プレゼンテーション蓄積検索システムにおける講義・講演音声情報を利用した適合度の改善

岡本　拓明[†]　仲野　亘[†]　小林　隆志[††]　直井　聡[†††,††]　横田　治夫[††,†]
岩野　公司[†]　古井　貞熙[†]

† 東京工業大学 大学院 情報理工学研究科 計算工学専攻
†† 東京工業大学 学術国際情報センター
††† 株式会社 富士通研究所
E-mail: †{okamoto,wnakano}@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp, †††, ††naoi.satoshi@jp.fujitsu.com,
†{yokota,iwano,furui}@cs.titech.ac.jp

あらまし　我々は，講義・講演のビデオとその中で使われたスライドをメタデータにより統合コンテンツとして蓄積するとともに，その統合コンテンツの特性を考慮したシーン検索機能を有する UPRISE (Unified Presentation slide Retrieval by Impression Search Engine) を提案してきた．これまで UPRISE では，スライド構造やスライドの提示時間，前後のスライドのコンテクストなどをその検索機能に利用してきた．本稿では，シーン検索の適合率を向上することを目的に，講義・講演ビデオ中の音声情報を利用する手法について述べる．これまでの UPRISE の検索機能に，音声認識によって抽出した音声情報を適用する 4 種類の方法を提案し，実験によりそれらの効果を評価する．
キーワード　情報統合，情報検索，e-learning

Unified Presentation Contents Retrieval using Voice Information

Hiroaki OKAMOTO[†], Wataru NAKANO[†], Takashi KOBAYASHI[††], Satoshi NAOI[†††,††],
Haruo YOKOTA[††,†], Koji IWANO[†], and Sadaoki FURUI[†]

† Department of Computer Science, Graduate School of Information Science and Engineering,
Tokyo Institute of Technology
†† Global Scientific Information & Computing Center, Tokyo Institute of Technology
††† FUJITSU LABORATORIES LTD.
E-mail: †{okamoto,wnakano}@de.cs.titech.ac.jp, ††tkobaya@gsic.titech.ac.jp, †††, ††naoi.satoshi@jp.fujitsu.com,
†{yokota,iwano,furui}@cs.titech.ac.jp

Abstract　We have proposed the UPRISE (Unified Presentation slide Retrieval by Impression Search Engine) to store lecture videos and presentation slides used in the lectures. It provides dedicated scene search functions for the metadata-based unified presentation contents, using the slide structure, scene duration, and context information. In this paper, we use the voice information in the lecture videos to improve the precision of scene retrieval using the speech recognition technique. We propose four ways to add the influence of voice to the scene search functions of the UPRISE, and evaluate the effect of them. The experimental results indicate that the voice information is effective to improve the precision.
Key words　information integration, information retrieval, e-learning

1. はじめに

近年，動画や文書，音声ストリームなどの複数のメディアコンテンツを統合し，それらを蓄積，検索するシステムが数多く研究，および提案されており [1]〜[3]，e-Learning を始めとする様々な用途に用いられている．特に e-Learning 用のコンテンツに対しては，利用者が希望するコンテンツを検索できるだけでなく，どのコンテンツのどの箇所から視聴するべきかを効果的

に発見することが重要である.

そのような検索を実現するために，我々は教育コンテンツの統合機構，および統合コンテンツに対する高度な検索機能を実現するシステムである UPRISE(Unified Presentation slide Retrieval by Impression Search Engine) を提案してきた [4]〜[11].

UPRISE では，メタデータによるコンテンツの統合のために動画ストリームをシーンの連続であると抽象化し，各シーンとそこで使用された資料とを対応付けることでそれらを統合する. また，各シーンに対して，対応する資料の文字/構造情報，シーンの長さ情報，レーザーポインタなどのポインティング情報から検索用インデクスを作成し，高度な検索を可能としている. スライドの切り替えタイミングによってシーン分割を行うため，単なるスライド検索とは異なり，同じスライドを用いていてもバックトラックを起こしたり巻き戻りがあった場合でも，違うシーンとして区別することができるといった利点がある.

しかしながら，従来の UPRISE では，そのようなシーンの差別化を行うために，シーンの継続時間情報や，前後にどのようなシーンが出現しているかといった情報を用いており，これらの情報だけではシーンで説明されている内容を考慮できないため，シーンの差別化が十分でないという問題があった.

そこで本研究では，講演者がそのシーン中に発言した音声情報に着目し，音声情報をシーンの差別化のために利用することを考える．自動音声認識分野の研究においては，話し言葉研究用のデータベース (日本語話し言葉コーパス：CSJ) [12], [13] の整備により，従来では認識率の低かった話し言葉主体の講義講演の動画に対して認識精度が向上してきた．そこで，我々はこれまでに自動音声認識エンジンによって抽出した音声情報をシーン毎に比較し，音声情報が UPRISE でのシーンの差別化に利用できることを示した [11].

音声認識の結果を利用した検索を行う既存研究に，ニュースなどの放送音声文書に対して汎用音声認識を行ない，検索を行うシステム [14] があるが，ニュースを選択的に視聴するための検索であり，詳細なシーンの検索には適していない．また，音声文書に対して発話のまとまり単位で検索を行うシステム [15] があるが，発話された内容に依存し，詳細なシーンの検索には適していない．さらにこれらは，音声文書のみを検索対象とし資料を検索の対象として用いていないため，音声認識の精度に依存してしまうといった問題がある.

本研究の対象である統合コンテンツに対する音声認識技術を利用する研究としては，音声認識による音声データを検索に利用する試み [16], [17] があるが，音声データによるシーン分割にとどまっており，その情報を用いてシーンの検索精度を向上させるには至っていない.

本稿では，UPRISE のシーンの検索精度を向上させるために，音声情報を統合するための格納方法と，それらの情報を用いた新しい検索用適合度計算手法を提案する．提案手法では音声認識により抽出した音声情報を UPRISE のデータベースに登録し，その情報と従来の適合度とを統合する.

以下では，まず 2. 節で，UPRISE の概要を示す．UPRISE において，従来用いている適合度について 2.2 節で簡単に述べる.

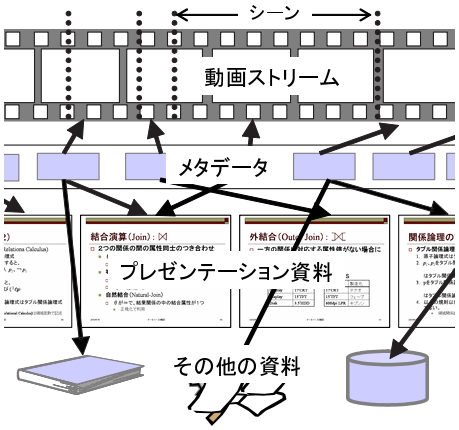

図 1　プレゼンテーション資料と動画の統合

3. 節では，本研究に用いる音声認識の概要と，音声情報のデータベースへの格納，従来の適合度との統合方法を説明し，最後に適合度の提案を行う．4. 節では，実際の講義をデータベースに格納し，検索実験を行う．最後に 5. 節において，まとめと今後の課題を述べる.

2. UPRISE の概要

以下では，UPRISE の概要について示す.

2.1 UPRISE のシステム

UPRISE のコンテンツ統合の概念図を図 1 に示す．メタデータには，動画のどの時刻にスライドの切り替えが起こったかというシーン情報と，その際にどのスライドを用いていたかという同期情報，スライドに含まれる文字列情報に対するインデクスを含める．これらの情報を保持するメタデータによってコンテンツを緩く結合することにより，個々のコンテンツが持つ情報に修正を加えることなくコンテンツの同期表示を実現し，柔軟な統合を可能にしている．また，このメタデータから得られるスライドの使用順序やスライドごとの説明に要した時間という情報を用いることによって各シーンの特性が具体化され，シーンの特性に基づいた検索が可能になる．UPRISE のシステムの詳細についてはこれまでの報告 [7] を参照されたい.

UPRISE では，動画中に同じスライドが複数回出現する場合にそれらを異なるシーンとして区別し，個別に適合度を算出する．これにより，それぞれのプレゼンテーションは対応する動画のシーンの集合として抽象化され，プレゼンテーション中の任意のシーンが検索可能になる.

以下では，UPRISE の検索において用いる，従来の適合度算出手法について簡単に述べる．詳細については, [6] を参照されたい.

2.2 従来の適合度算出方法

2.2.1 スライドの文書構造を考慮した適合度 I_p

適合度 I_p はスライドの文書構造を考慮した適合度であり，以下の式によって定義される.

$$I_p(s,k) = \sum_{l=1}^{L} P(s,l) \cdot C(s,k,l)$$

ここで，s はシーン，k はキーワード，l は行数であり，$P(s,l)$

はシーン s で用いられたスライドの行 l に与えられるポイント，$C(s,k,l)$ はシーン s で用いられたスライドの行 l にキーワード k が含まれる個数を表している．さらに $P(s,l)$ において行のインデントや文字の大きさに応じて重み付けをし，キーワードの出現回数だけではなく出現位置も考慮することができる．

2.2.2 シーンの時間情報を考慮した適合度 I_d

適合度 I_d は I_p にシーンの時間情報を付加した適合度であり，以下の式によって定義される．

$$I_d(s,k,\theta,u) = \left(\frac{T(s)}{u}\right)^{\theta} \cdot I_p(s,k)$$

ここで，$T(s)$ はシーン s の時間であり，θ は時間の影響の強弱を定めるパラメタ，u は単位時間を定めるパラメタである．これによって，長い説明を行っているシーンを重要視することができる．

2.2.3 シーンの前後関係を考慮した適合度 I_c

適合度 I_c は I_d にシーンの前後関係を付加した適合度であり，以下の式によって定義される．

$$I_c(s,k,\theta,u,\delta,\varepsilon_1,\varepsilon_2) = \sum_{\gamma=s-\delta}^{s+\delta} E(\gamma-s,\varepsilon_1,\varepsilon_2)\cdot I_d(s,k,\theta,u)$$

ここで，δ は考慮する前後シーンの範囲を定めるパラメタであり，$E(\gamma-s,\varepsilon_1,\varepsilon_2)$ は前後関係の強弱を定める関数である．$E(\gamma-s,\varepsilon_1,\varepsilon_2)$ は以下のように定義される．

$$E(x,\varepsilon_1,\varepsilon_2) = \begin{cases} \exp(\varepsilon_1 x) & (x<0) \\ \exp(-\varepsilon_2 x) & (x \geqq 0) \end{cases}$$

この適合度によって，適合度はそのシーンの前後 δ だけの範囲の影響を受け，ε が小さいほど影響を受けやすくなる．例えば $\delta=4, \varepsilon_1=5.0, \varepsilon_2=0.5$ の時，そのシーンの適合度は前後 4 シーンの適合度に影響を受け，後に続くシーンのほうにより強い影響を受ける．

2.2.4 その他の適合度

UPRISE では前述した I_p，I_d，I_c の他に，そのキーワードのどれだけシーンを特定できるかという性質 (特定性) を考慮した適合度 [9] や，レーザーポインタの情報を考慮した適合度 [10] を用いている．詳細については, [9], [10] を参照されたい．

3. 音声情報の利用手法

2. 節では UPRISE と従来の適合度について説明を行った．従来の適合度においてはシーンの差別化のために時間情報，前後関係を主に使用しており，そのシーンでキーワードが実際に説明されているかどうかは考慮していなかった．したがって，キーワードの説明がされていなくても，時間が長いシーンであれば順位が上位に表示されていた．

そこで，実際に説明されている内容を評価するための情報として音声情報に着目する．音声情報は実際に発話されている内容を示しているため，たとえ同じスライドを用いたシーンであっても発話されている内容により差別化を行うことができると考える．本研究はそのようなシーンに対して差別化を行うことにより，UPIRISE のシーン検索精度を向上させることを目的とする．

その目的を達成するために [11] ではシーン毎に音声認識による音声情報へのキーワード出現数と，実際に講義を聞き取ったキーワード出現数の比較を行い，音声認識による音声情報であっても，シーンの差別化に利用するには十分な精度であることを示した．一方，音声情報のみで検索を行う場合，誤認識の影響を受けてしまうため，十分な検索精度が実現できないと考える．そこで本論文では，音声情報を従来の適合度に統合した手法を提案することで，UPRISE のシーン検索精度向上を目指す．

以下では，講義・講演の音声情報を UPRISE の検索に利用する手法について提案する．まず，動画から自動音声認識によって音声情報を抽出する手法について説明し，抽出した音声情報をどのようなメタデータとしてデータベースに格納するかを示す．そして音声情報の統合方法について考察し，最後に従来の適合度に対して音声情報を統合した新しい適合度を提案する．

3.1 音声認識の概要

音声認識には連続音声認識ソフトウェア Julius (注1) [18] を用いる. Julius では，認識に用いる単語辞書の他に，音素ごとの音響特徴量をモデル化した音響モデルと，テキストコーパスから学習した言語モデルを用いて大語彙の汎用音声認識（トランスクリプション）を行うことができる．

講義や講演は自発性を持つ話し言葉であり，新聞などの文章から作成された言語モデルやその読み上げを用いて作成した音響モデルでは，精度の高い音声認識は難しい [19], [20]. そこで，本研究では山崎が [21] で作成した言語モデルと音響モデルを用いる. [21] では，日本語話し言葉コーパス (CSJ) [12], [13] の，学会講演 953 講演と模擬講演 1543 講演を学習データとして音響モデルを作成し，CSJ の学会講演 967 講演 (約 300 万単語) の書き起こしを学習データとして，バイグラムおよび逆向きトライグラムを言語モデルとして作成している. 言語モデル作成時の形態素解析には，茶筌(注2)，形態素解析用の辞書として，ipadic (注3) を用いている．また，認識用の辞書として，学習データ (約 300 万語) での出現頻度が高い順から選んだ 22,860 語を登録している.

本研究では，大学内の講義を撮影した動画ファイル (約 90 分) から，既存のエンコーダソフトウェアを用いて作成した，音声ファイルに対して音声認識を行う．単語辞書に登録されていない用語は音声認識に出現しないため, 山崎が [21] において作成した辞書に，各講義の資料より抽出した単語のうち辞書に含まれていない名詞を追加登録したものを単語辞書として使用する．

3.2 音声情報のデータベースへの格納

Julius を用いて音声認識を行うと，認識文章の候補 (単語区切り) と，その単語の発話に要した時間を含んだログファイルが得られる．そのログファイルから，単語と単語の発話された時刻 (秒単位) を計算し，XML ファイルを生成する．図 2 はその

(注1)：http://julius.source forge.jp/
(注2)：http://chasen.naist.jp/
(注3)：http://chasen.naist.jp/

XML の一部である．音声情報の XML は fragment タグの集合からなり，fragment タグはその時刻 (秒単位) に発話された単語を結合した文字列である string 属性を持ち，複数の voice タグを含んでいる．voice タグ 1 つは 1 つの音声情報を表している．voice タグの string 属性は音声情報の文字列であり，in 属性は発話された時刻 (秒単位)，consTimeMilli 属性は発話に要した時間 (ミリ秒単位)，consTimeSec 属性は発話に要した時間 (秒単位)，cmscore 属性は認識時のその単語の信頼度である．このファイルを用いて，単語と単語の発話された時刻を基に，単語の出現したシーンを計算し，検索テーブルに登録する．これにより音声データを考慮した，検索を行うことができる．また cmscore 属性は今回は使用しなかったが，この情報をデータベースに格納することによって，音声情報の信頼度を考慮した検索を行うことができると考える．これについては今後の課題とする．

```
- <fragment sec="4521" string="揺えることハザード">
    <voice in="4521" consTimeMilli="650" consTimeSec="0" string="揺える" cmscore="0.067" />
    <voice in="4521" consTimeMilli="270" consTimeSec="0" string="こと" cmscore="0.041" />
    <voice in="4521" consTimeMilli="600" consTimeSec="0" string="ハザード" cmscore="0.357" />
  </fragment>
- <fragment sec="4522" string="をえ～抑える">
    <voice in="4522" consTimeMilli="120" consTimeSec="0" string="を" cmscore="0.042" />
    <voice in="4522" consTimeMilli="100" consTimeSec="0" string="えー" cmscore="0.042" />
    <voice in="4522" consTimeMilli="330" consTimeSec="0" string="抑える" cmscore="0.085" />
  </fragment>
- <fragment sec="4523" string="方法他の">
    <voice in="4523" consTimeMilli="310" consTimeSec="0" string="方法" cmscore="0.346" />
    <voice in="4523" consTimeMilli="180" consTimeSec="0" string="他" cmscore="0.042" />
    <voice in="4523" consTimeMilli="310" consTimeSec="0" string="の" cmscore="0.108" />
  </fragment>
```

図 2　認識結果のログファイルから生成した XML 例

3.3　音声情報の統合方法

音声情報のみで検索を行うと，誤認識の影響を受けてしまうため十分な検索精度が得られないと考える．そこで資料のインデント情報やシーンの時間情報を考慮した，従来の適合度に統合することで精度を補完し，誤認識の影響を減らしながらシーンの差別化を行うことができると考える．以下では，3.2 節において登録した，キーワードとそのキーワードの発話されたシーンを基に，適合度として従来の適合度に統合する方法を示す．

まず，あるシーン s にキーワード k が発話された回数を $VA(s,k)$(Voice Appearance) とおく．この VA にシーン時間 $\left(\frac{T(s)}{u}\right)^{\theta}$ を積算することにより，従来の適合度に統合可能となる．次に，I_c のように音声情報に対しても音声の前後関係を考えることができる．ここで音声情報の前後関係を考える理由は，講義においてはキーワードは正解のシーンの前後においても，発話される確率が高くなると考えるからである．この関数を VC(Voice and Context) として以下のように定義する．

$$VC(s,k,\theta,u,\delta,\varepsilon_1,\varepsilon_2) = \sum_{\gamma=s-\delta}^{s+\delta} E(\gamma-s,\varepsilon_1,\varepsilon_2)\cdot\left(\frac{T(s)}{u}\right)^{\theta}\cdot VA(s,k)$$

これにより，あるシーンにおいてキーワードが発話された場合に，前後のシーンにもそのキーワードのポイントの一部が与えられる．

これらの適合度を考慮する際に音声のポイントを単に加算すると，スライド中にキーワードが存在しなくても検索対象となる．また湧き出し誤り (実際は発話されていないが，発話されていると認識される誤り) が起こると，そのシーンも検索の対象に含まれてしまうといった問題が起こると考える．

そこで，そのシーンで発話され，かつシーンで用いられているスライド中に出現する場合，つまり I_p のポイントが 0 でない場合のみ音声のポイントを与えるという方法を考える．また，発話されたシーンで用いられているスライドに存在していなくても，近傍のシーンのスライドに出現している，つまり I_c のポイントが 0 でない場合のみ音声のポイントを与えるという方法も考えられる．この組み合わせにより 3.4 節では 4 種類の適合度を提案する．

また，音声についても従来のテキスト検索技術における IDF [22] のような特定性を考慮する必要があると考える．これはキーワードに複数の単語を含む検索の時に，音声情報もテキスト中のキーワードと同様に，多くのシーンで発話されるようなキーワードはシーンを特定する能力が小さいからである．音声情報の特定性を考慮した関数 IVF(Inverse Voice Frequency) を以下のように定義する．

$$IVF(k) = \log\frac{\text{その講義に含まれるシーン数}}{\text{キーワード } k \text{ が発話されたシーン数}}$$

ここで，全検索範囲ではなくその講義に含まれるシーン数とした理由は，全検索範囲とすると対象の講義数が増えた場合に，分子にあたる総シーン数が大きくなるために，キーワード毎の差異が小さくなってしまい利点が失われてしまう，と考えるからである．この IVF を音声加算部分に積算することによって，音声の特定性を考慮することができる．

3.4　適合度の提案

3.3 節で示した統合方法の組み合わせにより，以下の 4 種類の適合度を提案する．

まず，該当シーンのスライドと音声の双方にキーワードが存在する場合にのみ音声の影響を考慮する適合度 I_{vas}(VA if keyword is in the Scene) として以下のように提案する．

$$I_{vas}(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2,\alpha) = \begin{cases} I_c(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2) + \alpha\left(\frac{T(s)}{u}\right)^{\theta}\cdot VA(s,k) & (I_p \neq 0) \\ I_c(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2) & (I_p = 0) \end{cases}$$

ここで，α は音声情報の影響度を決めるパラメタである．$\alpha = 6$ の時，1 回発話された重みがスライドのタイトルに 1 回出現した時の重みに相当する．$\alpha = 0$ の時，従来の適合度 I_c と一致する．

次に，I_{vas} の条件をすこし緩め，前後関係を考慮したスライドに関連するポイント I_c が 0 でなければ，音声の影響を考慮する適合度として，I_{vac}(VA if keyword is in Contexts) を以下のように定義する．

$$I_{vac}(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2,\alpha) = \begin{cases} I_c(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2) + \alpha\left(\frac{T(s)}{u}\right)^{\theta}\cdot VA(s,k) & (I_c \neq 0) \\ 0 & (I_c = 0) \end{cases}$$

次に，I_{vac} とは逆に，該当シーンのスライドにキーワードが登場し，音声の前後関係を考慮した関数 (VC) を利用する適合度として，I_{vcs} を以下のように定義する．

$$I_{vcs}(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2,\alpha)$$
$$= \begin{cases} I_c(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2) + \alpha VC(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2) & (I_p \neq 0) \\ I_c(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2) & (I_p = 0) \end{cases}$$

最後に最も緩い条件として，前後関係を考慮したスライドに関連するポイント (I_c) が 0 でなければ，音声の前後関係を考慮した関数 (VC) を利用する適合度として，I_{vcc} を以下のように定義する．

$$I_{vcc}(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2,\alpha)$$
$$= \begin{cases} I_c(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2) + \alpha VC(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2) & (I_c \neq 0) \\ 0 & (I_c = 0) \end{cases}$$

また 3.3 節で述べたように，音声加算部分に IVF を積算することによって特定性を考慮することができる．ここで，特定性を考慮した適合度を提案する際に，スライド内テキストの特定性を考慮した関数を導入する．ここで用いた特定性は [9] で提案した $ISFP$ ではなく，IVF に合わせて講義毎の頻度を用いたものを使用する．これを $ISFC$ とし以下に定義する．

$$ISFC(k) = \log \frac{\text{その講義に含まれるシーン数}}{\text{キーワード } k \text{ がスライドに出現するシーン数}}$$

これを上記で定義した 4 種類の適合度の I_c 部分に積算することにより，スライド内テキストの特定性を考慮することができる．これを $I_{x \cdot ISFC \cdot IVF}$ とする．ただし，x は vas，vac，vcs，vcc のいずれかとする．判定条件に I_p を用いるか I_c を用いるかは，適合度によって異なる. 以下では簡単化のため，I_c のパラメタは省略する．

$$I_{x \cdot ISFC \cdot IVF}(s,k,\delta,\theta,u,\varepsilon_1,\varepsilon_2,\alpha)$$
$$= \begin{cases} I_c \cdot ISFC(k) + \alpha I_x \cdot IVF(k) & (I_p \neq 0 \text{ もしくは } I_c \neq 0) \\ I_c \cdot ISFC(k) & (I_p = 0 \text{ もしくは } I_c = 0) \end{cases}$$

4. 実　　験

実際の講義のコンテンツを提案手法によって UPRISE に登録し，登録したコンテンツに対して各適合度ごとの検索実験を行う．その後，実験結果に対して考察を行う．

4.1 実験に用いたデータ

実際の講義コンテンツに対して音声認識を行い，音声情報を UPRISE のデータベースに格納した．今回はデータベースの講義 (全 11 回) と，計算機アーキテクチャの講義 (全 12 回) を用いた．音声認識に用いる辞書については，データベースの講義は山崎が [21] において作成した辞書に，講義から抽出した 1099 語のうち辞書に含まれていない名詞 134 語を追加した，22,994 語を登録した辞書を用いた．また計算機アーキテクチャの講義は同じく山崎が作成した辞書に，講義から抽出した 1109 語のうち辞書に含まれていない名詞 176 語を追加した，23,036 語を登録した辞書を用いた．認識の辞書を講義毎に変更した理由は，すべての講義に共通の辞書を作成するよりも，認識の精度が若干上がると考えたからである．なお，認識用の辞書に追加するためには，単語に読み付与 (どのような発音で読むか) を行なう必要があるが，今回の実験では読み付与が困難な英単語，記号等は追加しなかった．音声認識を行った結果，データベースの講義については延べ 85435 単語，計算機アーキテクチャの講義については延べ 89363 単語の音声情報が得られ，データベースに登録した．また今回の音声認識における音声認識率 (単語正解精度) は全講義平均で 25.4%であった．また検索に用いた各単語に対して, 各シーン毎の音声認識によって認識された回数と実際に発話された回数を比較することで, 適合度に影響のある湧き出し誤りと正しく認識された割合を計測したところ, 対象単語に対する音声認識結果の平均 46.7%が湧き出し誤りであり，正しく認識された割合は平均で 26.7%であった．

4.2 実　　験

提案手法の評価を行うため，4.1 節で登録したコンテンツに対し，キーワードについて説明しているシーンを実際に検索する実験を行った．

まず，音声情報と従来の適合度の統合方法の有効性を確認するために I_{vas}，I_{vac}，I_{vcs}，I_{vcc} のみで評価実験を行う．さらに，音声の特定性の影響を議論するために，前述の実験結果において結果が良かったものに対して，$ISFC$ と IVF を考慮した適合度を用意し比較を行う．各実験は以下の条件の下で行った．

- パラメタは $\theta = 0.5$，$u = 60$，$\delta = 4$，$\varepsilon_1 = 5.0$，$\varepsilon_2 = 0.5$ に固定した．
- 各適合度ごとに 124 種類のキーワードを検索した．
- 今回の検索範囲はキーワードの正解シーンの含まれる講義毎とした．
- キーワードに対し，最もよく解説していると判断したシーンをそのキーワードの正解シーンとした．
- 適合度の種類ごとに，正解シーンが何番目に順序付けされたかを記録した．
- I_{vas}，I_{vac}，I_{vcs}，I_{vcc} のパラメタ α の値を 0 から 10 まで 1 刻みで変化させた時と，0 から 100 まで 5 刻みで変化させた時の 2 種類の実験を行った．
- 上の実験で結果が良かったものに対して，$ISFC$ と IVF を考慮した適合度と，$ISFC$ のみを考慮した適合度を用意し，α の値を 0 から 20 まで 5 刻みで変化させて調査を行った．

パラメタを上記のように設定した理由は，前回までの実験による経験則である．今回はパラメタを変更しなかったが，パラメタの設定により影響が異なることは考えられ，これについては今後の課題とする．

検索範囲について各講義毎とした理由は，今回は音声認識を各講義毎に行なったからであるが，実験に用いたデータベースと計算機アーキテクチャに共通の用語は少なく，実験にはほとんど影響を与えないと考える．

なお，評価に際して指標となる再現率 (recall) と適合率 (precision) [23] について述べる．再現率は検索結果に含まれていた正解が全正解の中で占める割合，適合率は検索結果に対して正解が占める割合であり，今回の実験では，正解シーンを各キーワードに対して 1 つとしており，今回は全ての試行において検索結果には正解シーンが含まれているため，以下の結果における再現率は常に 1 である．

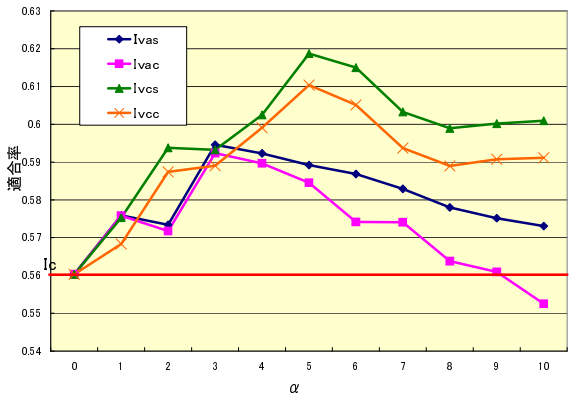

図 3　α を変化させた時の各適合度による適合率の推移 (1 刻み)

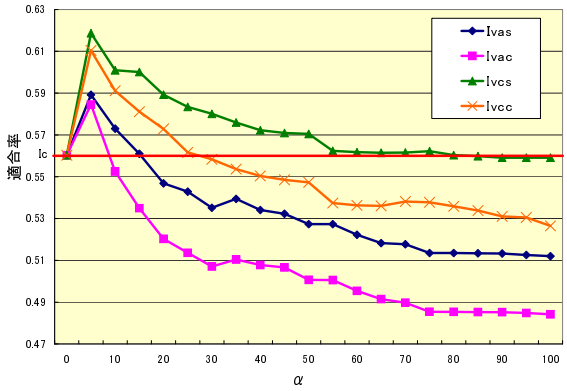

図 4　α を変化させた時の各適合度による適合率の推移 (5 刻み)

また，検索結果の範囲は正解シーンの順位までとする．これは，UPRISE の検索インターフェースでは適合度が 0 でないシーンが全て表示される仕様となっていること，および検索を行うユーザは検索結果の上位からシーンを見ると考えることから，検索精度の評価において正解シーン以下の順位にあるシーン群は無視できると考えるためである．

以上の前提のもとでは，適合率が以下の式で求まる．ただし，N は検索回数である．

$$\text{適合率} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{i\text{番目の検索での表示順位}}$$

4.3　実験結果と考察

4.3.1　統合方法による適合度の比較実験

表 1 は α を変化させた時の各適合度による適合率である．また，図 3 は表 1 をグラフ化したものである．

表 1　α を変化させたときの各適合度による適合率

α	I_{vas}	I_{vac}	I_{vcs}	I_{vcc}
0=I_c	0.560	0.560	0.560	0.560
1	0.576	0.576	0.575	0.568
2	0.573	0.572	0.594	0.587
3	0.595	0.592	0.593	0.589
4	0.592	0.585	0.602	0.599
5	0.589	0.574	0.619	0.610
6	0.587	0.574	0.615	0.605
7	0.583	0.563	0.603	0.594
8	0.578	0.561	0.599	0.589
9	0.575	0.553	0.600	0.591
10	0.573	0.520	0.601	0.591

従来の適合度 I_c による適合率は 0.560(α = 0 の時) であるため，各適合度による適合率は I_c より最高で I_{vas} が 0.035，I_{vac} が 0.032，I_{vcs} が 0.059，I_{vcc} が 0.05 上回った．これにより，音声情報を考慮する事によって検索の精度が上昇していることがわかる．その一方で，I_{vac} では α = 9, 10 において従来の適合度 I_c による適合率を下回った．これは音声情報を過剰に考慮する事によって逆に精度が落ちていることを示している．

このことは α を 0 から 5 刻みで 100 まで変化させた図 4 からも明らかである．最も精度の上がっていた I_{vcs} であっても α が 80 より大きくなると，従来の適合度 I_c による適合率を下回った．これは音声情報の影響度を上げることによって，正解シーン以外のキーワードがよく発話されているシーンが上位に来てしまうためと考える．表 1 より，音声情報の比率は α = 5〜6 くらいが最適だと考える．この時，発話 1 回が I_p におけるタイトル 1 回の出現と同等になる．

また，音声の前後関係を考慮したものと，そうでないものを比較すると，音声の前後関係を考慮したほうが適合率で上回っていることがわかる．これは，講義においては正解シーンの前後のシーンにおいてもキーワードが発話される確率が高まるためと考える．スライドに共通して出現しないとポイントにしない場合と，スライドに出現していなくても前後関係を考慮したポイントがある場合は音声のポイントを加える場合の比較では，前者のほうが適合率で上回っている．後者を用いるとスライドにキーワードを含んでいなくても音声情報の影響を受ける可能性があるため，スライドにキーワードを含まずに，たまたまキーワードの発話が正解シーンより多かったシーンが上位にくる可能性があるためと考える．また，複数の単語を含むキーワードの検索においては，片方の単語しか含まれていないスライドを用いているシーンは検索対象にはならないが，後者を用いることにより検索対象となる可能性がある．検索対象を広げることは再現率を広げることにあたり，今回のような正解をただ 1 つとする評価方法においては適合率が下がってしまう．今回用いた以外の評価方法については検討の必要性があると考える．

表 2　キーワードの順位の例 (I_{vcs})

キーワード	α = 0	α = 10	α = 100	α = 10000
スキーマ	14	2	2	2
ハザード	3	10	17	17
衝突，ベクトル	1	3	4	4

次に，最も精度の良かった適合度 I_{vcs} において，α の値を変化させたときの検索への影響をいくつかのキーワードについて考察する．表 2 は，’スキーマ’，’ハザード’，’衝突，ベクトル’ という 3 つのキーワードについて α の値を 0，10，100，10000 と変化させたときの検索順位を示したものである．

’スキーマ’ というキーワードでは I_c による順位が 14 位であったにもかかわらず，α が 10 以上になれば順位が 2 位に上昇している．これは，図 5 のようにスキーマがスライド上で重

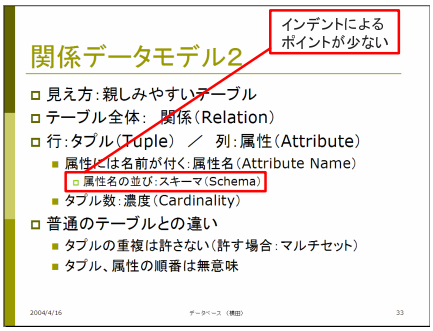

図 5　検索語：’ スキーマ’ の正解シーン

要な位置に書かれていないため，正解シーンにおいて I_p 等のテキスト情報だけではあまり重要視されなかったと考える．一方，このシーンではスキーマの説明がされているため，スキーマという単語は頻繁に発話され，音声情報の影響度を上げることによって正解シーンの順位が上昇したと考える．

一方，’ ハザード’ というキーワードでは音声情報の比率を上げることによって順位が下降した．これはハザードを用いた複合語 (例えば制御ハザード，データハザードなど) が多いため，ハザードは多数のシーンで発話されている．そのため正解シーンの順位が下がってしまったと考える．このように音声を考えることにより，複合語の多いキーワードでは，順位が下がる例も考えられる．そこでパラメタ α の値をキーワードの種類や特性によって変えることにより，適合度を上昇させることができると考える．

また，’ 衝突，ベクトル’ というキーワードでも，音声情報の比率を上げることによって順位が下降してしまった．この原因として出現頻度の違いがある．衝突は全シーン中 7 回発話されているのに対して，ベクトルは 70 回も発話されている．このことから音声情報の比率を上げることによって，発話されている回数の多い’ ベクトル’ の影響が大きくなってしまったと考える．この影響を減らすためには音声にも特定性を考慮する必要があると考える．

4.3.2　特定性を考慮した適合度の比較実験

I_{vcs} に $IS\,FC$ および IVF を考慮した適合度 $I_{vcs\cdot IS\,FC\cdot IVF}$ と $IS\,FC$ のみを考慮した適合度 $I_{vcs\cdot IS\,FC}$ との適合率の比較結果を表 3 に示す．まず $\alpha = 0$(音声を考慮しない場合) の値より，スライド中のキーワードの特定性 $IS\,FC$ を考慮することによって適合率が上昇していることがわかる．$IS\,FC$ のみを考慮した適合度 $I_{vcs\cdot IS\,FC}$ は，$\alpha = 5$ の時は適合度 I_{vcs} を下回っているが，α の値が大きくなると適合度 I_{vcs} を上回り，$\alpha = 20$ の時には適合率 0.623 を記録した．このことから，音声を考慮した適合度においても，スライド中のキーワードの特定性を考慮することは有用であるということがわかる．

一方，音声の特定性を考慮した適合度 $I_{vcs\cdot IS\,FC\cdot IVF}$ では，全体的に適合度 I_{vcs} を上回り，$\alpha = 15$ の時には従来の適合度 I_c とくらべて 0.063(6.3%) 高い，適合率 0.623 を記録した．このことからスライド中のキーワードの特定性に加えて，音声情報の特定性を考慮することは有用であると考える．

表 3　音声の特定性を考慮した適合度による適合率

パラメタ	I_{vcs}	$I_{vcs\cdot IS\,FC}$	$I_{vcs\cdot IS\,FC\cdot IVF}$
$\alpha = 0$	0.560	0.580	0.580
$\alpha = 5$	0.619	0.587	0.619
$\alpha = 10$	0.601	0.606	0.621
$\alpha = 15$	0.600	0.611	0.623
$\alpha = 20$	0.589	0.623	0.613

特定性を考慮しない適合度では $\alpha = 5$ 付近で最大値をとっていたのに対して，特定性を考慮した適合度では $\alpha = 10$ 以上で最大値をとる．これは特定性の関数は 0 から 1 の値をとるため，全体を除算することになり，適合率のピークが遅くなるためと考える．音声の特定性を考慮しない適合度 $I_{vcs\cdot IS\,FC}$ では，さらに適合率のピークが遅くなると考える．これはスライド情報の部分の影響が大きくなることによって，相対的に音声の影響が減るためと考える．

表 4　特定性を考慮した適合度によるキーワードの順位の例 ($\alpha = 5$)

キーワード	I_{vcs}	$I_{vcs\cdot IS\,FC\cdot IVF}$	$I_{vcs\cdot IS\,FC}$
要求，デイジーチェーン	4	3	4
関係，データモデル	3	2	3
衝突，ベクトル	2	2	1

表 5　キーワードの音声情報への出現数の例 (各講義毎)

キーワード 1	キーワード 2	キーワード 1 の出現数	キーワード 2 の出現数
要求	デイジーチェーン	50	0
関係	データモデル	161	37
衝突	ベクトル	7	70

次に，適合度 I_{vcs} において，特定性を考慮した時の検索への影響をいくつかのキーワードについて考察する．表 4 は，’ 要求，デイジーチェーン’，’ 関係，データモデル’，’ 衝突，ベクトル’ という 3 つのキーワードについて $\alpha = 5$ の値の通常の適合度，スライドと音声の両方の特定性を考慮した適合度，スライドの特定性のみを考慮した適合度における検索順位を示したものである．

’ 要求，デイジーチェーン’ というキーワードの検索においては，音声の特定性を考慮した $I_{vcs\cdot IS\,FC\cdot IVF}$ のほうが，順位が上昇した．これは表 5 の出現数からもわかるように，特定性を考慮することにより，2 語の間の出現格差を是正することができたと考える．その一方で，デイジーチェーンの音声情報への出現数が 0 である理由として，読み方や発音の違い (ディジーやチェイン) があげられる．このようなキーワードに対する読み付与・認識は難しく，今後の課題である．

’ 関係，データモデル’ というキーワードの検索においても，音声の特定性を考慮した適合度 $I_{vcs\cdot IS\,FC\cdot IVF}$ のほうが，順位が上昇した．これは’ 関係’ という言葉は出現数が多いため特定性が低く，特定性を考慮することにより改善したと考える．

’ 衝突，ベクトル’ というキーワードの検索においては，特定性のみを考慮した適合度 $I_{vcs\cdot IS\,FC}$ では順位が上昇したが，適合度 $I_{vcs\cdot IS\,FC\cdot IVF}$ では変化しなかった．これは改善幅が小さいために，特定性のみを考慮したもののみ順位が上昇したと考える．改善幅が小さい理由として，現在は音声認識を行い抽出した音声情報を用いていることがある．音声認識はシーンの差別化を行うためには十分な精度であると考えるが，特定性を考慮する

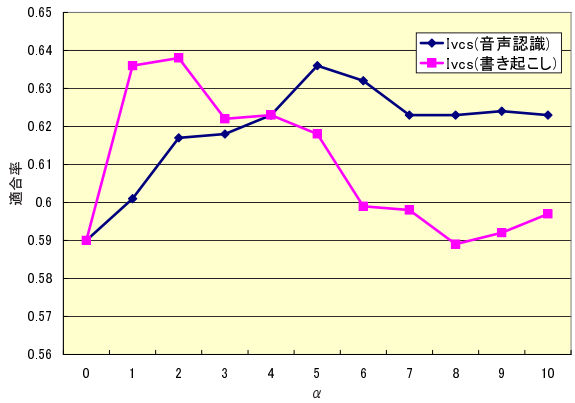

図 6　音声認識と書き起こし情報の比較

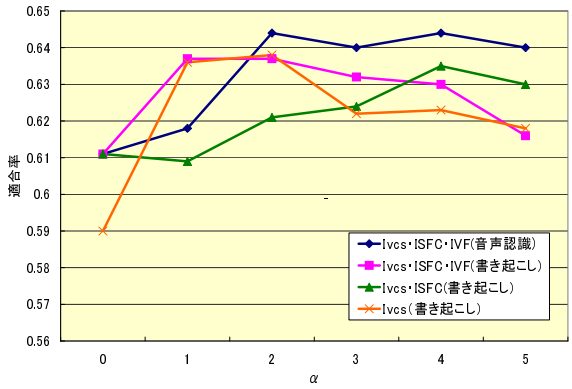

図 7　音声認識と書き起こし情報の比較 (音声の特定性)

場合にはキーワードの再現率 (認識率) が影響してしまうと考える．キーワードの再現率は [11] で延べたように，言語モデルに存在するかどうかや，単語の認識し易さに大きく影響する．そこで，特定性を考慮することによる有効性の詳細の評価のために，実際に講義音声を書き取り，そのデータから作成した音声情報を用いた適合度との比較が必要であると考える．

また IDF [22] などの特定性の関数は，キーワードに単語 1 語のみ含まれる検索の場合は影響しない．ところが，*IVF* の場合音声部分のみに特定性の関数を乗算するために，キーワードに単語 1 語のみ含まれる検索においても影響してしまう．この問題を解決するためには，キーワードに複数の単語を含む検索の場合のみ特定性を考慮するか，提案した式を改良する必要があると考える．これについても今後の課題とする．

4.3.3　書き起こしから抽出した音声情報を用いた場合との比較実験

これまでの実験においては，音声認識処理によって抽出を行なった音声情報を用いていた．しかしながら，音声認識処理の精度が向上した場合にも同様の効果が得られるかを確認する必要がある．そこで，講義を実際に聞き取ったテキストから抽出したデータを用いた場合との比較実験を行う．

実験のパラメタや評価方法については 4.2 と同様とした．ただし，書き起こしを作成できなかった 4 回分の検索範囲を除き，その範囲に正解シーンを含むキーワードを除外したため 104 種類のキーワードで検索を行った．ここで，書き起こしが作成できなかった理由は，録音が途中で途切れるなどしたためである．録音が途切れている区間以外は検索に有効であると判断したため，4.2 節の実験においては使用したが，書き起こし作成の対象からは除いた．前の実験で最も精度の高かった I_{vcs} を用いて実験を行い，音声の影響度を決めるパラメタ α は 1 から 10 まで 1 刻みで変化させた．また，テキスト情報の特定性のみを考慮した $I_{vcs\cdot ISFC}$ と音声の特定性を考慮した $I_{vcs\cdot ISFC\cdot IVF}$ についても書き起こし情報を用いた場合の実験を行い，パラメタ α は 1 から 5 まで 1 刻みで変化させた．

I_{vcs} を用いた比較実験の結果のグラフを図 6 に示す．グラフからわかるように，音声認識処理の結果を用いた場合は，音声影響度パラメタ $\alpha = 5$ のときに最大値 0.636 を取っているのに対し，書き起こし情報を用いた場合は $\alpha = 3$ のときに最大値 0.638 を取っている．最大値を取るパラメタの値が小さくなったのは，音声認識処理に対し書き起こしを用いた場合は，認識できていなかった部分が認識されたことでキーワード数が増加したためと考える．また最大値がそれほど変化しない理由として，3.3 節で説明した統合方法は脱落誤りを考慮したものであり，書き起こしを使用した利点が少なくなったためと考える．今回の統合方法を用いれば音声認識処理の結果を用いた場合でも，書き起こしと同程度の精度が得られることが確認された．

音声の特定性を用いた書き起こしの比較実験のグラフを図 7 に示す．グラフからわかるように，書き起こしを用いた場合でも 4.3.2 節の結果と同様の傾向が得られている．また音声の特定性においては，音声認識処理の結果を用いた場合のほうが，若干精度が高いという結果が得られた．これは今回の統合方法が脱落誤りを考慮した統合方法であるためと考える．そのため音声の精度が向上した場合の統合方法を検討する必要があると考える．

5.　おわりに

5.1　まとめ

本稿では UPRISE の検索精度を向上させるために，自動音声認識によって抽出した講義・講演の音声情報をシーンの差別化に利用する手法を提案した．まず，自動音声認識によって講義・講演の音声情報を抽出し，その音声情報の格納方法について述べた．さらに格納した音声情報を従来の適合度に統合した新しい適合度計算手法を提案した．また，実際の講義を UPRISE に登録し，評価実験を用いて提案手法の有効性を示した．

評価実験の結果，全ての適合度で従来の適合度 I_c(適合率 0.560) よりも適合率が上昇し，適合度 I_{vcs} においては最大 0.059(5.9%) 改善し，適合率 0.619 を記録した．またいくつかのキーワードについて詳細に考察することにより，音声情報による正解シーンの順位の変化を考察した．また音声における特定性 (キーワードがどれだけ文書を特定できるかという性質) を考慮した適合度を提案し，従来の適合度 I_c(適合率 0.560) にくらべて最大 0.063(6.3%) 改善し，適合率 0.623 を記録した．

さらに講義を実際に聞き取りデータ化した (書き起こし) テキストの結果と比較し，現在の統合方法においては，音声認識処理結果を用いた場合でも，書き起こしとあまり差のない精度が

得られることを示した．また，音声の特定性についても書き起こしを用いた比較実験を行い，同様の傾向が得られた．

5.2 今後の課題

本研究の今後の課題を以下に述べる．まず，本論文ではシーンの差別化を行うためには十分な精度であることを示したが，音声文書検索に対して誤認識の影響を減らすための手法 [24] や，音素の情報を利用して未知語 (辞書中に存在しない単語) に対応する手法 [25] を取り入れることで，音声認識時の精度低下が検索に与える影響を減らす必要があると考える．さらに，音声認識処理の精度が向上した場合の統合方法や利用方法についても考える必要がある．また 4.3 節で述べたように，読み付与の難しいキーワードへの対策や，キーワード単語 1 語のみを含む場合の検索時に，影響を与えないような特定性の式の改善も今後の課題である．今回の実験ではパラメタを経験則に基づいて決定したが，キーワードの種類を分類することによって，通常の適合度のパラメタや音声情報の比率パラメタを変化させることができれば，精度がさらに上昇すると考える．さらに，今回実験における評価方法では正解シーンを 1 つとしていた．しかし，実際の検索においては正解シーンが 1 つとは限らない．そこで，正解シーンを正解の度合い別に複数設定し評価する必要があると考える．

謝　　辞

本研究で用いた Julius と音響/言語モデルの使用にあたりご協力頂いた，東京工業大学大学院情報理工学研究科計算工学専攻の山崎裕紀氏に感謝致します．なお，本研究の一部は，文部科学省科学研究費補助金特定領域研究 (15017233,16016232)，独立行政法人科学技術振興機構 CREST，および東京工業大学 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の助成により行われた．

文　　献

[1] R. Müller and T. Ottmann. The ”Authoring on the Fly” system for automated recording and replay of (tele)presentations. *Multimedia Syst.*, Vol. 8, No. 3, pp. 158–176, 2000.

[2] Carnegie Mellon University The Informedia Project. Informedia ii digital video library. `http://www.informedia.cs.cmu.edu/`.

[3] G. D. Abowd. Classroom 2000: an experiment with the instrumentation of a living edu cational environment. *IBM Syst. J.*, Vol. 38, No. 4, pp. 508–530, 1999.

[4] 横田治夫. 東工大学術国際センターの情報蓄積・活用—教育コンテンツの統合とその手法—. 情処研報 DBS-125-58, 情報処理学会, July 2001.

[5] 村木太一, 吉田誠, 小林隆志, 直井聡, 横田治夫. メタデータによる講演資料と動画の統合と検索. In *Proc. of DBWeb2002*, pp. 97–104. 情報処理学会, 2002.

[6] Haruo Yokota, Takashi Kobayashi, Taichi Muraki, and Satoshi Naoi. UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine. *IEICE Transactions on Information and Systems*, Vol. E87-D, No. 2, pp. 397–406, February 2004.

[7] 小林隆志, 村木太一, 直井聡, 横田治夫. 統合プレゼンテーションコンテンツ蓄積検索システムの試作. 電子情報通信学会論文誌, Vol. J88-D-I, No. 3, pp. 715–726, March 2005.

[8] 岡本拓明, 小林隆志, 横田治夫. プレゼンテーション蓄積検索システムにおける適合度計算の改善. データ工学ワークショップ論文集, pp. DEWS2004–1–B–3. 電子情報通信学会 DE 研, March 2004.

[9] Hiroaki Okamoto, Takashi Kobayashi, and Haruo Yokota. Presentation Retrieval Method Considering the Scope of Targets and Outputs. In *Proc. of WIRI2005*, pp. 47–52, April 2005.

[10] Wataru Nakano, Yuta Ochi, Takashi Kobayashi, Yutaka Katsuyama, Satoshi Naoi, and Haruo Yokota. Unified Presentation Contents Retrieval Using Laser Pointer Information. In *Proc. of SWOD*, pp. 170–173, April 2005.

[11] 岡本拓明, 小林隆志, 直井聡, 横田治夫, 古井貞熙. 講義講演シーン検索における音声データの利用. 研究報告 2005-dbs-137 (ii) (78), pp.585-591, 電子情報通信学会, July 2005.

[12] 国立国語研究所. 日本語話し言葉コーパス. `http://www2.kokken.go.jp/~csj/public/`.

[13] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. Spontaneous speech corpus of japanese. In *Proc. LREC2000*, Vol. 2, pp. 947–952, Athens, Greece, May 2000.

[14] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P.Woodland. The Cambridge University spoken document retrieval system. In *Proceedings of ICASSP'99*, pp. 49–52, 1999.

[15] 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, March 2002.

[16] 中澤聡, 佐藤研治, 奥村明俊. 講演音声とプレゼンテーション資料の対応付けによる講演検索. Technical Report 情処研報 2005-SLP-55-12, 情報処理学会, February 2005.

[17] 中澤聡, 佐藤研治, 奥村明俊. 講演音声-プレゼンテーション資料アライメントによる講演検索. 言語処理学会第 10 回年次大会ワークショップ「e-Learning における自然言語処理」, March 2004.

[18] 河原達也, 李晃伸. 連続音声認識ソフトウエア Julius. 人工知能学会誌, Vol. 20, No. 1, pp. 41–49, 2005.

[19] 篠崎隆宏, 斎藤洋平, 堀智織, 古井貞熙. 話し言葉音声の認識を目指して. 信学技報 SP2000-96, 電子情報通信学会, Dec 2000.

[20] Takahiro Shinozaki, Chiori Hori, and Sadaoki Furui. Towards automatic transcription of spontaneous presentations. In *Proc. Eurospeech2001*, Vol. 1, pp. 491–494, Aalborg, Denmark, Sep 2001.

[21] 山崎裕紀. 講義音声認識の高精度化に関する研究. 東京工業大学工学部 卒業論文, February 2005.

[22] G. Salton. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley, 1988.

[23] D. A. Grossman and O. Frieder. *Information Retrieval Algorithm and Heuristics*. Kluwer, 1998.

[24] A. Singhal and F. Pereria. Document expansion for speech retrieval. In *Proc. of ACM SIGIR 99*, pp. 34–41, 1999.

[25] S. Srinivasan and D. Petkovic. Phonetic confusion matrix based spoken document retrieval. In *Proc. of ACM SIGIR2000*, pp. 81–87, 2000.