3D2-05

音声検索を用いたe-Learningシステムの評価

Evaluation of retrieval system of the on-line lecture using speech recognition

橋本 幸司[*1] 速水 悟[*2]
Koji Hashimoto Satoru Hayamizu

[*1]岐阜大学大学院工学研究科 Graduate School of Engineering, Gifu University
[*2]岐阜大学工学部 Faculty of Engineering, Gifu University

This paper describes retrieval system of the on-line lecture using speech recognition. Many lectures in universities are becoming available through Internet in the form of video and audio contents by streaming technology. When a user takes an on-line lecture, a user has to view and listen to unnecessary scenes. However, if a user can search a scene to see, a user does not need to look at unnecessary scenes. Our purpose is reducing the burden of users' search time using speech recognition. We verified the validity of the retrieval using speech recognition by comparing two systems, one system with speech recognition by continuous speech recognizer and the other system with speech retrieval by keyword spotting. We tested two systems by questionnaire to users.

1. はじめに

近年，我が国ではインターネットの普及，特にADSLなどのブロードバンドの急速な普及に伴う，インターネット接続環境の高速化や低価格化により，データ量が大きい音楽や画像等のマルチメディアの配信なども幅広く行われるようになってきている．また，画像などの配信の普及と，家庭やその他の施設でも，インターネットのネットワークにつながる環境があれば，どこからでも視聴できるという利点から，インターネットを利用した遠隔教育である，ネットラーニングも大学などの高等教育や教育産業などで普及しつつある．

本研究は，ネットラーニングの主な利点である，時間短縮という点に着目して，音声認識を利用したオンライン講義の検索を扱う．また，音声認識結果のテキストと音声の特徴量に対して検索を行うシステムの評価を行い，講義における音声検索の有効性を示す．

2. システムの構成

本研究で使用したデータは，実際の講義をカメラ，マイクで収録することで作成した．作成されたコンテンツは，ブラウザで視聴可能であり，カメラで撮影した動画，マイクで収録した音声，講義の際使用したスライドの画像，講義の章構成を含んでいる (Fig. 1).

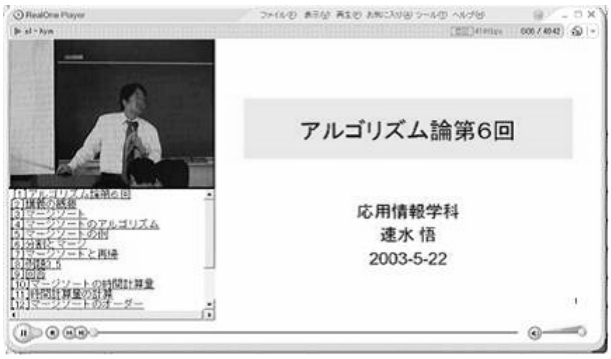

Fig. 1 実際のオンライン講義コンテンツ

連絡先: 橋本幸司，岐阜大学大学院工学研究科応用情報学専攻，岐阜市柳戸 1-1，koji@hym.info.gifu-u.ac.jp

この作成したコンテンツをWeb，ストリーミングサーバの二つのサーバを使用してネットワークを構成し，講義を配信している．

本研究では，２つの音声認識を利用した検索手法を用いたシステムを作成した．音声認識を利用した検索を行う上で，以下の共通の手順により検索対象となるインデックスデータを作成する．

1. 講義の撮影
2. 講義の動画データから音声を抜き出す
3. 抜き出した音声データに対して音声認識
4. 音声認識を行った結果をインデックスデータとして保存

作成した2つのシステムいずれも，上記の方法でインデックスデータを作成しているが，インデックスデータが音声認識した結果のテキストデータであるか，音声特徴量データであるかに違いがある．各システムはCGIにより，ユーザがブラウザからキーワードを打ち込んで検索を行える．

1つ目の検索システムは，講義を行っている講師の声を音声認識したものを仮名漢字混じりのテキストデータとして保存し，そのテキストデータをインデックスデータとして，キーワード検索を行うものである．この手法の検索対象は講義のスライドごとの章の検索が可能である．

2つ目はメディアドライブ社で開発された音声検索を利用する．これは，講義を行っている講師の声を，特徴量に変換したものをインデックスデータとして，入力されたキーワードの発話特徴量と比較をして検索を行う手法である．この手法は，講師がキーワードを発話した時刻を検索可能である．

2.1 音声認識結果のテキストに対してキーワード検索を利用したシステム

講義データはスライドごとの章に分けられているため，音声認識の結果とスライドのテキスト情報によってTF・IDF法を用いて情報検索を行い，キーワードがどの章に含まれているのかを割り出す．音声認識はJuliusを利用する [河原 2000]．言語モデル・認識用辞書をそれぞれ用意する必要がある．

このシステムでは章を検索するのでスライドのテキスト情報も利用して検索を行っている．スライドのテキスト情報を利用するために，あらかじめスライドのテキスト情報は抜き出しておいて，音声認識の結果と共にインデックスデータとして使用する．

インデックスデータは XML ファイルであり，講義の音声認識の結果テキスト，スライドのタイトル，スライドのテキストを保存している．

TF・IDF は，索引語頻度 (Term Frequency) と文章頻度の逆数 (Inverse Document Frequency) という 2 つの考えにより行われる [伊藤 2001, 藤井 2002].

2.2 音声の特徴量に対して検索を利用したシステム

このシステムは音声の特徴量を独自の形式で保存して，それに対して、ブラウザから入力されたキーワードの特徴量をもとに検索を行うものである．このシステムの検索結果は，キーワードが発話されたところからの頭出しを行う．また，テキストデータに比べてファイルサイズの大きい音声の特徴量に対して検索を行うので，単純にテキストに対して検索を行うキーワード検索よりも時間がかかる．

3. 実験・評価

実際に被験者を用意し，アンケートにより，2 つの手法を利用したシステムを評価した．構築した 2 つのシステムを比較するために,「どちらのシステムが検索結果に必要な情報を高い候補に含んでいるか？」という有効情報,「どちらのシステムが検索結果不要な情報を多く含んでいるか?」という不要情報，主に，この 2 点に付いてアンケートを行う．

アンケートの対象者は 12 人で，検索対象はアルゴリズム論の講義半年分とし，検索する総章数は 601 である．評価する内容は，音声認識結果のテキストデータとスライドのテキスト情報をインデックスデータとして検索を行うシステムをシステム A，音声認識結果を音声特徴量としてインデックスデータとして検索を行うしシステム B の 2 つを比べて有効情報，不要情報の評価をする．

アンケートの方法は、ユーザに 3 つの指定したキーワードと 7 つのキーワードを自由に選んでもらった，計 10 のキーワードで検索してもらい, 各キーワードについて上記の 2 つの観点から評価してもらう．評価基準は，+3(システム A に有効情報が多い/不要情報が多い)〜-3(システム B 有効情報が多い/不要情報が多い) の 7 段階で，10 のキーワードについて 12 人にアンケートを行い，120 件の評価データを得た.

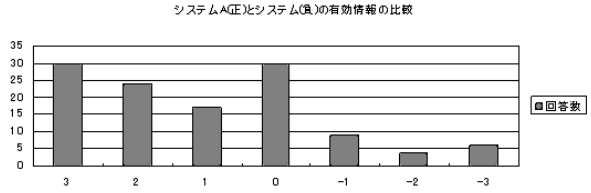

Fig. 2 有効情報の評価

有効情報は明らかに，音声認識結果のテキストをインデックスデータとするシステム A が多く含んでいる (Fig. 2)．そのようになった要因は，音声の特徴量をインデックスデータとするシステムは，不要情報が多いとともに，スライドのテキスト情報を使用を使用していないためであると思われる．一方，音声認識結果のテキストをインデックスデータとするシステムは語の並びの統計情報である言語モデルとスライドのテキスト情報を使用しているので，講義の内容にまとまりがあり，検索したときに上位に検索されると思われる．

不要情報は明らかに，音声の特徴量をインデックスデータとするシステム B の方が，システム A より含んでいる (Fig. 3)．そのようになった要因は，音声の特徴量をインデックスデータ

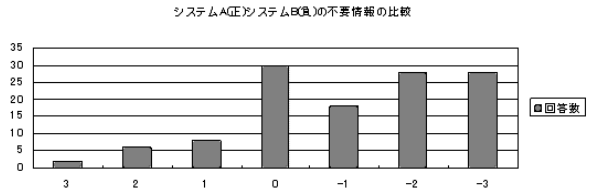

Fig. 3 不要情報の評価

とするシステムは，発音が似たような語を検索してしまう可能性が高いからであると思われる．

4. 結果・考察

以上の結果から，音声の特徴量に対して検索を行うシステムの方は，キーワードを発話した部分は検索はされるものの，語彙の統計情報である言語モデルや講義のまとまりを示すスライドのテキストを使用していないために，検索精度をあげることが難しいと思われる．テキストを検索するシステムは，語彙の並びの統計である言語モデルとスライドのテキスト情報を使用しているため，音声認識の結果に，講義内容のまとまりがあり，検索精度を上げていると思われる．また，音声認識の結果テキストに対して単純なキーワード検索のみでは，精度よく検索することができない．理由は，キーワードに関する重要な章の講義を行っていたとしても，そのキーワードを講師が発話するとは限らない．また，スライドのテキストのみで検索を行うと，図のみのスライドはたとえ重要であっても検索されない．

5. おわりに

本研究の一番の課題は音声認識精度の向上である．本研究で作成したシステムは，いずれも音声認識を使用しているため，キーワードによって音声認識の精度に大きな違いがある場合がある．その結果音声認識されやすいキーワードはよく検索されるが，音声認識されにくいキーワードは検索されないといったことが起こる．音声認識を行うシステムでは，検索を行うキーワードに強いキーワード，弱いキーワードがあり，認識精度の差をなくすことは重要である．

講義のような構造をもったものを音声認識の結果だけから検索するだけではなく，講義の構造を利用するなどの工夫が必要である．

謝辞

プログラムを提供していただいたメディアドライブの八村謙治氏，橋本恭貴氏に感謝いたします．また，実験に協力していただいた，上澤泰氏，日比野哲也氏，諸炯氏，アンケートに協力していただいた皆様に感謝いたします.

参考文献

[伊藤 2001] 伊藤，藤井，石川，“音声文章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-39，165-170，2001

[藤井 2002] 藤井　敦，伊藤克亘，石川徹也，“音声文書検索の応用によるオンデマンド講演システム,”言語処理学会第 8 回年次大会発表論文集，pp.192-195，Mar.2002

[河原 2000] 河原，李，小林，武田，峯松，伊藤，山本，宇津呂，鹿野，“日本ディクテーション基本ソフトウェア” 日本音響学会誌 56 巻 4 号，pp.255-259，2000