2005－DBS－137（Ⅱ）（78）
2005／7／15

講義講演シーン検索における音声データの利用

岡本拓明[†]　小林隆志[††]　直井聡[†††,††]
横田治夫[†,††]　古井貞熙[†]

我々はこれまで，講義講演のプレゼンテーションにおける資料とその撮影動画をメタデータによる統合コンテンツとして蓄積し，その特性を利用して高度なシーン検索を提供するUPRISEを提案してきた．本稿では，このUPRISEのシーン検索機能の性能を向上させるために，講義講演の動画から抽出した音声データを利用する手法を提案する．音声認識エンジンにより抽出を行った音声データと従来の適合度手法をシーン毎に比較することにより，シーン検索に音声データを利用することの有用性を示す．

Application of Voice Data for Retrieving Unified Presentation Contents

Hiroaki OKAMOTO,[†] Takashi KOBAYASHI,[††] Satoshi NAOI,[†††,††]
Haruo YOKOTA[†,††] and Sadaoki FURUI[†]

We have proposed unifying presentation contents, such as lecture video and presentation slides used in lectures, using metadata. For the unified contents, we have also proposed an advanced searching scene mechanism named UPRISE(Unified Presentation Slide Retrieval by Impression Search Engine). In this paper, for the performance , we pay attention to voice data abstracted from lecture videos. To confirm the availability of the voice data, we compared voice data with existing impression indicators.

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[4–10]．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，レ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，その音声情報をシーンの差別化を行うために利用することを考える．音声認識の結果を利用した検索を行う既存研究に，音声認識結果を動画のインデクスとして利用し検索を行うもの[11–15]があり，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

† 東京工業大学 大学院 情報理工学研究科 計算工学専攻
Department of Computer Science Graduate School of Information Science and Engineering in Tokyo Institute of Technology
†† 東京工業大学 学術国際情報センター
Global Scientific Information & Computing Center in Tokyo Institute of Technology
††† 株式会社 富士通研究所
FUJITSU LABORATORIES LTD.

る音声データを検索に利用する試みが，[16, 17] などでなされている．しかしながら，音声データによるシーン分割にとどまっており，その情報を用いた検索には至っていない．

話し言葉研究用のデータベース (CSJ) [18, 19] 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，実用的な音声認識精度を達成できるようになってきており，発話される音声はシーンによって異なるため，シーンごとの特性が音声データに現れると考える．また，発話する内容はスライドにまとめられているため，スライド構造を考慮している従来の適合度と連動していると予想でき，時間が長いシーンのほうが発話される確率が高くなるため，時間情報を考慮している従来の適合度とも連動していると予想できる．その事から音声データを考慮した適合度と統合した場合でも，従来の適合度と反発し合うことなく，シーンの差別化を行うことができると考える．

本稿では，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も，2.2 節で簡単に述べる．3 節では，本研究に用いる音声認識の概要と，キーワードに対する音声再現率を検証する．4 節においては，代表的なキーワードに対する，シーンごとの音声出現数と従来の適合度を比較することにより，音声データを考慮した適合度が有用であることを示し，4.2 節では，音声データの今後の利用可能性を述べる．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．

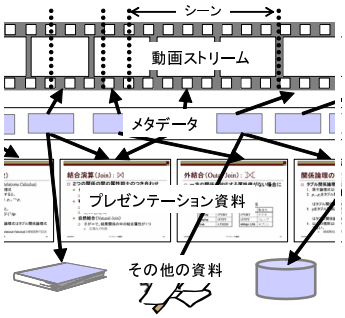

図 1 プレゼンテーション資料と動画の統合

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 従来の適合度算出手法

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 その他の適合度

UPRISE では前述した I_p，I_d の他に，シーンの前後関係を考慮した適合度 I_c と，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) を考慮した適合度を用いている．詳細については，[6, 9] を

参照されたい．

3. 音声データのデータベースへの格納

2 節で説明した UPRISE においては，バックトラックを起こしたり巻き戻りがあれば，違うシーンとして適合度が計算される．このようなシーンの差別化を行うために，2.2.2 節で説明した時間情報を考慮した適合度 I_d や，前後関係を考慮した適合度 I_c を用いていたが，たとえシーンで用いられていなくても，シーンの時間が長いシーンのほうが重要であるとされてしまうという問題があった．そこで，本稿では音声データを検索に利用することで，この問題を解決することを目指す．

音声データを検索に利用するためには，音声認識を行い，講義講演の動画から音声データを抽出し，データベースに格納する必要がある．

この節では，3.1 節で，本研究で行う音声認識の概要と，音声認識を行った結果を検索に利用できる形式への変換の過程を示す．さらに，3.2 節では，実際に音声認識実験を行った結果を基に，キーワードに対する音声認識率について考察する．

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 [20, 21] 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され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 [22, 23]. そこで，本研究では山崎が [24] で作成した言語モデルと音響モデルを用いる． [24] では, 日本語話し言葉コーパス (CSJ) [18, 19] 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 [25]，形態素解析用の辞書として，ipadic [26] 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, 大学内のデータベース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．単語辞書として, 山崎が [24] において作成した辞書に，実験に用いた講義の資料より抽出を行った 1099 語のうち，辞書に含まれていない名詞 134 語を，未知語として追加登録した，22,994 語を登録した辞書を使用する．

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，単語と単語の出現した時間を計算し，XML ファイル形式に変換する．その例を図 2 に示す．このファイルを用いて，キーワード名とキーワードの出現した時間を，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．

表 1 キーワードの音声再現率

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	×

```
<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 2 認識結果のログファイルから変換した XML 例

3.2 キーワードに対する音声再現率の評価

ここで，実際に 90 分の講義を撮影した動画に対し，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．湧き出し誤り回数とは，実際には音声に出現していない時に，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．

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

例として挙げたキーワードの内，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未知語として加えたものである．

表 1 からわかるように，言語モデルに存在する語’

トランザクション’ のほうが再現率が高い．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、ある程度長い語のほうが，似た言葉が少なく，他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，さら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．

また，キーワード例の一つである’ コミット’ が，実際に音声に出現していた数と音声認識された数の比較のグラフを，図 3 に示す． このグラフから分かるよ

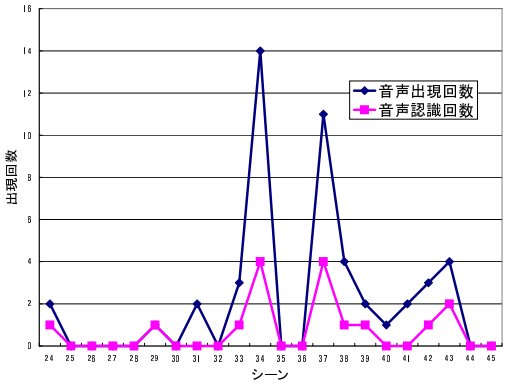

図 3　音声出現回数と音声認識回数の比較 (コミット)

うに, 認識回数と実際の出現回数のグラフは増減の傾向が同じであり, 他のキーワードでも同様であった. このことから現在の再現率であっても，認識結果の増減の傾向を利用することで，実際の音声出現回数と同等の情報として，十分利用できることがわかる．

CSJ は日本語の自然発話音声を大量に集めた話し言葉研究用のデータベースであり，質・量ともに世界最大である [24]．そのため，話し言葉に対する認識率は高い．しかしながら，ユーザの検索要求に出現するような，専門用語は言語モデルに存在しない事が多く認識率が低いため，専用の言語モデルを作成する等のチューニングが必要であると考える．しかし，本研究ではチューニングは行わずに，音声データを考慮した検索の有用性を考える．

4.　音声データの検索への利用

以下ではまず 4.1 節において，各キーワード例に対し，音声データの有用性の評価を行う．その評価を踏まえて，4.2 節において，今後の検索への利用可能性について述べる．

4.1　音声データの有用性の評価

認識率がさらに向上した場合の影響を見るため，本研究では 90 分の講義音声を実際に聞き取り，シーンごとの出現回数を数え上げたものとも比較を行う．ユーザの検索要求にあると想定したキーワードの内，今回は’ トランザクション’，’ コミット’，’ アボート’，’ ハッシュ’，’ チェックポイント’ を例として取り上げた．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なった．その出現回数と従来の適合度の値の比較を図 4〜図 8 に示す．実際に音声を聞いてシーン毎に数え上げを行なったものを音声出現回数，正しく認識された回数をシーン毎に数え上げたものを音声認識回数とする．また，シーン毎の従来の適合度 I_p，I_d の値を比較に用いる．I_d のパラメータ $\theta = 1.0$，$u = 60$ とした．

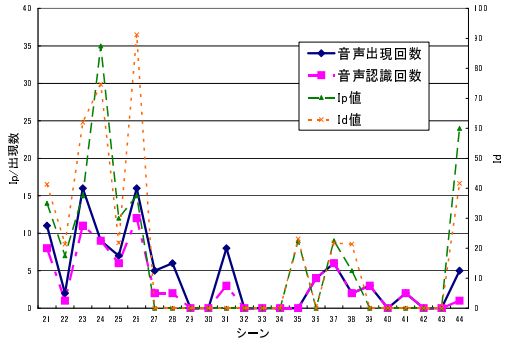

図 4　音声出現回数と従来の適合度の比較 (トランザクション)

それぞれの図から，音声の出現回数は従来の適合度と連動している事を確認した．このことは，講演者はスライド中の文字を読むことが多いために，スライドに出現する時は，発話されやすい傾向があることが理由である．また，シーンの時間が長ければ長いほど，

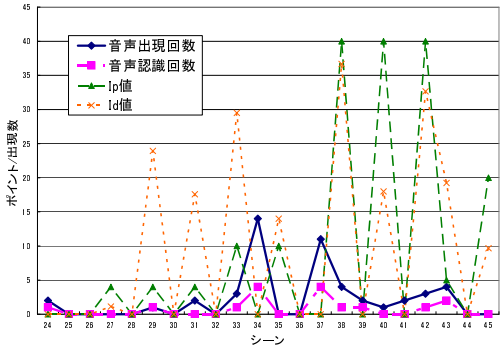

図 5　音声出現回数と従来の適合度の比較 (コミット)

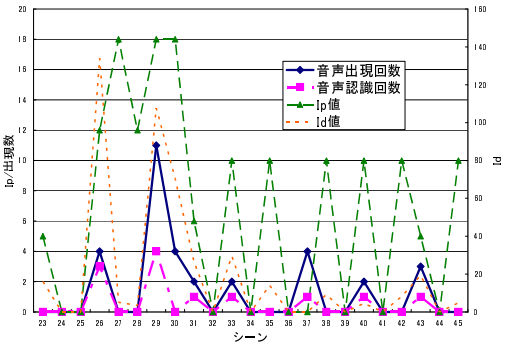

図 6　音声出現回数と従来の適合度の比較 (アボート)

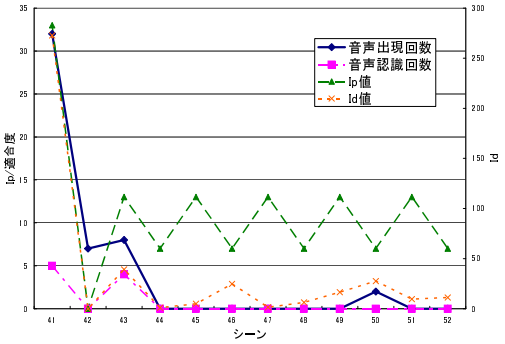

図 7　音声出現回数と従来の適合度の比較 (ハッシュ)

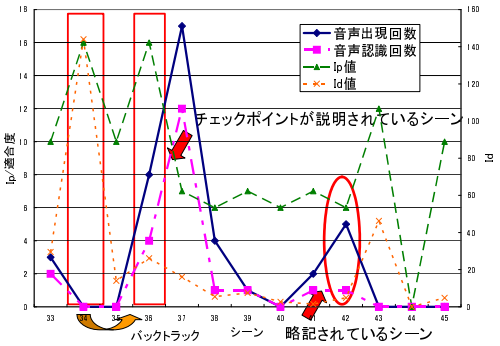

図 8　音声出現回数と従来の適合度の比較 (チェックポイント)

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．

ここで図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幾つか存在する．これには大きく分けて 2 つの場合が存在する．

1 つは，そのキーワードが’コミット’，’アボート’，’チェックポイント’ などの説明に多く用いられる単語であるため，スライド中には出現せずに，図を説明するシーンに多く用いられている場合である．

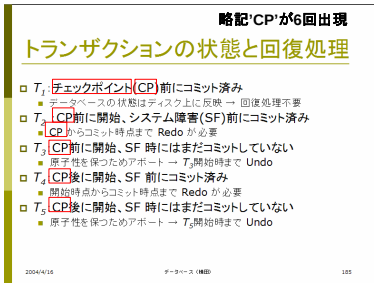

図 9　キーワードが異表記されている例

もう 1 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 9 のように，’チェックポイント’ が’CP’ と略記されていたり，’ハッシュ’ が’ ハッシング’ と表記されている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，このように対応する辞書を持たなくても，キーワードの表記揺れを吸収し，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．

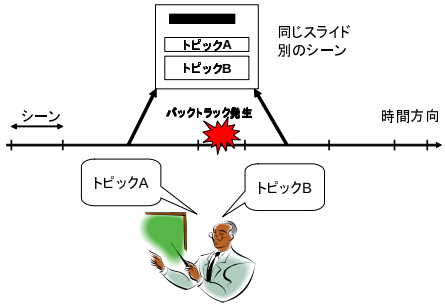

図 10　バックトラックによるシーンの分割

その他の注目すべき特性として，1 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，シーンの差別化がある．UPRISE では，図 10 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．

図 11 1 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

例として，図 11 を取り上げる．このスライドには全て’チェックポイント’に関係するトピックではなく，複数話されているトピックの 1 つとして’チェックポイント’の定義が説明されていた．このような例の場合，従来の適合度においては，図 8 からもわかるように，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，前半のほう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ほう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わかる．

これらの事から，

- 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差別化

の場合に，特に有効であることを確認した．

4.2 今後の音声データの利用可能性

この節では，4.1 節で述べた音声データの特性以外に，今後検索に利用可能であることについて述べる．

1 つめは音声による特定性の考慮である．現在のシステムでは，主にキーワードが目的のシーンを特定する性質 (特定性) の考慮として，プレゼンテーション中での出現頻度を用いていた [9]．しかし，講義講演においては，資料は簡潔にまとめられている場合が多く，出現頻度の差はあまり大きくならない傾向がある．そこで，音声のデータへの出現回数を用いて頻度を考慮することで，出現頻度の差を明確にできると考える．例えば，データベースの講義においては，’データベース’という言葉は非常に特定性の低いキーワードであるが，資料中での出現頻度は低いため，従来の特定性の考慮方法では，特定性の高いキーワードであると判断される．一方，音声データを用いて特定性を考慮すれば，’データベース’は頻繁に発話される言葉であるため，特定性が低いと判断することができると考える．

また，我々はこれまでに，検索表示単位である Output Range と，この検索表示単位を検索キーワードに対応した，トピックに合うように動的に変更する手法を提案 [9] している．この検索表示単位の決定の際に，音声データへの出現回数を考慮することによって，検索表示単位をよりトピックに合った形にできると考える．

さらに複数キーワードの AND 検索の場合，音声の出現回数による適合度を考慮すれば，シーンで用いられているスライド中に，キーワードが全て出現していなくても，出現していないキーワードが音声データに出現していれば，検索候補とすることが考えられる．但し，検索候補を広げることは再現率を上げることを意味するが，精度を下げる可能性もある．それでも検索候補を広げるかどうかが，今後の大きな課題の 1 つである．

5. おわりに

本稿では，講義講演を撮影した動画のシーンを検索するシステムにおいて，講義講演の音声データを検索に利用することが有効であることを示した．まず，講義講演を撮影した動画から抽出した音声ファイルを用いて，音声認識を行い音声データの抽出を行った．音声認識により抽出した音声データは，実際にキーワードの出現回数を数え上げたものと比較を行い，シーン毎の増減傾向が同じであるため，現段階でも十分利用可能である事を確認した．また，従来の適合度との比較により検索への有効性を確認した．特にスライド中に出現するキーワードの表記揺れの吸収や，スライドで複数の話題が含まれている場合の検索精度の向上について確認した．また今後の音声データの利用可能性として，音声データを用いた特定性の考慮，検索表示範囲を決定する時の指標，複数キーワード検索における検索候補の拡大を挙げた．これらを実際にシステムに実装し，評価を行うことが今後の課題である．

また，その他の今後の課題として，音声を利用した適合度と従来の適合度の統合比率，ポインタ情報との統合が挙げられる．前者は，音声と従来の適合度の統合時における，最適なパラメータを決定するもので，実験による精度の評価を必要とする．後者は，[10] で提案した，ポインタ位置付近の文字列の情報の利用手法と音声データを連携させることにより，よりシーンの特性を正確に抽出することができると考えるものである．さらに，これらの実験を行う際には，より正確な音声認識が必要不可欠であるため，類似する講義を学習データとして加える等，言語モデルの改良を行うことも必要であると考える．

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岩野公司助手に感謝致します. 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 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.

参 考 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) 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.
12) 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.
13) G.Jones, J.Foote, K.S. Jones, and S.Young. Retrieving spoken documents by combining multiple index sources. In *Proc. of ACM SIGIR 96*, pp. 30–38, 1996.
14) A.Singhal and F.Pereria. Document expansion for speech retrieval. In *Proc. of ACM SIGIR 99*, pp. 34–41, 1999.
15) S.Srinivasan and D.Petkovic. Phonetic confusion matrix based spoken document retrieval. In *Proc. of ACM SIGIR2000*, pp. 81–87, 2000.
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) 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/.
19) 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.
20) 大語彙連続音声認識システム Julius. http://julius.sourceforge.jp/.
21) 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.
22) 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.
23) 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.
24) 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.
25) 形態素解析システム 茶筌. http://chasen.naist.jp/.
26) ipadic. http://chasen.naist.jp/.