次回以降なるべく検索しないようにすれば，少ない情報によって効用を最大化することが期待できる．これは，ユーザがある文書を閲覧したときに，ユーザが抱える情報要求のうち，まだ満足していない部分を特定し，その部分に対して効果的な別の文書を提示する問題と捉えることができる．

松村ら [6] は，ユーザが文書 d を閲覧しても依然として満たされない情報要求を，検索質問と文書 d との差分ベクトルとして表現し，次回以降の検索に利用する手法を提案した．ここで，検索質問と各文書があらかじめ索引語のベクトルとして表現されていることが前提である．

我々は，この手法を応用し，検索質問ベクトルと一度検索されたパッセージに対応するベクトルとの差分を取りながら再帰的に検索を繰り返すことで，冗長な内容の再生を回避する．また，検索を繰り返すたびに，検索質問ベクトルが縮退するので，式 (1) で計算されるスコアは次第に低下する．これはユーザの要求が次第に満たされていく過程をモデル化している．そこで，スコアに対する閾値を設定し，スコアが閾値を下回った時点でユーザの要求が十分満たされたと判断して処理を終了する．

現在は，スコアの閾値に対して決定的な値や一般的な範囲は分かっていない．テストデータを用いて実験を繰り返しながら，経験的に設定する必要がある．

4. 実行例と考察

3 章で説明したシステムを実装し，以下に示す複数のコンテンツを対象に試験運用を行った．

- 教科書が市販され，その内容に基づく講義が放映されているテレビ番組（45 分間）
- 学会でのチュートリアル講演（予稿あり）を再現したビデオ（30 分間）

前者では，CATV から受信したビデオデータを DV に録画し，後者では，DVCAM を用いて聴衆がいない状態でスタジオで撮影した．

構築したシステムは，ウェブブラウザで検索および視聴できるように実装した．検索処理はサーバで実行される．しかし，現状ではビデオデータのサイズが大きくサーバから短時間でダウンロードすることが困難である．そこで，ビデオデータはクライアントの PC 上に保存している．この点は今後改善する必要がある．

システムインタフェースの外観を図 2 に示す．この図では，学会チュートリアル講演を視聴している．

ユーザは画面左の資料を読みながら，画面下の入力ボックスに講演資料からコピーした内容や任意のキーワードを入力することができる．検索を実行すると，検索されたパッセージの書き起こしが複数提示され，ユーザが選択した内容が画面右に再生される．

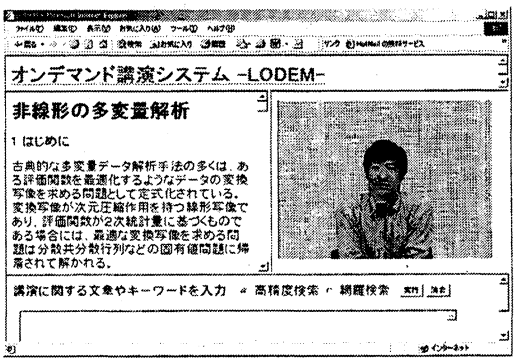

図 2: オンデマンド講演システムのインタフェース

以下，実行例に利用したコンテンツを対象にいくつかの考察を行った．

まず，話し言葉と書き言葉の違いについて調べるために，講演資料テキストと講演発話の文字数を比較した．ここで，発話の文字数は，実際には音声認識によって生成された書き起こしに含まれる文字数である．音声認識には認識誤りがあるものの，文字数に関しては正解とそれほどの差はない点に注意を要する．結果を以下に示す．

	講演資料	書き起こし
テレビ番組	3,489	11,038
学会講演	3,342	6,017

この結果から分かるように，話し言葉は，書き言葉に比べると 2〜3 倍程度に冗長であった．話し言葉と書き言葉の違いは，講演のスタイルなどに依存するため一般化することは困難であるものの，今回対象にしたコンテンツでは，書き言葉固有の表現として以下のようなものが顕著に現れた．

- 話の進行に必要なメタ表現（あらすじなど）
- 具体例（例えば，教科書では「公害事件」と表記されている箇所に対して「水俣病」の例が挙げられている）
- 聞き手に語りかけるような表現（「皆さんご存知のように」など）
- 丁寧な表現（「ですます調」など）

話し言葉と書き言葉の違いに関して，テレビ番組の例を示す．太字が話し言葉のみに現われた表現である．

> 刑法とは，**皆さんご存知のようにどのような行為が犯罪となり，その行為にどのような刑罰が科されるかを定めた法律のことを指します．六法を開くとどの六法にも刑法と名前のついた法律，すなわち刑法典がのっています．**

DV で録画されたデータの分割は，パワーに基づいて 400ms 程度の無音区間を検出したところで音声データを区切り，それを一つの発話とした．さらに，発話 3 つで一つのパッセージを構成した．その結果，45 分の講義が 34 のパッセージに分割された．ここで，パッセージあたり最長で 582 文字，最短で 96 文字，平均 312 文字が含まれていた．

さらに，テレビ番組の冒頭 10 分間に対して音声認識精度を評価した．論文抄録や新聞記事から作成した言語モデルを使い分けた結果，単語誤り率（Word Error Rate: WER）は 20〜30%という結果が得られ，音声文書検索には十分な認識精度であることが分かった．

今回対象としたテレビ番組は，法律や裁判に関する講義だった．教科書から「人の健康に係る公害犯罪の処罰に関する法律」という一部を抜き出して検索質問とした場合の検索結果（書き起こし）を以下示す．

> また，1960 年代には，熊本の水俣病事件を，はじめとする．公害事件が多発しました．今の安全率で得るのは，公害事件の，民事裁判に関する映像です．このような発光が事件の多発を契機として，1970 年には向上などから，ヒトの健康がする物質を鞏出して，ヒト腺眼，身体に実験を調査する行為を処罰する．人・高にかかる血行が**犯罪の処罰に関する法律**で稼いでされています．

当該検索結果において「人の健康に係る公害犯罪の処罰に関する法律」は最後の太字部分である．この部分は正しく音声認識されていないものの，周辺の文脈に現れる語によって正しく検索されている．このように，検索質問や検索対象パッセージをある程度の長さにすることで，音声認識誤りに対しても頑健な検索が可能となった．

5 おわりに

同一の内容に関するテキスト情報とビデオ（音声・画像）情報を併用し，テキスト入力によってビデオ内容を選択的に視聴するオンデマンドシステムを実現した．現在までにテレビの講義番組，学会チュートリアルなどを対象に実験を行った．その他，新聞とニュース番組，レシピのウェブページと料理番組のようなマルチメディアコンテンツの利用が考えられる．現状では画像はユーザが講演内容を理解する助けにはなるものの，処理の対象にはなっていない．今後は，画像解析を応用したビデオパッセージ分割などについて研究を行う予定である．また，評価方法などについても今後検討する必要がある．

参考文献

[1] John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text REtrieval Conference*, pp. 83–91, 1997.

[2] Pierre Jourlin, Sue E. Johnson, Karen Spärck Jones, and Philip C. Woodland. Spoken document representations for probabilistic retrieval. *Speech Communication*, Vol. 32, pp. 21–36, 2000.

[3] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[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. 情報処理学会研究報告 2001-SLP-39, pp. 165–170, 2001.

[5] 伊藤克亘, 秋葉友良, 藤井敦, 石川徹也. 音声入力型テキスト検索システムのための音声認識. 日本音響学会講演論文集, pp. 193–194, Oct. 2001.

[6] 松村真宏, 大澤幸生, 谷内田正彦. AAS: 文書の組み合わせによってユーザの興味を満足する検索システム. 人工知能学会誌, Vol. 14, No. 6, pp. 1177–1185, 1999.

[7] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄（編）. 音声認識システム. オーム社, 2001.

[8] 藤井敦, 伊藤克亘, 秋葉友良, 石川徹也. 音声言語データの構造化に基づく講演発表の自動要約. ワークショップ「話し言葉の科学と工学」, pp. 173–177, 2001.