論文 / 著書情報
Article / Book Information

論題(和文)	音声情報を統合したプレゼンテーションコンテンツ検索
Title(English)	Presentation-Content Retrieval Integrated with the Speech Information
著者(和文)	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
Authors(English)	Hiroaki OKAMOTO, Wataru NAKANO, Takashi KOBAYASHI, Satoshi NAOI, Haruo YOKOTA, Koji IWANO, Sadaoki Furui
出典(和文)	電子情報通信学会和文論文誌, Vol. J90-D, No. 2, pp. 209-222
Citation(English)	IEICE Journal, Vol. J90-D, No. 2, pp. 209-222
発行日 / Pub. date	2007, 2
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2007 Institute of Electronics, Information and Communication Engineers.

音声情報を統合したプレゼンテーションコンテンツ検索

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[†]

Presentation-Content Retrieval Integrated with the Speech Information

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

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プレゼンテーションコンテンツとして蓄積するとともに，そのプレゼンテーション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，前後のシーンのコンテクストなどをその検索機能に利用してきた．本論文では，シーン検索の精度を向上させることを目的に，講義・講演ビデオ中の音声情報を，これまでのUPRISEの検索機能に統合する．講義ビデオから音声認識によって音声情報を抽出し，その音声情報のコンテクストへの影響を考慮した4種類の統合手法と，音声情報の特定性を考慮した統合方法を提案する．更に，実際の講義コンテンツを用いた実験によりそれらの効果を評価する．

キーワード　e-learning，情報検索，情報統合，音声情報，音声認識

1. まえがき

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，及び提案されており[1]～[5]，e-Learningをはじ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，我々は講義や講演などのプレゼンテーションコンテンツの統合機構，及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[6]～[14]．

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づ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，レーザポインタなどのポインティング情報から検索用インデックスを作成し，高度な検索を可能としている．スライドの切換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，バックトラックしたり，スライドを再利用したりした場合であっても，同じスライドを用いる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．

従来のUPRISEでは，そのようなシーンを適切に順位付けするために，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いた．しかしながら，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，これらのシーンの順位付けが十分できないという問題があった．

そこで本研究では，シーンの適切な順位付けのために講演者がそのシーン中に発言した音声情報に着目し，

[†] 東京工業大学大学院情報理工学研究科計算工学専攻，東京都
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152–8552 Japan
[††] 東京工業大学学術国際情報センター，東京都
Global Scientific Information and Computing Center, Tokyo Institute of Technology, Tokyo, 152–8550 Japan
[†††] (株) 富士通研究所，川崎市
Fujitsu Laboratories LTD., Kawasaki-shi, 211–8588 Japan

音声情報を統合した検索を考える．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）[15], [16] 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認識精度が向上してきた．そこで，本研究では，音声認識エンジンによって抽出した音声情報をシーンの適切な順位付けのために利用することを考える．

本論文では，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，その情報と音声情報を考慮していないこれまでの適合度とを統合する．

以下では，まず **2.** で，UPRISE の概要を示し，UPRISE においてそのシーンが与えられたキーワードに対しどの程度適合しているかを表す指標である，適合度の従来の計算方法について **2.2** で簡単に述べる．**3.** では，本研究に用いる音声認識の概要と，音声認識による音声情報の有用性を評価する．次に，**4.** において，音声情報のデータベースへの格納方法と音声情報を考慮していないこれまでの適合度との統合方法を説明し，音声情報を統合した新しい適合度の提案を行う．**5.** では，提案手法の有効性を確認するために行った，実際の大学での講義をコンテンツ化したプレゼンテーションコンテンツを対象とした検索実験に関して説明する．最後に **6.** において関連研究に関して述べた後，**7.** において，まとめと今後の課題を述べる．

2. UPRISE の概要

2.1 UPRISE のシステム

UPRISE では，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を検索対象としている．

UPRISE における，メタデータを用いた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換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とその構造，更にはレーザポインタの照射位置・時間などの情報を含める．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も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 [8] を参照されたい．

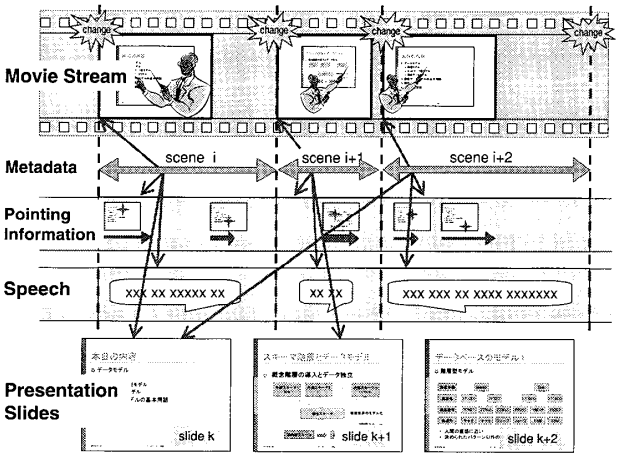


図 1　プレゼンテーション資料と動画のメタデータによる統合
Fig. 1　Unifying a presentation video and slides.

UPRISE では，動画中に同じスライドが複数回出現する場合であってもそれらを異なるシーンとして区別し，それぞれ個別に，与えられたキーワードに対しどの程度適合しているかを表す指標である適合度を算出する．

2.2 音声情報を統合しない場合の適合度算出方法

以下では，UPRISE の検索において用いる，音声情報を考慮していないこれまで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，[7] を参照されたい．

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．

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

ここで，s はシーン，k はキーワード，l はスライド中の行位置であり，$L(s)$ はシーン s で用いられたスライドの行数，$P(s,l)$ はシーン s で用いられたスライドの l 行目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 l 行目にキーワード k が含まれる個数を表している．更に $P(s,l)$ 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．

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

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ータである．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．

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

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

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

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 $\delta = 4, \varepsilon_1 = 5.0, \varepsilon_2 = 0.5$ のとき，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，後に続くシーンの方により強い影響を受ける．なお，この I_c のパラメータ群 $(s,k,\theta,\delta,\varepsilon_1,\varepsilon_2)$ は UPRISE における適合度関数の基本パラメータ群であるため，本論文ではこれを Φ として簡略化表現する．

2.2.4 その他の適合度

前述した I_p，I_d，I_c に加えて，我々は従来の文書検索技術における IDF [17] のような，複数キーワード間の特定性の差を考慮した適合度として，すべてのシーンを対象とした特定性や，一連の講義ごとの特定性など，様々な範囲での特定性を考慮できるような適合度 $iafr$（Inverse keyword appeared scene frequency in a range）を提案している [10], [14].

また，我々は講師が用いるレーザポインタの情報を考慮した適合度についても提案を行ってきた [11]．これらの詳細については，[10], [11], [14] を参照されたい．

3. 音声認識と音声情報の有用性

本章では，まず，本研究に用いる音声認識の概要を説明し，音声認識による音声情報の有用性に関して議論する．

3.1 音声認識の概要

本研究での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ほかに，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もつ話し言葉であり，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）[15], [16] 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，バイグラム及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 音声情報の有用性の評価

3.2.1 キーワードごとの音声再現率

本研究では，まず音声認識によって抽出した音声情報がプレゼンテーションコンテンツの検索に利用できるかどうかを検討するために予備実験を行った．

実際にデータベースに関する 90 分の講義を撮影した動画中の音声に対して音声認識を行った．得られたキーワード例を表 1 に示す．ここで，音声出現回数とは，実際に講義を聞いてキーワードが出現していた回数，音声認識回数とは，実際に認識された数である．

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

表 1 キーワードの音声再現率
Table 1 Recall of speech recognition for some keywords.

キーワード	音声出現回数	音声認識回数	湧き出し誤り回数	音声再現率	言語モデルに存在
トランザクション	102	72	0	0.706	○
コミット	51	17	0	0.333	×
アボート	32	12	0	0.375	×
ハッシュ	49	9	1	0.163	×
チェックポイント	40	22	0	0.550	×

わき出し誤り回数とは，実際には音声に出現していないときに，出現していると認識された回数である．また，音声再現率は以下により算出した値である．

$$\text{音声再現率}=\frac{\text{音声認識回数}-\text{わき出し誤り回数}}{\text{音声出現回数}}$$

例として挙げたキーワードのうち，言語モデルに存在していたのは‘トランザクション’のみであり，残りのキーワードは本研究で新たに加えたものである．

表 1 から分かるように，言語モデルに存在する語‘トランザクション’の方が再現率が高い．また，言語モデルに存在していなくても，‘チェックポイント’は 5 割以上認識している．これは，ある程度長い語の方が，似た言葉が少なく，ほかの言葉に間違えられる確率が低くなるためと考える．一方，‘ハッシュ’は著しく再現率が低かった．これは無声音を含み，更に話者の違いによって発音が変わる単語であることが影響していると考える．

3.2.2 シーンごとの出現回数と認識回数の比較

前項において示した音声認識結果が，シーン検索にどの程度有効であるか調査するために，シーンごとの出現回数を数え上げたものとの比較を行った．表 1 の各キーワードの説明が特に集中している回の講義を選び，その回の講義動画を用いて音声認識を行った．これらのキーワードにおける音声出現回数と音声認識回数，音声情報を考慮していないこれまでの適合度 I_p 及び I_d との，ピアソンの乗率相関係数を表 2 に，‘トランザクション’，‘チェックポイント’における音声出現回数と音声認識回数，I_p，I_d の値のシーンごとの比較結果を図 2，図 3 に示す．それぞれの I_d のパラメータ $\theta = 1.0$ とした．

表 2 から，音声認識回数と音声出現回数の間には非常に強い相関関係があることが分かる．このことから，現時点の音声認識精度であっても十分音声情報として利用できることが分かる．

表 2 音声出現回数と認識回数，従来の適合度との相関係数
Table 2 Pearson product-moment correlation coefficient between the number of occurrences in the speech and existing impression indicators.

キーワード	音声出現回数 vs 音声認識数	音声出現回数 vs I_p	音声出現回数 vs I_d
トランザクション	0.942	0.592	0.791
コミット	0.954	−0.011	0.038
アボート	0.890	0.468	0.765
ハッシュ	0.865	0.775	0.958
チェックポイント	0.965	0.017	−0.134

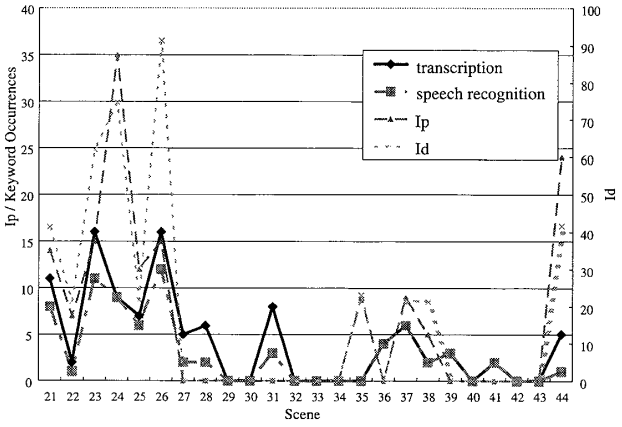


図 2 音声出現回数と従来の適合度の比較（トランザクション）
Fig. 2 Comparison between the number of occurrences in speech and existing impression indicators for ‘transaction.’

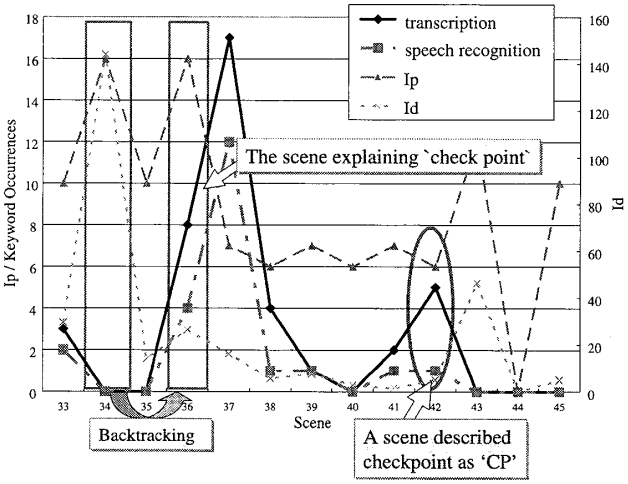


図 3 音声出現回数と従来の適合度の比較（チェックポイント）
Fig. 3 Comparison between the number of occurrences in speech and existing impression indicators for ‘check point.’

3.2.3 音声出現回数独自の特性

更に，表 2 から多くのキーワードにおいて音声出現回数と従来の適合度は強い相関があることも分かる．

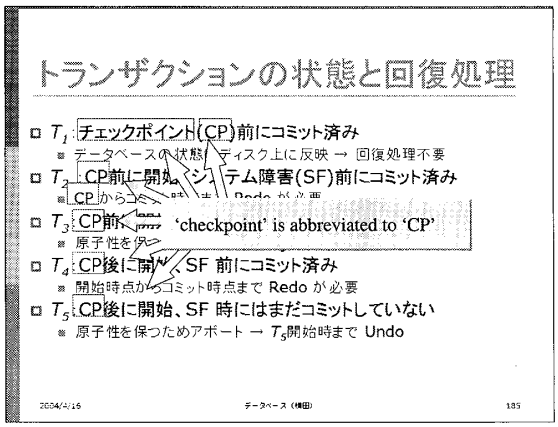


図 4　キーワードが異表記されている例
Fig. 4　An example of variants of keywords.

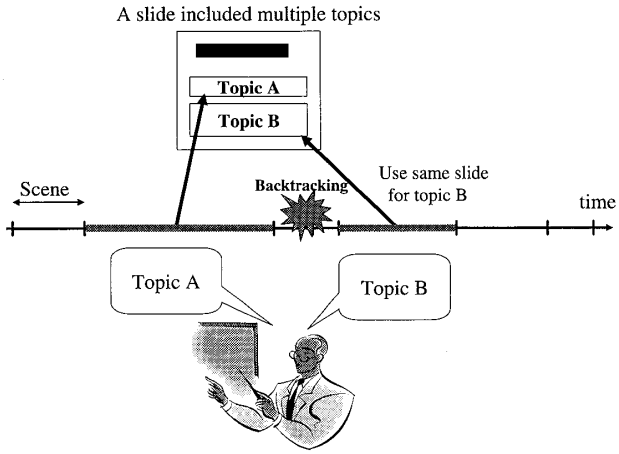


図 5　バックトラックによるシーンの分割
Fig. 5　The scene segmentation caused by backtrack.

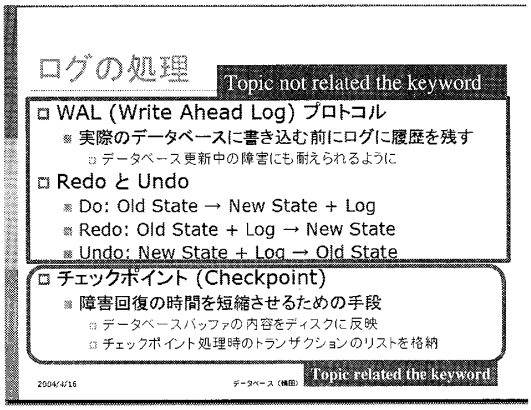


図 6　一つのスライドに複数のトピックが存在するスライドの例
Fig. 6　An example of a slide having multiple topic.

これは，講演者はスライド中の文字を読むことが多く，スライドに出現するときは発話されやすい傾向があるためである．また，シーンの時間が長ければ長いほど，発話される確率が高くなることも理由の一つであると考える．

しかし，キーワード‘コミット’や‘チェックポイント’のように，従来の適合度とほとんど相関のないキーワードも存在する．音声出現回数独自の特性を調べるために図 2，図 3 を細かく見てみると，音声には出現しているが，従来の適合度の値がないシーンがいくつか存在する．これには大きく分けて二つの場合が存在する．

一つは，‘コミット’，‘チェックポイント’などのキーワードがデータベースの講義内で一般的な用語であり，データベースに関連する他の概念を説明する際に用いられる単語であるため，スライド中の文字には出現せずに，図等を説明する際に多く発話されている場合である．

もう一つは，キーワードが異表記されている場合である．例えば，図 4 のように，‘チェックポイント’が‘CP’と略記されているような場合がこれにあたる．このようなシーンは，スライドの構造を考慮して算出する，従来の適合度 I_p ではポイントが小さくなるが，実際には，そのキーワードがそのシーンで重要な意味で用いられていると考える．そこで，音声データを検索に用いれば，各シーンにおける，キーワードの重要度を正確に判断できると考える．

その他の注目すべき特性として，一つのスライドに複数のトピックが存在し，それぞれ別のシーンで説明されているような，シーンの適切な順位付けを行う点がある．UPRISE では，図 5 のように，バックトラックや巻き戻りで同じスライドが複数回出現する場合には，それぞれ別のシーンとして適合度が算出される．従来の適合度においては，このようなシーンを区別するために時間情報や前後情報を用いていた．しかし，これらはシーンの継続時間や前後関係のみ着目しているため，シーンで話されているトピックとは無関係に適合度を決定していた．ここで，音声データを考慮した検索を行えば，どのシーンで，キーワードにあったトピックが話されているかを考慮できる．

例として，図 6 を取り上げる．このスライドには‘チェックポイント’の説明を含む複数のトピックが混在しており，それぞれのトピックを複数のシーンで説明していた．このような例の場合，従来の適合度においては，図 3 からも分かるように，前半部分の‘チェックポイント’と無関係のトピックが話されているシーンと，後半部分の‘チェックポイント’が実際に話され

ているシーンの違いは，時間のみで判定されるため，従来の適合度 I_d を用いると，前半の方が適合度が高かった．これに対し，音声は後半のシーンのみ出現するため，音声データを検索に用いれば後半のシーンの方が重要であると判断できる．よって，このような例の場合，音声データを考慮して検索を行えば，検索精度が向上することが分かる．

これらのことから，

- キーワードの表記揺れの吸収
- 同じスライドを用いているシーンの適切な順位付け

の場合に，音声情報は特に有効である．

4. 音声情報を統合した検索手法

3. で述べたように，予備実験により音声情報を用いることで複数トピックスを扱うスライドを含むシーン間での適切な順位付けやキーワードの表記揺れの吸収ができることが確認できた．本章では，この音声情報を実際に UPRISE の検索に統合する手法に関して説明する．

以下ではまず，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．次に，音声情報の統合方法について考察し，音声情報を統合した新しい適合度を提案する．

4.1 音声情報のデータベースへの格納

前出の 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 7 はその 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 属性をもち，複数の speech タグを含んでいる．speech タグ一つは一つの音声情報を表している．speech 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もとに，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="6" string="っ">
<voice in="0" consTimeMilli="6880" consTimeSec="6" string="っ" />
</fragment><fragment sec="7" string="ん">
<voice in="6" consTimeMilli="250" consTimeSec="0" string="ん" />
</fragment><fragment sec="9" string="ですって">
<voice in="7" consTimeMilli="230" consTimeSec="0" string="です" />
<voice in="7" consTimeMilli="1910" consTimeSec="1" string="って" />
</fragment><fragment sec="10" string="今日の内容">
<voice in="9" consTimeMilli="340" consTimeSec="0" string="今日" />
<voice in="9" consTimeMilli="140" consTimeSec="0" string="の" />
<voice in="9" consTimeMilli="490" consTimeSec="0" string="内容" />
</fragment><fragment sec="11" string="は分かれ">
<voice in="10" consTimeMilli="180" consTimeSec="0" string="は" />
<voice in="10" consTimeMilli="490" consTimeSec="0" string="分かれ" />
```

図 7 認識結果のログファイルから生成した XML の例
Fig. 7 An example of XML generated from a logfile of speech recognition.

4.2 音声情報の統合方法

3.2 で述べたように，音声情報は同じスライドに複数のトピックが存在するような場合の適切な順位付けなどに特に有効であるため，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音声の情報を統合することで，これまで適切に順位付けができなかったシーンの順位付けへの効果が期待できる．しかし，我々は音声情報として，自動音声認識技術によって生成されたものを利用するため，誤認識の影響を考慮する必要がある．本研究では，誤認識の影響を削減するために，該当シーンや近傍シーンのスライド中の文字列に出現するかどうかの条件を用い，音声情報を統合する．

統合に際しまず，あるシーン s にキーワード k が発話された回数を $skc(s,k)$ (Spoken Keyword Count) とおく．本研究で提案する手法では，この $skc(s,k)$ と，スライド中の文字情報を考慮するポイント $I_p(s,k)$ とを統合する．音声の情報は各シーンにおけるスライド文字列の情報とは独立しているため，近傍シーンへの影響計算の方法を考慮する必要がある．本研究では，近傍シーンへ与える影響を計算する際に，あるキーワードがシーン中で発話された影響と使用しているスライド中に存在する影響を個別に考慮する方法と，スライド中の文字情報と区別せずにまとめて考慮する方法の 2 種類を提案する．

また，単に音声の登場によるポイントをこれまでの適合度に加算すると，スライド中にキーワードが存在しなくてもそのシーンは検索結果となる．この場合，わき出し誤りが起こると，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．

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．この組合せにより **4.3** 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 [17] や我々が提案してきた iafr 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際に，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．本研究では，iafr と同様に特定性を考慮する際に，すべてのシーンを対象とした特定性だけでなく，様々な範囲での特定性を考慮できるよう，音声情報の特定性を考慮した関数 $isfr(k,\lambda)$（Inverse keyword spoken scene frequency in a range）を，S_λ を範囲 λ に含まれるシーンの集合，$S_{\lambda/spoken}(k)$ を範囲 λ 内で，キーワード k が発話されたシーンの集合として，以下のように定義する．

$$isfr(k,\lambda) = \log \frac{|S_\lambda|}{|S_{\lambda/spoken}(k)|}$$

λ は，全シーンを表現する *all-scenes* の他に，講義コンテンツであれば，講義科目単位 (*course*)，1 回の講義単位 (*class*) などの範囲指定パラメータを用いる．この $isfr(k,\lambda)$ 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．

4.3 適合度の提案

4.2 で示した統合方法の組合せにより，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 $I_{c+skc/p}$ として以下のように提案する．

$$I_{c+skc/p}(\Phi,\psi) = \begin{cases} I_c(\Phi) + \psi \cdot T(s)^\theta \cdot skc(s,k) & (I_p \neq 0) \\ I_c(\Phi) & (I_p = 0) \end{cases}$$

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

次に，$I_{c+skc/p}$ の条件を少し緩め，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，$I_{c+skc/c}$ を以下のように定義する．

$$I_{c+skc/c}(\Phi,\psi) = \begin{cases} I_c(\Phi) + \psi \cdot T(s)^\theta \cdot skc(s,k) & (I_c \neq 0) \\ 0 & (I_c = 0) \end{cases}$$

次に，$I_{c+skc/c}$ とは逆に，該当シーンのスライドにキーワードが登場する場合に，音声の影響を近傍シーンのものも含めて考慮する適合度として，$I_{c[p+skc/p]}$ を以下のように定義する．

$$I_{c[p+skc/p]}(\Phi,\psi) = \begin{cases} I_{c[p+skc]}(\Phi,\psi) & (I_p \neq 0) \\ I_c(\Phi) & (I_p = 0) \end{cases}$$

ただし，

$$I_{c[p+skc]}(\Phi,\psi) = \sum_{\gamma=s-\delta}^{s+\delta} E(\gamma - s, \varepsilon_1, \varepsilon_2) \cdot T(\gamma)^\theta \cdot \{I_p(\gamma,k) + \psi \cdot skc(\gamma,k)\}$$

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，音声の影響を近傍シーンのものも含めて考慮する適合度として，$I_{c[p+skc/c]}$ を以下のように定義する．

$$I_{c[p+skc/c]}(\Phi,\psi) = \begin{cases} I_{c[p+skc]}(\Phi,\psi) & (I_c \neq 0) \\ 0 & (I_c = 0) \end{cases}$$

また **4.2** で述べたように，音声加算部分に $isfr$ を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．

音声の特定性に加え **2.2.4** で述べた $iafr$ を用いスライド内テキスト上のキーワードの特定性を考慮する適合度を $I_x \cdot iafr \cdot isfr$ とし，以下のように提案する．ただし，x は $c+skc/p$，$c+skc/c$，$c[p+skc/p]$，$c[p+skc/c]$ のいずれかとする．

$$I_x \cdot iafr \cdot isfr(\Phi,k,\lambda)$$

$$= \begin{cases} I_c(\Phi) \cdot iafr(k,\lambda) & (I_p \cdot I_c = 0) \\ I_c(\Phi) \cdot iafr(k,\lambda) & \\ +\psi \cdot I_x \cdot isfr(k,\lambda) & (\text{otherwise}) \end{cases}$$

5. 実　　験

本研究では，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った．以下ではその実験に関して説明し，実験結果に対して考察を行う．

5.1 実験に用いたデータ

実際の講義をコンテンツ化したプレゼンテーション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．認識の辞書を講義ごとに変更した理由は，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，各シーンごと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わ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．

5.2 実　　験

提案手法の評価を行うため，**5.1**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$I_{c+skc/p}$，$I_{c+skc/c}$，$I_{c[p+skc/p]}$，$I_{c[p+skc/c]}$のみで評価実験を行う．更に，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，$iafr$と$isfr$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もとで行った．

- パラメータは$\theta = 0.5$，$\delta = 4$，$\varepsilon_1 = 5.0$，$\varepsilon_2 = 0.5$，$\lambda = course$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ごととした．
- 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．
- $I_{c+skc/p}$，$I_{c+skc/c}$，$I_{c[p+skc/p]}$，$I_{c[p+skc/c]}$のパラメータψの値を0から10まで1刻みで変化させたときと，0から100まで5刻みで変化させたときの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，$iafr$と$isfr$を考慮した適合度を用意し，ψの値を0から20まで5刻みで変化させて調査を行った．

パラメー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．今回はパラメータを変更しなかったが，パラメー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ごととした理由は，今回は音声認識を各講義ごとに行ったからであるが，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．

評価に際しては，今回の実験では，正解シーンを各キーワードに対して一つとしていることから，平均逆数順位（Mean reciprocal rank：MRR）を用いた[22]．MRRは質問応答システムの評価に用いられることが多く[23]，質問ごとに最初に出現した正解の順位の逆数を求め，それらを全質問にわたって平均することで定義される．

本実験のMRRは，Nを検索回数とすると以下の式で求めることができる．

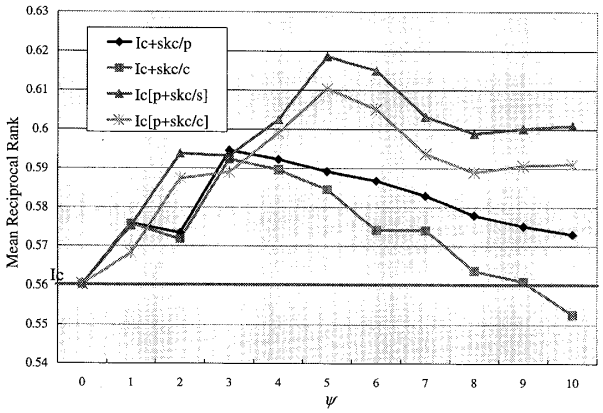


図 8　ψ を変化させたときの各適合度による MRR の推移（1 刻み）

Fig. 8　Comparison of the four indicators varying ψ per 1 point.

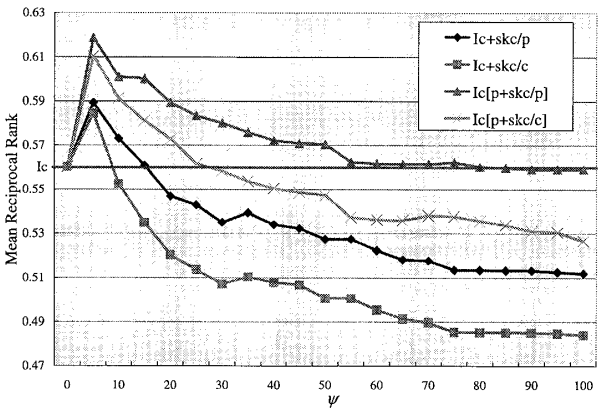


図 9　ψ を変化させたときの各適合度による MRR の推移（5 刻み）

Fig. 9　Comparison of the four indicators varying ψ per 5 points.

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

MRR は，すべて 1 位の際には 1 であるが，すべて 2 位の際には 0.5 と，情報検索分野で評価に際して指標となる正解が複数ある場合の適合率（precision）[24] と比べると比較的低い値になる傾向がある点に注意されたい．

5.3　実験結果と考察

5.3.1　統合方法による適合度の比較実験

図 8 は ψ を変化させたときの各適合度による MRR である．

従来の適合度 I_c による MRR は 0.560（$\psi = 0$ のとき）であるため，各適合度による MRR は I_c より最高で $I_{c+skc/p}$ が 0.035，$I_{c+skc/c}$ が 0.032，$I_{c[p+skc/p]}$ が 0.059，$I_{c[p+skc/c]}$ が 0.05 上回った．これにより，音声情報を考慮することによって検索の精度が上昇していることが分かる．その一方で，$I_{c+skc/c}$ では $\psi = 9, 10$ において従来の適合度 I_c による MRR を下回った．これは音声情報を過剰に考慮すること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 ψ を 0 から 5 刻みで 100 まで変化させた図 9 からも明らかである．最も精度の上がっていた $I_{c[p+skc/p]}$ であっても ψ が 80 より大きくなると，従来の適合度 I_c による MRR 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きてしまうためと考える．図 8 より，音声情報の比率は $\psi = 5$ 程度が最適だと考える．本実験では，$\psi = 6$ の際に発話 1 回が I_p におけるタイトル 1 回の出現と同等になる．

表 3　‘チェックポイント’ での検索結果

Table 3　The ranks in search results for ‘checkpoint.’

適合度	正解の順位	上位のシーン
I_c	3	34, 33, **36**
skc のみ	2	37, **36**
$I_{c+skc/p}$	1	**36**, 37, 34
$I_{c+skc/c}$	3	37, 33, **36**, 34
$I_{c[p+skc/p]}$	1	**36**, 37, 34
$I_{c[p+skc/c]}$	3	37, 33, **36**, 34

各統合方法を比較すると，まず音声の前後関係を考慮したもの ($I_{c[p+skc/x]}$) と，そうでないもの ($I_{c+skc/x}$) では，前者が MRR で上回っていることが分かる．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．また，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 (skc/p) 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 (skc/c) の比較では，前者の方が MRR で上回っている．これは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，偶然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，**3.2.3** で示した特徴を，提案する適合度が適切に考慮できているかどうかを確認するために，検索語 ‘チェックポイント’ で検索した場合，正解シーン（シーン 36）と同一スライドのシーン（シーン 34）が提案した四つの適合度，I_c，skc のみ においてどのように順位が変化するかを調査した（表 3）．

表 4 キーワードの順位の例 ($I_{c[p+skc/p]}$)
Table 4 The ranks in search results for several keywords.

キーワード	$\psi = 0$	$\psi = 10$	$\psi = 100$	$\psi = 10000$
スキーマ	14	2	2	2
ハザード	3	10	17	17
集合，和	7	11	13	14

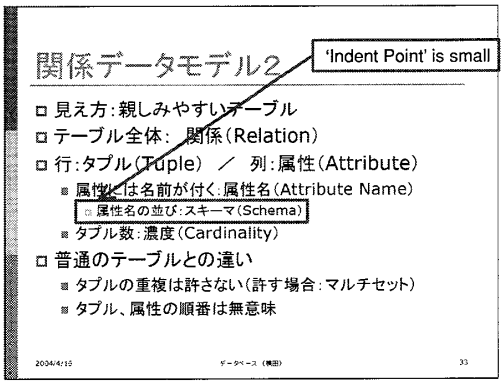


図 10 検索語：'スキーマ' の正解シーン
Fig. 10 A correct scene for keyword 'schema.'

図 3 とこの表から，提案手法を用い音声を統合した場合は，同一スライドを用いているシーン 34 より，正解シーンであるシーン 36 が上位にきており，I_c や skc のみでは適切に順位付けできなかった問題を解決していることが分かる．

なお，skc/c によって音声を考慮した場合は，skc 値の高いシーン 37 と I_c で上位のシーン 33 の影響を受け，I_c と同等の順位となっているが，この点は前後関係を考慮する際の重み付けを調整することでより向上させることができると考える．

次に，最も精度の良かった適合度 $I_{c[p+skc/p]}$ において，ψ 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 4 は，'スキーマ'，'ハザード'，'集合，和' という三つのキーワードについて ψ 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，ψ が 10 以上になれば順位が 2 位に上昇している．これは，図 10 のようにスキーマがスライド上で重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．

同様に，'集合，和' 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．'和' は全シーン中 7 回発話されているのに対して，'集合' は 68 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 '集合' 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．特定性の考慮に関する実験は **5.3.3** にて後述する．

5.3.2 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った音声情報を用いていた．本項では，音声認識における誤認識を考慮した提案手法の有効性を確認するために，実際に人手で書き起こした音声認識を用いた場合との比較実験を行う．

実験のパラメータや評価方法については **5.2** 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，**5.2** 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 $I_{c[p+skc/p]}$ を用いて実験を行い，音声の影響度を決めるパラメータ ψ は 1 から 10 まで 1 刻みで変化させた．

$I_{c[p+skc/p]}$ を用いた比較実験の結果を図 11 に示す．グラフから，最大値はそれほど変化せず，提案手法は音声認識結果を利用した場合でも書き起こしと同程度の精度が得られていることが分かる．このことから，提案する統合方法を用いれば，非常に作業コストが高い書き起こしを行わずとも，音声情報を利用することが可能となるといえる．

また，音声認識処理の結果を用いた場合は，音声影

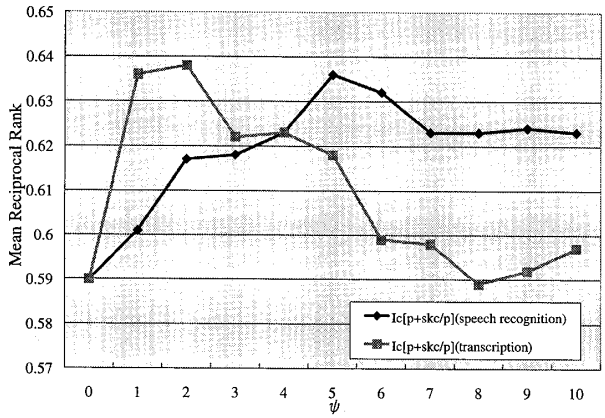


図 11 音声認識と書き起こし情報の比較 ($I_{c[p+skc/p]}$)
Fig. 11 Comparison between speech recognition and transcription using $I_{c[p+skc/p]}$.

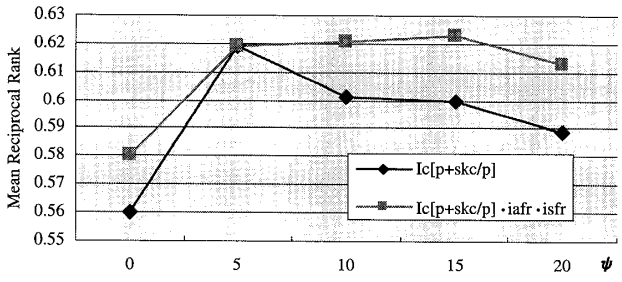


図 12 音声の特定性を考慮した適合度による MRR
Fig. 12 Comparison of Mean reciprocal rank of three impression indicators considering the specificity of speech.

響度パラメータ $\psi = 5$ のときに最大値 0.636 をとっているのに対し，書き起こし情報を用いた場合は $\psi = 3$ のときに最大値 0.638 をとっている．最大値をとるパラメータの値が小さくなった理由として，書き起こしを用いた場合は，音声認識で認識できていなかった発話回数分のキーワード数が増加したためであると考える．

5.3.3 特定性を考慮した適合度の比較実験

音声の特定性を考慮した適合度の有効性を確認するために，適合度 $I_{c[p+skc/p]}$ と $I_{c[p+skc/p]} \cdot iafr \cdot isfr$ との比較実験を行った．実験のパラメータや評価方法は **5.2** と同様とし，ψ を 0 から 20 まで 5 刻みで変化させた．図 12 はその結果を示したグラフである．

特定性を考慮した適合度 $I_{c[p+skc/p]} \cdot iafr \cdot isfr$ は全体的に適合度 $I_{c[p+skc/p]}$ を上回り，$\psi = 15$ のときには従来の適合度 I_c と比べて 0.063 (6.3%) 高い，MRR: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．

次に，特定性を考慮し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 5 は，‘集合，和’，‘拡張，ハッシュ’，‘関係，データモデル’ という三つのキーワードについて $\psi = 5$ 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における検索順位を示したものである．

表 5 特定性を考慮した適合度によるキーワードの順位の例 ($\psi = 5$)
Table 5 The rank of search result for some keywords by impression indicators considering the specificity.

キーワード	$I_{c[p+skc/p]}$	with $iafr \cdot isfr$
集合，和	11	8
関係，データモデル	3	2
拡張，ハッシュ	5	4

表 6 キーワードの音声情報への出現数の例
Table 6 The number of occurrences in speech for some keywords.

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
集合	和	68	7
拡張	ハッシュ	4	22
関係	データモデル	161	37

また，表 6 は各キーワードごとに，音声認識結果中の出現数を比較したものである．これらの三つの例のように，各キーワードの出現数の差が大きい場合は，提案手法によって特定性を考慮することで表 5 のように順位が改善するものと考える．

6. 関連研究

本研究の対象である講義講演のプレゼンテーションコンテンツに対し，音声認識技術を利用する既存研究は非常に少なく，音声認識による音声情報を検索に利用することを目的とした試み [25], [26] がなされているが，音声データによるシーン分割にとどまっており，その情報を用いて検索を行うに至っていない．

プレゼンテーションコンテンツに対して音声認識を利用した検索を行う既存研究に，オンデマンド講演システム LODEM [27] がある．LODEM では，資料テキストの情報と発話内容を利用して，パッセージと呼ばれる発話のまとまり単位での検索を可能としている．しかしながら，本研究と異なり資料テキストの前後情報や，パッセージの長さを考慮していないため，バックトラックが起こるなどして，同じ資料を使用した場合には適切なパッセージを選択できないと考える．また，音声の誤認識への対応もなされていない．

なお，対象ビデオ中のシーン情報に加え，音素イン

デックスファイルを生成し，検索に利用する研究[28]もなされており，製品化もされているが，性能評価などの報告はなされていない．

7. むすび

7.1 まとめ

本論文ではUPRISEの検索精度を向上させるために，音声認識によって抽出した講義・講演の音声情報を統合した検索手法を提案した．まず，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．次にキーワードの音声認識回数と音声出現回数を比較することで，現在の音声認識精度であっても検索に十分に利用可能であることを示した．また，音声認識への出現回数と従来の適合度の値を比較することにより，音声データの有用性を示した．

更に音声情報の有用性を十分発揮できるように，音声情報の統合方法について考察し，その統合方法に基づいて，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，すべての適合度で従来の適合度 I_c(MRR:0.560) よりもMRRが上昇し，適合度 $I_{c[p+skc/p]}$ においては最大0.059（5.9%）改善し，MRR:0.619を記録した．また講義を実際に聞き取り書き起こした音声情報を用いた結果と比較し，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ることが可能となり，非常に作業コストが高い書き起こしを行わずとも，音声情報を有効に利用することが可能となることを示した．

更に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(MRR:0.560) に比べて最大0.063（6.3%）改善し，MRR:0.623を記録した．

7.2 今後の課題

本研究の今後の課題を以下に述べる．まず，単語1語のみの場合の特定性関数の問題がある．IDF [17] などの特定性の関数は，単語1語のみの検索の場合は影響しない．ところが $isfr$ の場合，音声部分のみに特定性の関数を乗算するために，単語1語のみの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，複数の単語による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．

また，今回の実験ではパラメー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ータや音声情報の比率パラメータを変化させることができれば，精度が更に上昇すると考える．更に，今回の実験における評価方法では正解シーンを一つとしていた．しかし，実際の検索においては正解シーンが一つとは限らない．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御協力頂いた，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，及び東京工業大学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] 森本容介，室田真男，清水康敬，"教育用動画像検索システムと時間情報同期方法の開発，" 信学論（D-I），vol.J88-D-I, no.10, pp.1515–1524, Oct. 2005.

[5] 戈 指夷，角谷和俊，"遠隔会議システムにおける資料操作ログに基づくアーカイブコンテンツ作成支援方式，" データ工学ワークショップ論文集，ISSN 1347-4413, DEWS2006–6C–i5., March 2006.

[6] 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，" 情処学研報，DBS-125-58, July 2001.

[7] H. Yokota, T. Kobayashi, T. Muraki, and S. Naoi, "UPRISE: Unified presentation slide retrieval by impression search engine," IEICE Trans. Inf. & Syst., vol.E87-D, no.2, pp.397–406, Feb. 2004.

[8] 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，" 信学論（D-I），vol.J88-D-I, no.3, pp.715–726, March 2005.

[9] 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，DEWS2004-1-B-3., March 2004.
[10] H. Okamoto, T. Kobayashi, and H. Yokota, “Presentation retrieval method considering the scope of targets and outputs,” Proc. WIRI2005, pp.47–52, April 2005.
[11] W. Nakano, Y. Ochi, T. Kobayashi, Y. Katsuyama, S. Naoi, and H. Yokota, “Unified presentation contents retrieval using laser pointer information,” Proc. SWOD, pp.170–173, April 2005.
[12] 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，” 情処学研報，2005-DBS-137-78, July 2005.
[13] 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，“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，” データ工学ワークショップ論文集，ISSN 1347-4413, DEWS2006-6C-o1, March 2006.
[14] H. Yokota, T. Kobayashi, H. Okamoto, and W. Nakano, “Unified contents retrieval from an academic repository,” Proc. International Symposium on Large-scale Knowledge Resources LKR2006, pp.41–46, 2006.
[15] 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/
[16] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, “Spontaneous speech corpus of Japanese,” Proc. LREC2000, vol.2, pp.947–952, Athens, Greece, May 2000.
[17] G. Salton, Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley, 1988.
[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, Dec. 2000.
[20] T. Shinozaki, C. Hori, and S. Furui, “Towards automatic transcription of spontaneous presentations,” Proc. Eurospeech2001, vol.1, pp.491–494, Aalborg, Denmark, Sept. 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，Feb. 2005.
[22] 酒井哲也，“よりよい検索システム実現のために，” 情報処理，vol.47, no.2, pp.147–158, Feb. 2006.
[23] 岸田和明，岩山　真，江口浩二，“検索実験の方法と実際：Ntcir ワークショップでの試み，” Pre-meeting Lecture at the NTCIR-3 Workshop, Oct. 2002.
[24] D.A. Grossman and O. Frieder, Information Retrieval Algorithm and Heuristics. Kluwer, 1998.
[25] 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, Feb. 2005.
[26] G.J.F. Jones and R.J. Edens, “Automated alignment and annotation of audio-visual presentations,” Proc. 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'02), pp.276–291, Springer-Verlag, Rome, Italy, Sept. 2002.
[27] A. Fujii, K. Itou, and T. Ishikawa, “Lodem: A system for on-demand video lectures,” Speech Commun., vol.48, no.5, pp.516–531, May 2006.
[28] Fuji Xerox, ビデオアクセス技術 (MediaDEPO). http://www.fujixerox.co.jp/company/technical/mediadepo/

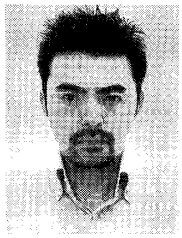
（平成 18 年 5 月 16 日受付，8 月 25 日再受付）

岡本　拓明

平 16 東工大・工・情報工卒．平 18 同大大学院・情報理工・計算工・修士課程了．複合メディアコンテンツの管理・検索に関する研究に従事．

仲野　亘

平 17 東工大・工・情報工卒．同年より同大大学院・情報理工・計算工・修士課程在学中．複合メディアコンテンツの管理・検索に関する研究に従事．日本データベース学会学生会員．



小林　隆志　（正員）

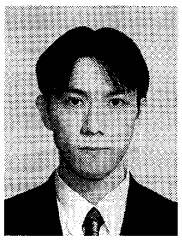
平 9 東工大・工・情報工卒．平 16 同大大学院・情報理工・計算工・博士課程了．平 14 より同大学術国際情報センター助手，現在に至る．工博．ソフトウェア開発方法論，ソフトウェア再利用技術，複合メディアコンテンツの管理・検索，Web サービス連携などの研究に従事．情報処理学会，日本ソフトウェア科学会，日本データベース学会，ACM 各会員．

直井　聡　（正員）

昭 58 慶大・工・電工卒．昭 60 同大大学院・工・電気工学・修士課程了．同年より（株）富士通研究所．現在，部長．平 13 東工大・学術国際情報センター・客員助教授．平 17 より同センター・客員教授．工博．文字パターン処理，画像処理，文字認識や e ラーニングの研究に従事．情報処理学会会員．

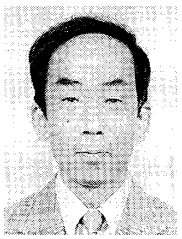
横田　治夫　（正員）

昭 55 東工大・工・電物卒．昭 57 同大大学院・情報・修士課程了．同年富士通（株）．同年 6 月（財）新世代コンピュータ技術開発機構研究所（ICOT）．昭 61（株）富士通研究所．平 4 北陸先端大・情報・助教授．平 10 東工大・大学院情報理工・助教授．平 13 より同大学術国際情報センター教授，現在に至る．工博．主として分散インデキシング，データ工学向けアーキテクチャ，高機能ストレージシステム，ディペンダブルシステム等に関する研究に従事．日本データベース学会理事．ACM SIGMOD 日本支部評議委員．情報処理学会，人工知能学会，IEEE，ACM 各会員．



岩野　公司　（正員）

平 7 東大・工・電子情報卒．平 12 同大大学院・工学系・情報工・博士課程了．同年東工大・大学院情報理工・計算工・助手，現在に至る．工博．音声認識，話者認識，音声合成，マルチメディア情報処理等の研究に従事．IEEE，ISCA，情報処理学会，日本音響学会各会員．



古井　貞熙　（正員：フェロー）

昭 43 東大・工・計数卒．昭 45 同大大学院修士課程了，NTT 研究所入社．ベル研究所客員研究員，NTT 基礎研究所第四研究室長，ヒューマンインタフェース研究所音声情報研究部長，古井特別研究室長を経て，現在，東京工業大学大学院情報理工学研究科計算工学専攻教授．工博．音声認識，話者認識，音声合成，マルチメディア情報処理等の研究に従事．IEEE，米国音響学会（ASA）各フェロー．ISCA，情報処理学会，日本音響学会等各会員．