音声ドキュメントのセマンティックアーカイブ化

中川聖一（第４工学系、センター長）　北岡教英（名古屋大学、客員助教授）　土屋雅稔(情報メディア基盤センター)

１　はじめに

ネットワークの速度・容量増加に伴い、インターネット上で利用できる音声情報を含むコンテンツ（映画、放送、ビデオ等）が増大の一途を辿っている。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、音声内容をより効率的に把握できるようにするため、音声自動認識や音声要約、自動インデキシング、トピックセグメンテーション等の技術によるメタデータ付与の研究が注目を集めている。本研究では、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、効率的な視聴を支援するシステムを構築することを目的としている。

２　音声ドキュメントの収録と分析

我々の大学で実施されている大学院向けの音声言語情報処理や情情報工学に関する6名の講義を収録している。講義は１回 150 分で、前後半 75 分ずつである。その際、録音機材が認識結果に与える影響を調査するため、そのうちの 4 講義について、ハンドマイク（ソース A）,有線のピンマイク（ソース B）,ワイアレスのピンマイク（ソース C)の３種類の装置を同時に使用して収録を行った。

講義音声は、予想に反して、講演音声よりも言い直し、倒置が少ないことがわかった。フィラーなどの話し言葉の諸現象を扱う言語モデルの検討も行っている。

３　連続音声の認識

この音声資料を、我々の研究室で開発し大語彙連続音声認識システム SPOJUS[１]を用いて音声認識実験を行なった。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、音声情報処理に関するトピックが主であるため、今回の音声認識対象である講義の内容とドメインは近く未知語率（音声認識語彙サイズ 17000 語）は比較的小さい。

マイクによる差という観点からソース A、B、C を比較すると、各講義において、A>B>C という順に認識性能が劣化した。実用的には、ピンマイクにより収録した音声を対象とする必要がある。現段階では、単語認識率は３０％～７０％程度である。音声ドキュメントの検索やビデオの頭出し、音声要約のためには、この程度の認識でも十分であるが、書き起こし結果をユーザに提示するには、７０－８０％程度の認識率が必要である。

高精度に音声認識を行うために、コンテキスト依存音節単位音響モデル、単語単位の音響モデル、音響伝達特性の正規化、音響モデルの話者適応化、言語モデルのトピック適応化、などの研究を行っている。

４　音声要約

韻律情報と表層的言語情報の両方を用いた音声要約の研究を行ってきた。またそれらを組み合わせた手法も行った。ここで用いた特徴量の説明を以下に記す[２]。

tf　各文中の名詞の tf を計算し、スコアの高い文から 25%を抽出する。

頻出単語　出現頻度の高い方から、その語を 2 つ以上含んでいる文の数が全体の 25%になるように語を選び、文を抽出する。

slide-title　スライドのタイトルに含まれる名詞が出現する文を重要文として抽出する。

slide-tf　スライド中に三回以上現れる名詞を頻出単語とし、その頻出単語が一回以上含まれる文を重要文として抽出する。

F０　基本周波数の高い文の順に 25%を抽出。(スライド情報不使用時、slide-tf の代替)

パワー　パワーの強い順に 25%を抽出。(スライド情報不使用時、slide-title の代替)

発話時間長　発話時間の長い文から 25%を抽出

重要表現　重要文に含まれる重要表現（形態素列および品詞列）を機械学習（CRF：条件付確率場）であらかじめ抽出し利用。

組合せ実験　上記の特徴量に加え、話速の遅い文、発話時間長の短い文といった棄却特徴（非重要文である特徴）も組み合わせる。

単独特徴量の中では頻出単語および発話時間長によるものが性能が良かった。発話時間の長い文が抽出されているので、時間的な要約率は５０％程度である。特徴量の組合せで、人間の要約結果と大差ない結果が得られた。

スライド情報を使用していない場合の要約実験では、若干精度が低下し、スライド情報が有効であることがわかった。

５　インデキシング

スライド中のキーワードと発話単語をマッチングし、キーワード単位での対応付けを行った。キーワードの抽出には tf・idf を用い、tf 値は各スライドから、idf 値は CSJ の 264 講演文の書き起こしから取得した。最終的に、

画面上のスライドにある対応付けられたキーワードをクリックすることで、講義ビデオをその単語が発話された位置へジャンプさせることができるシステムを構築した。図 1 にシステムのインターフェース画面を示す。画面右側のスライド画面中にある下線を引かれた単語は、インデキシングによって音声認識結果へと対応付けられており、クリックすると対応する位置へビデオがジャンプする。また画面右下のキーワード一覧には、音声認識結果に現れたキーワードを時系列順に並べ（五十音順も可能）、スライド中のリンクと同様に対応する位置へビデオをジャンプさせることができる。前者の方法は、認識誤りによりスライド中でキーワードとならない単語が存在する。一方、後者の方法は、音声認識の誤りに頑健であり、必ず頭出しできる利点がある。

6　話者認識

音声ドキュメントには、話者情報の付与が有用である。我々は、以前、音声を話者別に自動クラスタリングし話者情報を付与する技術を開発したが、最近は、話者認識率の向上に取り組んでいる。代表的な統計的話者識別法である HMM と GMM の組み合わせ、および従来利用されてこなかった位相情報の導入により、識別誤り率を半分以下に抑える技術を開発している。

7　むすび

本報告では、収録した講義ビデオ音声をコンテンツ化するための一連の処理－音声認識・要約・セグメンテーション・インデキシングの研究を紹介した。

大規模な音声ドキュメントを有効に利用するために、このほかの研究として、音声対話システムや音声ドキュメントの検索、質問応答システムの研究を行っている。また、応用研究として、ニュース映像の語学学習教材への利用、発音学習システムの研究などを行なっている。

参考文献

[1] 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-D Ⅱ No.3、pp.799-807、2004

[2] 小林聡、山口優、中川聖一、“表層的言語情報と韻律的情報を用いた講義音声の重要文抽出”、自然言語処理 Vol.12 No.5、pp.43-68、2005

発表文献

[1] 富樫慎吾、山口優、北岡教英、中川聖一、“講義音声の認識・要約・インデックス化の検討”、情処学研報 2006-SLP-62-11、2006.7

[2] 小暮悟、西崎博光、土屋雅稔、中川聖一、“講義コンテンツ収集・分析および講義音声における認識手法に関する検討、” 第 1 回音声ドキュメント処理ワークショップ論文集、pp.1-8、2007.2

[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”、第 1 回音声ドキュメント処理ワークショップ論文集、pp.17-24、2007.2

[4] 浅川康平、中川聖一、“MFCC と位相情報を用いた話者認識”、日本音響学会講演論文集、１－P―17、2007.3

[5] 太田圭、中川聖一、“日本人の英語文発声の発音評価法”、日本音響学会講演論文集、3－8－12、2007.3

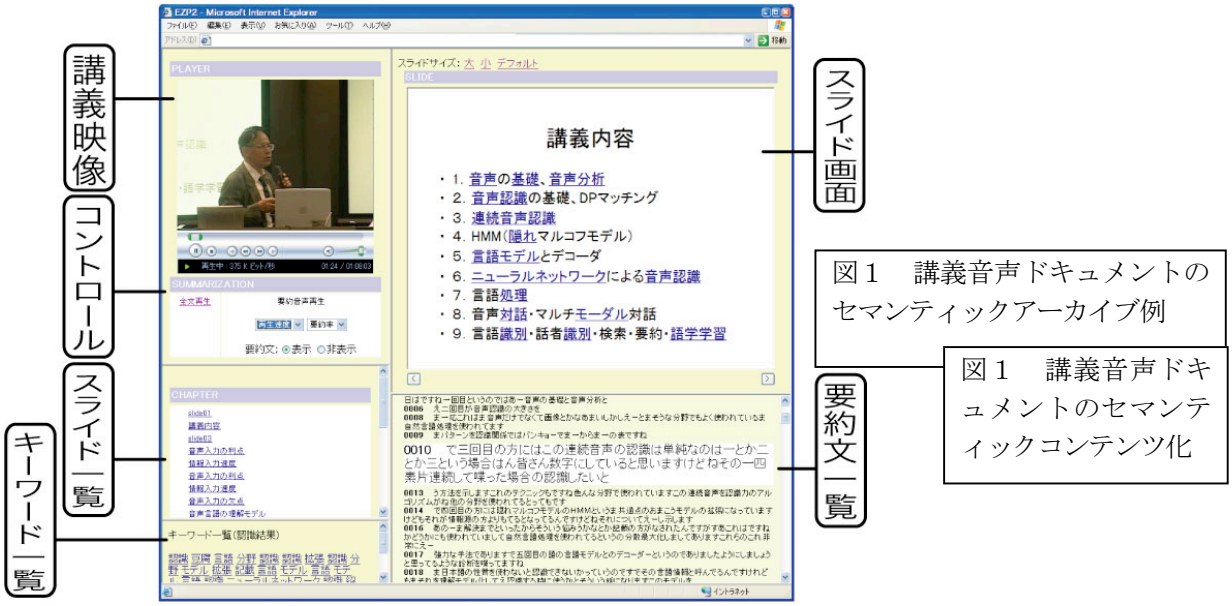

図１　講義音声ドキュメントのセマンティックアーカイブ例

図１　講義音声ドキュメントのセマンティックコンテンツ化