音声ドキュメントのコンテンツ化

中川聖一　　情報工学系／センター長

1　はじめに

ビデオオンデマンドによる e-Learning の長所はいつでも好きな時に学習、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、音声自動認識・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。我々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、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し、被験者実験にて有効性を確認した。

2　講義音声認識

2.1　認識方法

表 1 に示すように 2 種類の音響モデルを用意した SPOJUS〔2〕を用いて講義音声の認識実験を行った。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大学院講義（「電子計算機応用特論Ⅰ、Ⅱ」、「音声情報処理工学特論」）の内容とドメインは近い。

SPOJUS は 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と bigram 言語モデルを用い、得られた N-best 候補をトライグラムでリスコアする〔2〕。実験では N=200 とした。

コンテキスト独立（Context Independent, CI）音節モデルは、直前の情報を考慮していないコンテキスト独立モデルである。しかし、直前にどのような発音を行ったかによって音は変わってくる。コンテキスト依存（Context Dependent, CD）音節モデルはそうした発声の変化を考慮してモデリングしたものである。今回使用した CD 音節モデルは、直前の単語の末尾の音節が、/a/、/i/、/u/、/e/、/o/、/N/（撥音）、/SIL/（無音）、/qs/（促音）の 8 種類のうちのどれだったかということを受けて、異なるモデルとなる。この 928 種類の CD 音節モデルを使用した認識システムを講義音声に対して適用し、CI 音節モデルとの比較を行った。

2.2　認識結果

音声認識の評価結果を表 2 に示す。講義によって差があるものの、CD 音節モデルを使用したシステムは、CI 音節モデルと比較して性能が向上することが分かった。1pass 目による結果同士を比較すると、CD 音節モデルは CI 音節モデルに対し、Accuracy の平均で 2.39%、Correct では平均で 4.90%、向上した。2pass 目の結果では、Accuracy 平均で 1.34%、Correct 平均で 5.44%の向上となっている。

表1　認識システム

表記	デコーダ	音響モデル	言語モデル
CI	SPOJUS	音節（133音節）*1)	CSJ *2)
CD	SPOJUS	コンテキスト依存音節（928音節 *3)）	CSJ *2)

*1 CSJ　最終版797講演（男性話者）より学習、コンテキスト独立特徴パラメータは25次元
*2 CSJ　最終版で学習、トライグラム言語モデル（17625語彙）
*3 CSJ　最終版814講演（男性話者）より学習、コンテキスト依存特徴パラメータは38次元

3　講義コンテンツの改良

3.1　機能

我々は以上に述べた音声自動認識を元にして、要約、インデキシングなどの自動処理を行い、講義教材をより高度に利用できるシステムを構築した。ベースとなるシステムは日立 EZ プレゼンテーターである[1]（以下, EZP）。これは、ビデオと音声が入力された状態でプレゼンテーションソフト[2] を起動すると、スライドの切替えタイミングを保存しておき、後にビデオと同時に自動的に切り替わるスライドをネットワーク配信できるシステムである。EZP に追加した主な機能は以下の通りである。

・再生速度変更機能
Windows Media Player の基本的な機能だが、ブラウザ上からでも細かく再生速度を変更できるようにセレクトボックスを追加した（図 1 の「要約率・話速設定」部分）。

・要約再生機能
図 1 の「要約率・話速設定」部分にあるセレクトボックスを操作することで、任意の要約率で再生が可能。

・スライド一覧から当該箇所へビデオをジャンプ
図 1 のスライド一覧部のリンクをクリックすることで、対応する箇所へビデオをジャンプさせる。

・キーワード一覧
図 1 の書き起こしからのキーワード一覧部にある単語をクリックすることで、その単語が発話されている文頭の箇所へビデオをジャンプさせる。

---

[1] http://www.hitachi-ad.co.jp/ezp/
[2] Microsoft®PowerPoint®のみに対応

・スライド画面
現在のビデオ位置で実際に表示されていたスライドを表示。
・スライド画面のキーワードリンク
スライド中に現れるキーワードをクリックすることで、その単語が発話されている文頭の箇所へビデオをジャンプさせる。
・要約文一覧
全重要文を一覧表示し、現在発話されている重要文を強調表示する。

3.2 キーワード抽出の改善

従来のスライドからのキーワード抽出では、ChaSen で形態素解析した結果、名詞と判定された形態素をキーワード候補として抽出していた。その結果、一般的な名詞もキーワードとして大量に抽出され、利用者がキーワードを選択しづらいコンテンツとなっていた。一般的に、講義の際に使用される専門用語は複合語である場合が多い。本研究では専門用語抽出用 Perl モジュールである TermExtract〔4〕を使用して複合語を含めたキーワード候補を抽出し、各キーワード候補について tf･idf スコアを求め、スライド中の平均 tf･idf スコアよりも高かった

表 2　認識評価結果（Correct[%]）

講義	CI_1pass	CD_1pass	CI_2pass	CD_2pass
1	46.23	51.50	46.26	51.75
2	47.85	55.24	47.92	56.34
3	89.03	63.24	59.86	64.51
4	54.71	57.86	55.01	59.36
5	66.19	70.68	66.10	70.41
平均	54.80	59.70	55.03	60.47

表 3　従来のキーワード単位との比較

文	従来	複合語処理
0001	大/学/項/計算/量/音/内容/音声/言語/人/理解/音声/言語/音/処理/母音	音声言語/音声言語
0002	内容/境界/応用/木/音声/基礎/音声/分析	音声分析
0003	内容	
0004	音声/認識	音声認識
0005	人/動的/時/列	
0006	音声/自然/言語/処理	言語処理
0007	隠れ/マルコフ/モデル/モデル/情報	隠れマルコフモデル
0008	自然/言語/処理/最大/化	言語処理
0009	法/語/言語/モデル	言語モデル
0010	日本語/室/認識/言語/情報/理解/モデル/化/認識/時/話/モデル	理解モデル

図 1　教材動作画面例

ものをキーワードとして抽出した。従来の抽出単位との比較を表 3 に示す（音声認識結果からの抽出）。複合語に整理されることでキーワード数が削減され、より利用しやすい単位になったと思われる。

4　むすび

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基盤となる音声認識について実験を行い，コンテキストを考慮したモデルを用いることで性能向上を得た。また、当研究室で開発している講義コンテンツについて、キーワード抽出手法の改善、キーワード認識法の改善に関する検討を行った。その結果、キーワードの出現確率を強調することによってキーワード認識率は改善することができた。今後の課題としては、より自然なキーワードモデルの構築や、サブワード単位でのキーワードスポッティングによる未知語への対応などを考えている。

参考文献

[1] 日本語講義音声コンテンツコーパス http://www.slp.ics.tut.ac.jp/CJLC/

[2] 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

[3] 富樫慎吾、藤井康寿、北岡教英、中川聖一、“講義音声ドキュメントのコンテンツ化と視聴システムの試作”、音声ドキュメント処理ワークショップ論文集, pp. 17-24、2007

[4] 専門用語自動抽出用perlモジュール "termextract" http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html/

[5] 富樫慎吾、北岡教英、中川聖一、“スライド情報を用いた言語モデル適応による講義音声認識”、日本音響学会、春季講演論文集、pp. 1-P-24、2006

発表文献

[1] 富樫慎吾、中川聖一、“講義収録映像の音声情報を利用した講義コンテンツの構築と評価”、日本音響学会、秋季講演論文集、pp.1-3-3、2007

[2] 中川聖一、富樫慎吾、山口優、藤井康寿、北岡教英、“音声ドキュメントのコンテンツ化と視聴システム”、電子情報通信学会論文誌、第 J91-DⅡ巻、pp.238-249、2008

[3] 富樫慎吾、中川聖一、“講義音声ドキュメントのコンテンツ化とブラウジングシステムの改良”、第 2 回音声ドキュメント処理ワークショップ講演論文集、pp.155-160、2008

[4] J. Wang、L. Wang、S. Nakagawa、“LVCSR based on context-dependent syllable acoustic models”、Asian Workshop on Speech Science and Technology、SP2007-200、pp.81-86、2008