音声ドキュメントのセマンティックコンテンツ化と音声対話による高度利用化の研究

中川聖一（代表者）
秋葉友良（代理発表）

SCOPE最終報告会　2008．6．11

研究組織

・代表者

中川聖一：豊橋技術科学大学・教授

・分担者

新田恒雄：豊橋技術科学大学・教授

増山繁：　豊橋技術科学大学・教授

秋葉友良：豊橋技術科学大学・准教授

北岡教英：豊橋技術科学大学・講師（申請時）

土屋雅稔：豊橋技術科学大学・助教

山本和英：長岡技術科学大学：准教授

小暮悟：　静岡大学：助教

西崎博光：山梨大学：助教

申請時の研究項目

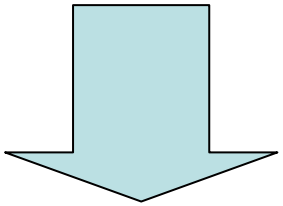
- 講義・講演・会議・ニュース等の音声・マルチメディアドキュメントの作成と利用（全員）
- 音声認識（音響モデル、言語モデル）（中川、北岡、新田）
- ドキュメントのインデキシング・メタ化（中川、西崎、新田）
- ドキュメントの要約（増山、山本、中川）
- ドキュメントの検索（中川、西崎、土屋）
- ドキュメントに対する質問応答（秋葉、中川、土屋）
- 音声対話による検索・質問応答（中川・小暮）
- ドキュメントの高度コンテンツ化（新田、中川）

主：サブテーマ

講義音声の認識・要約・コンテンツ化・ブラウジングシステム

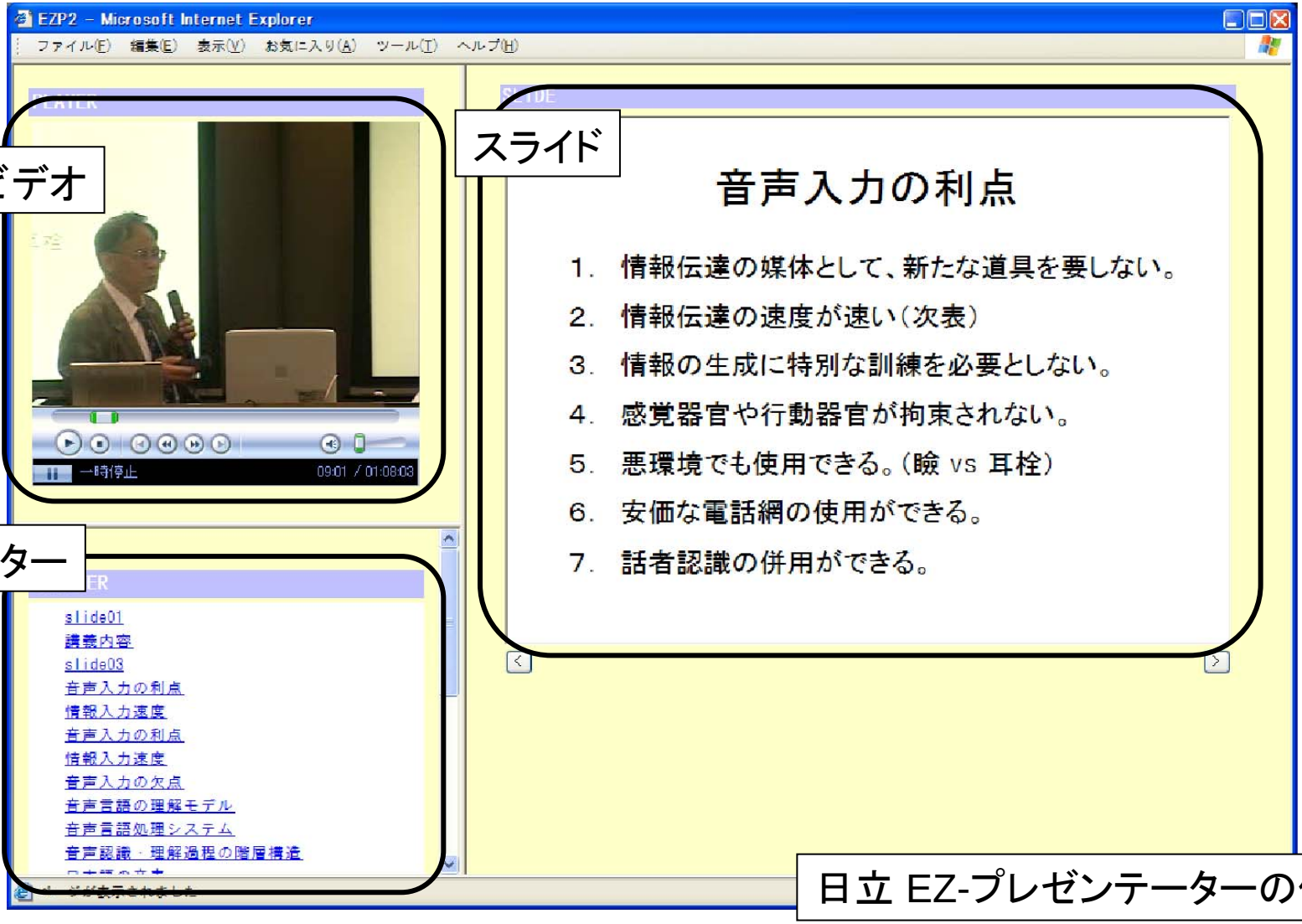
研究背景・目的

- 好きな時間に好きな場所で講義を受講することのできる*e-Learning*が、各教育機関や企業の研修などにおいて採用されてきている
- 現在の一般的な教材では、動画・および動画と同期したスライド画面といった、基本的な機能しかもたない



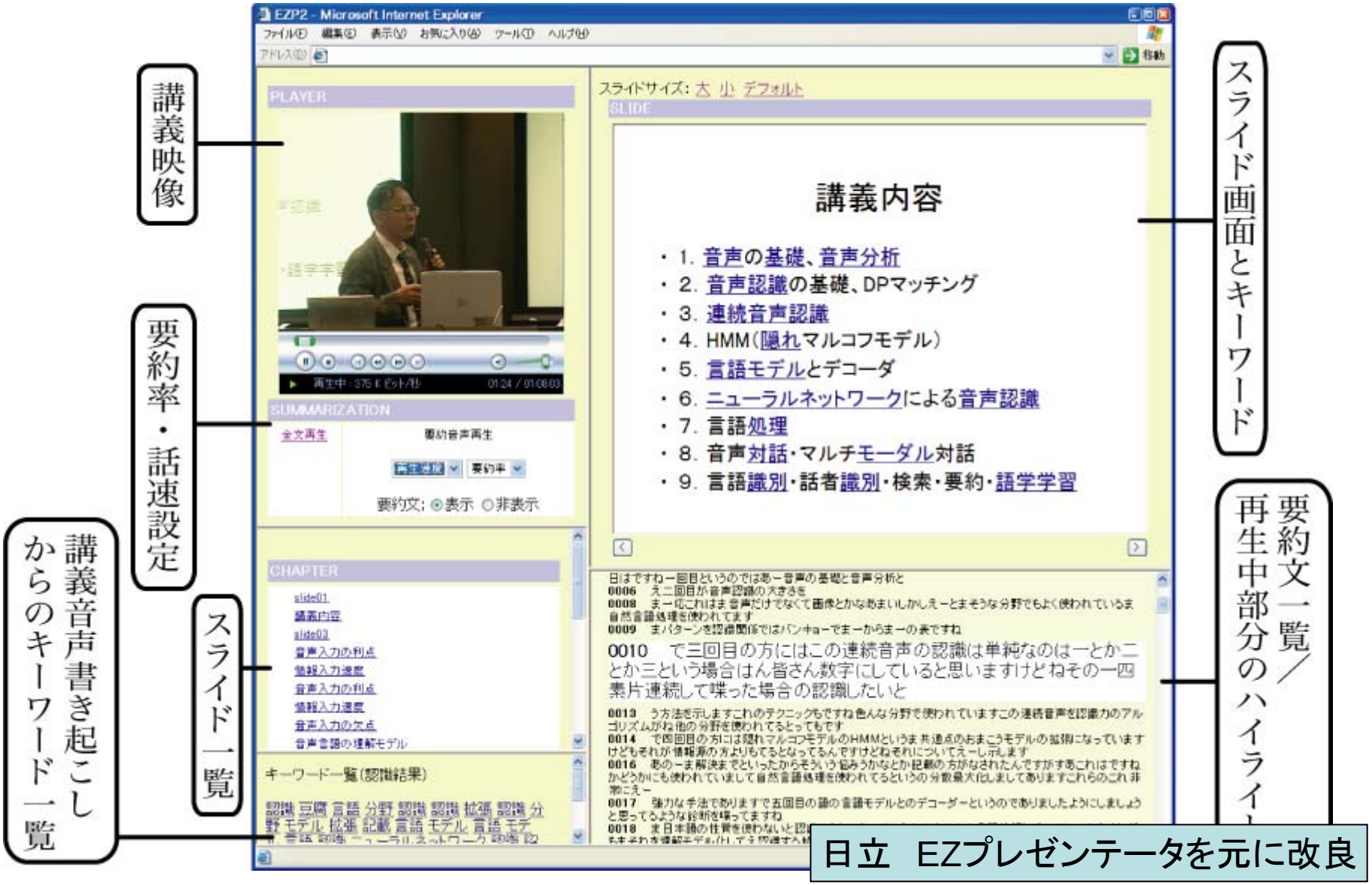
- 学習者にとって便利な付加情報を、音声情報から自動生成して提示する教材を作成

ネットワーク配信教材の例

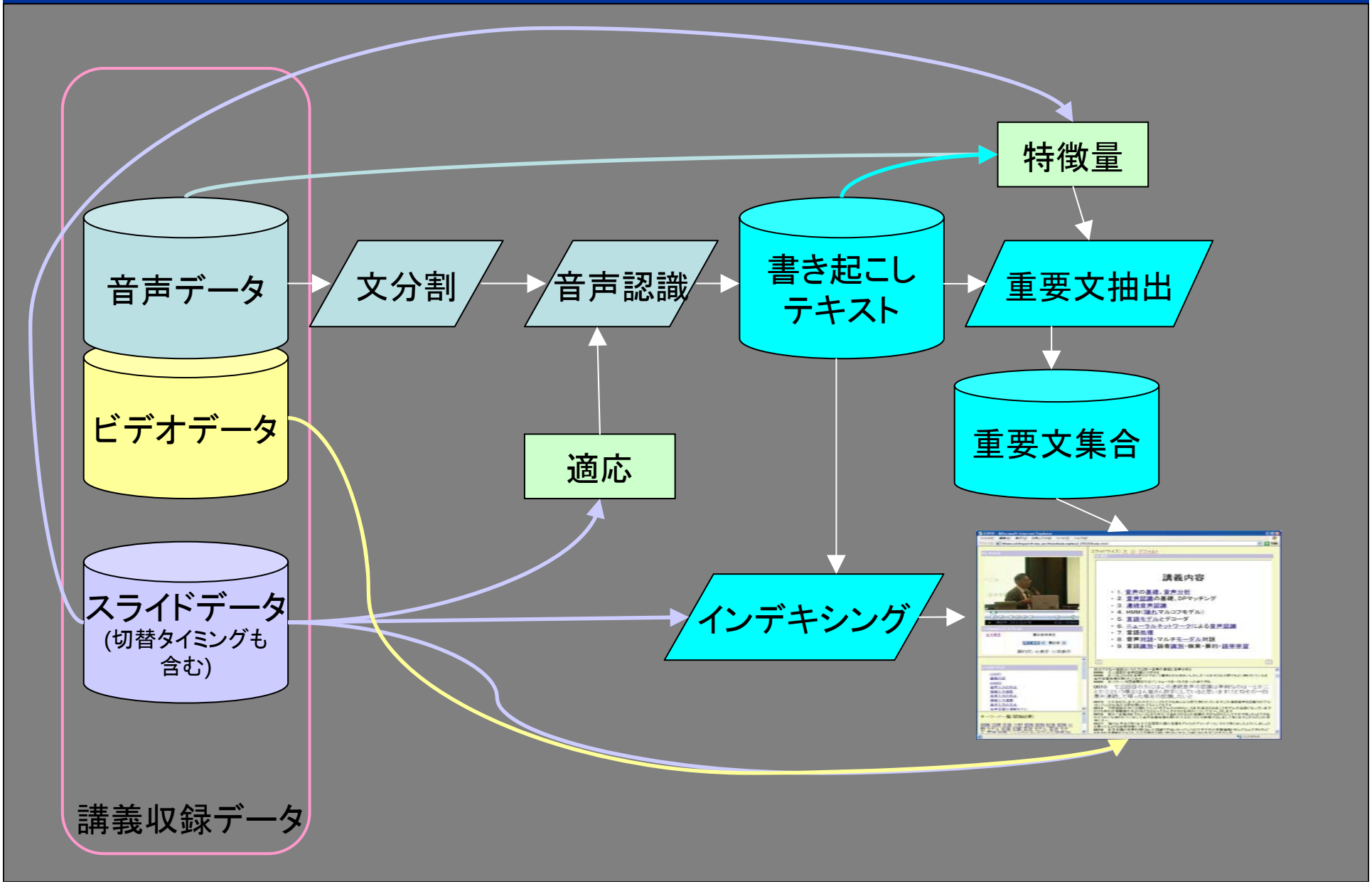


日立 EZ-プレゼンテーターの例

音声認識を高度に利用した教材の例



システムの俯瞰図


特徴量
音声データ
文分割
音声認識
書き起こし
テキスト
重要文抽出
ビデオデータ
適応
重要文集合
スライドデータ
(切替タイミングも
含む)
インデキシング
講義内容
講義収録データ

必要な要素技術

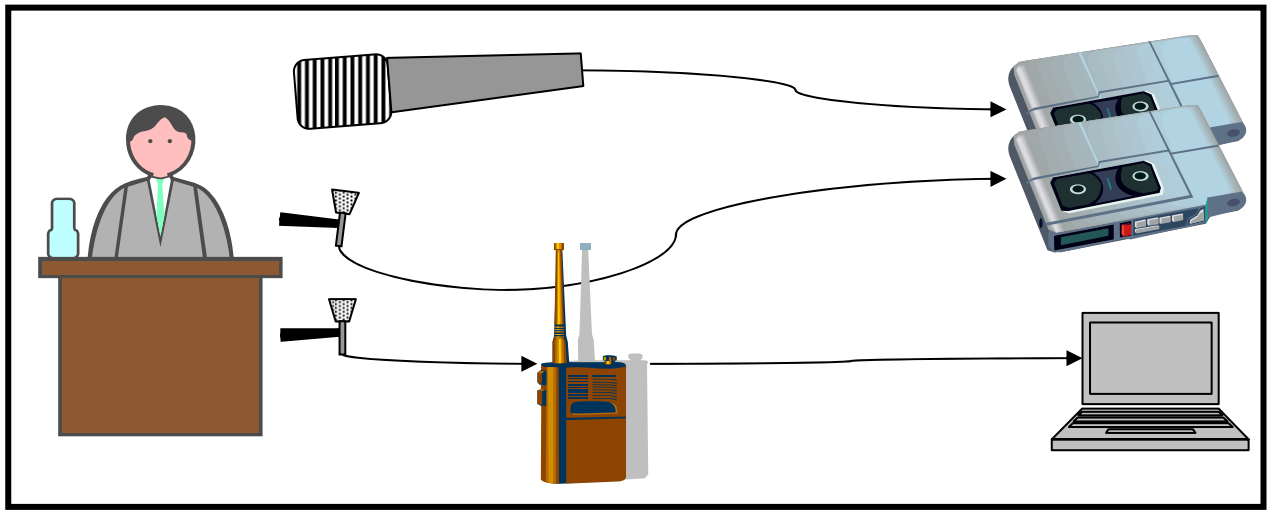
- 講義音声の自動認識
- 重要文抽出要約
- 自動インデキシング

必要な要素技術

- 講義音声の自動認識
- 重要文抽出要約
- 自動インデキシング

比較要素(1/2)

- 3種類の収録装置
  - A.指向性ハンドマイク（有線・DAT入力）
  - B.ピンマイク1（有線・DAT入力）
  - C.ピンマイク2（ワイヤレス・PC入力->WMA圧縮）



比較要素(2/2)

- 認識システム

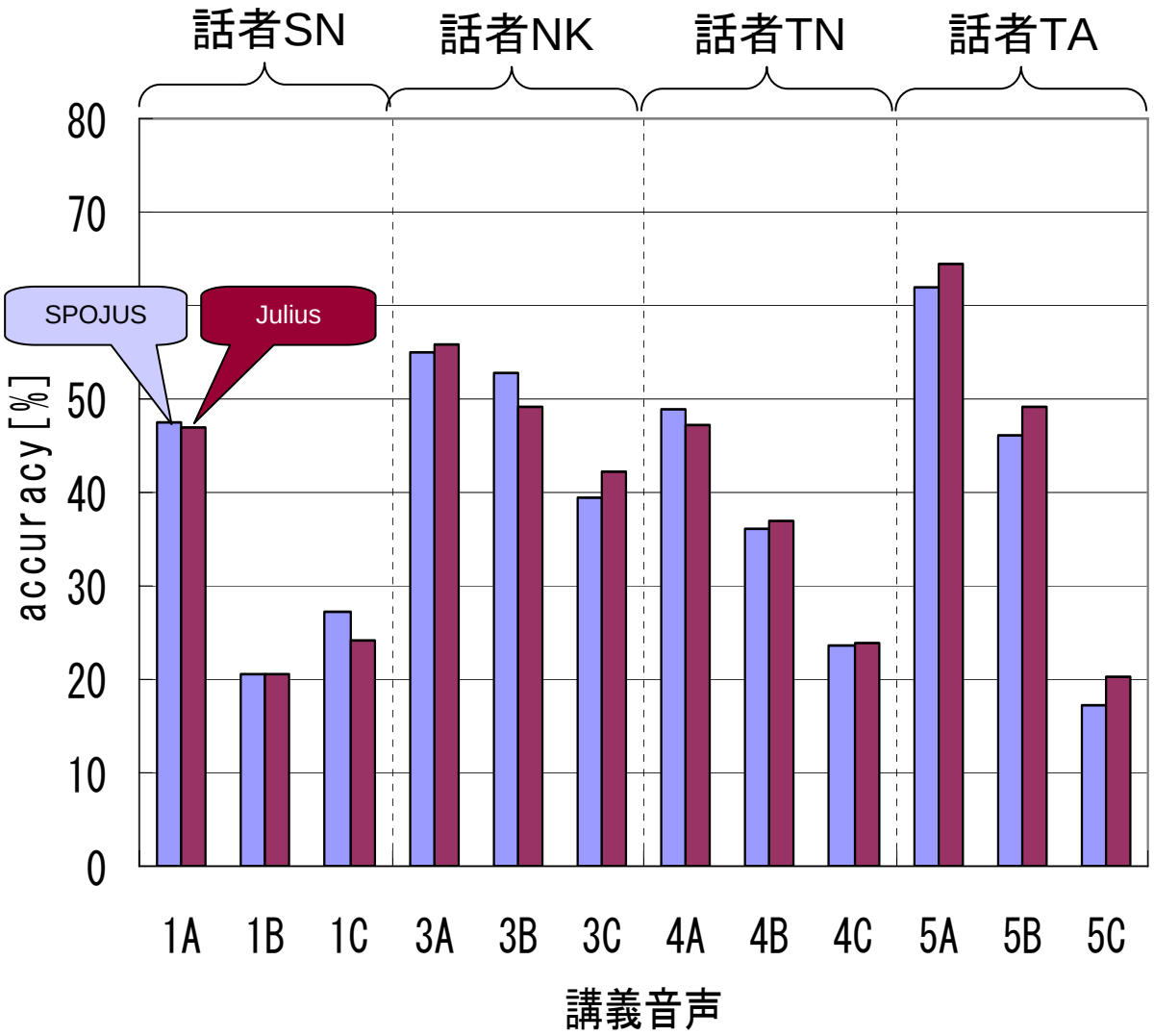
デコーダ	音響モデル	言語モデル
SPOJUS	CSJ 133音節	CSJ 17k Trigram
Julius		

特徴量： 16kHzサンプリング音声 ⇨ MFCC(12)、ΔMFCC(12)、ΔPOWER(1)、計25次元
HMM ： 音節(133)モデル ⇨ 混合数32、状態数3(母音)～5(子音)、対角共分散行列

- 認識対象講義

講義ID	話者	時間長	PP	未知語率[%]	講義内容
1	SN	1:07:56	186.4	0.37	音声情報処理の概要
3	NK	1:05:49	177.7	1.88	音声対話
4	TN	1:09:28	285.6	1.94	音声認識とパターン認識
5	TA	1:10:02	176.4	1.89	自然言語処理

認識実験結果



SPOJUS⇔Julius

SPOJUS 39.7%
Julius 40.0%

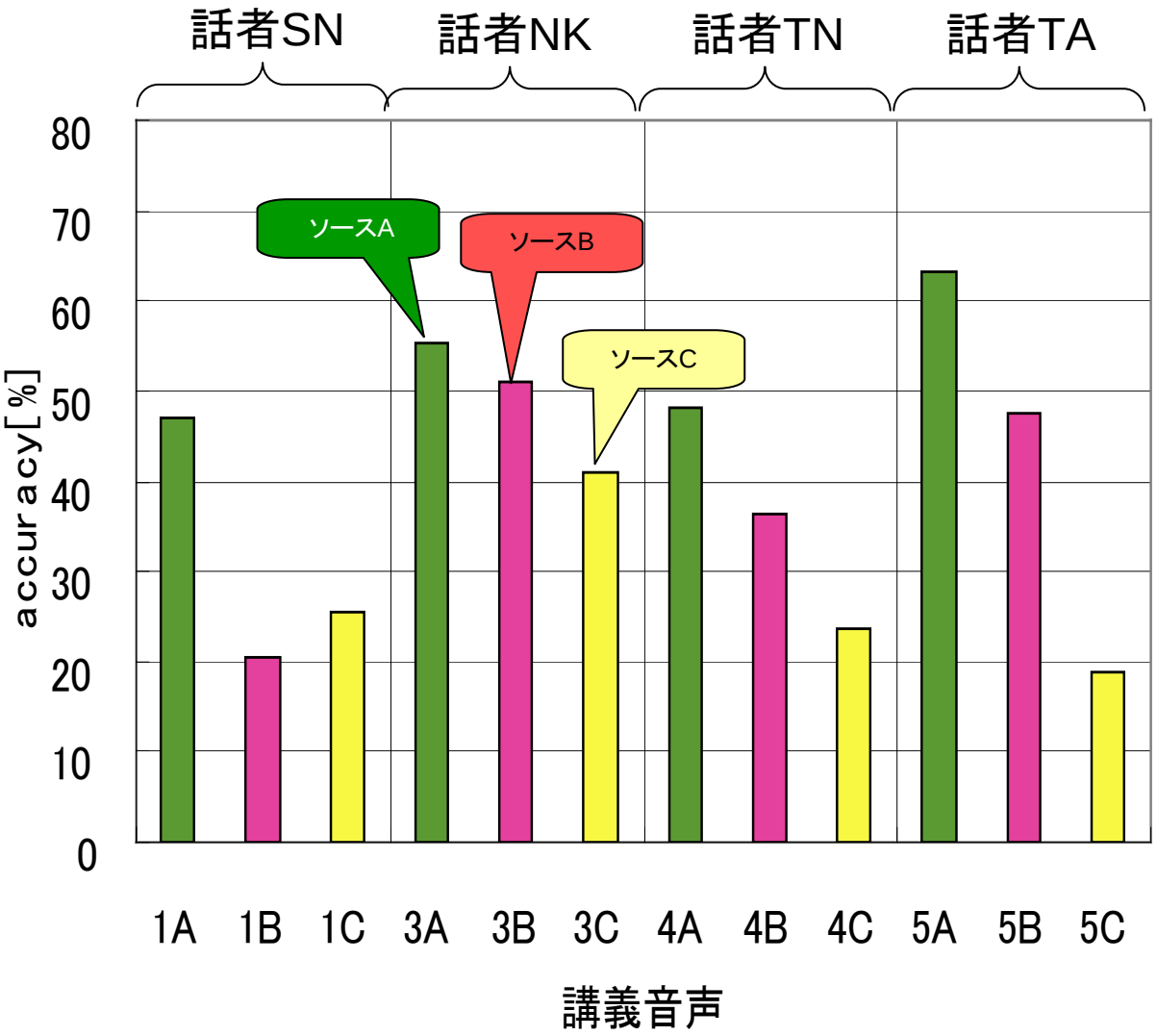
ソースA,B,C(収録装置)

ソースA(ハンドマイク)が
すべての場合において優れている

話者

話者NKがもっとも高い
話者SNが全体的に低い

認識実験結果



SPOJUS⇔Julius

SPOJUS 39.7%
Julius 40.0%

ソースA,B,C(収録装置)

ソースA(ハンドマイク)が
すべての場合において優れている

話者

話者NKがもっとも高い
話者SNが全体的に低い

必要な要素技術

- 講義音声の自動認識
- 重要文抽出要約
- 自動インデキシング

音声要約(重要文抽出)

- 講義全体をいくつかの区間(文)に区切り、その文が重要であるかどうかを自動的に判断する。

- 重要であると判断された文のみを提示する

文ごとの認識結果や、韻律情報などを特徴量として利用

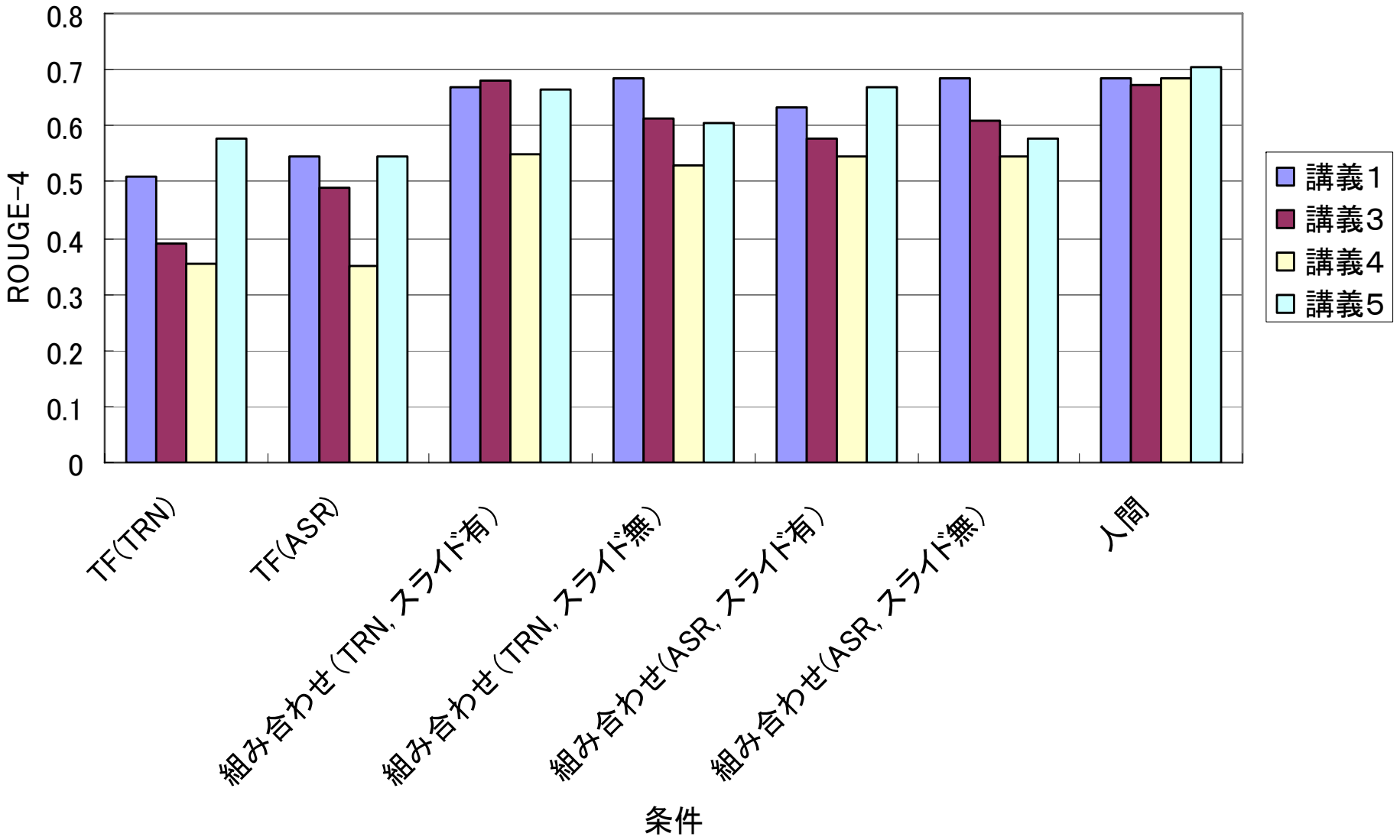
文の提示は音声+自動認識結果で行うため、
利用者は認識誤りによる悪影響を受けにくい

自動抽出手法

- 韻律情報
  - F0の平均の高い文
  - パワーの平均の高い文
  - 発話時間長の長い文
- 表層的言語情報(自動認識結果より)
  - 高頻出単語を含む文
  - スライド(PPT)情報
    - スライド中の高頻出単語を含む文
    - スライド中の見出し語を含む文
- 非重要文としての特徴
  - 各文のF0/パワーの平均の低い文、発話時間長の短い文

ベクトルモデルとして統合し、線形SVMによる判別を行う

重要文抽出結果　ROUGE-4(設定要約率25%)



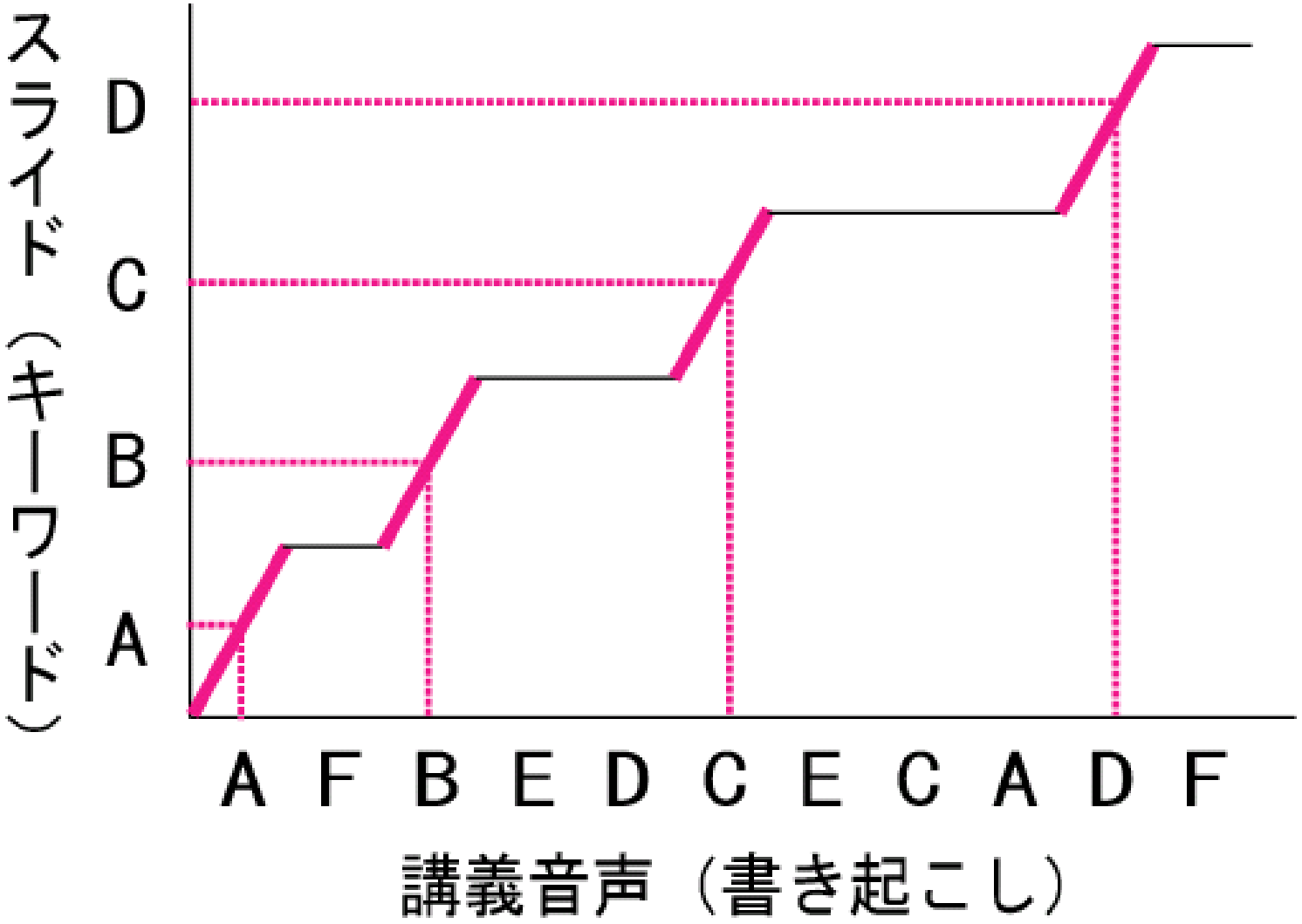
必要な要素技術

- 講義音声の高精度な自動認識
- 重要文抽出要約
- 自動インデキシング

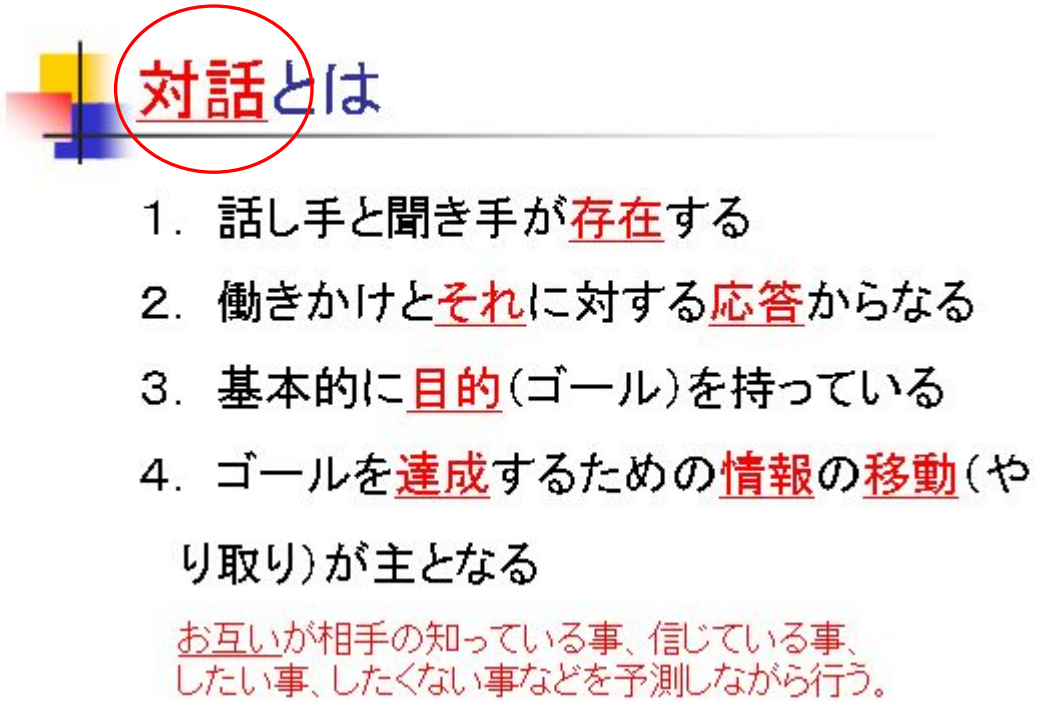
インデキシング

- スライド中のキーワードと発話中の単語をDPマッチングし，キーワード単位での対応付けをする
- スライド中のキーワードは、tf･idfスコアが平均以上の単語を選択する
- キーワードは名詞のみを対象とする
  - chasenで形態素解析を行った結果を用いる

スライドキーワードと発話文とのDPマッチング



PPT（html）への埋め込み



- これらのキーワードをクリックすると、対応付けられた発話時間へビデオがジャンプする
- 聞きたいところがより探しやすくなる

複合語の抽出

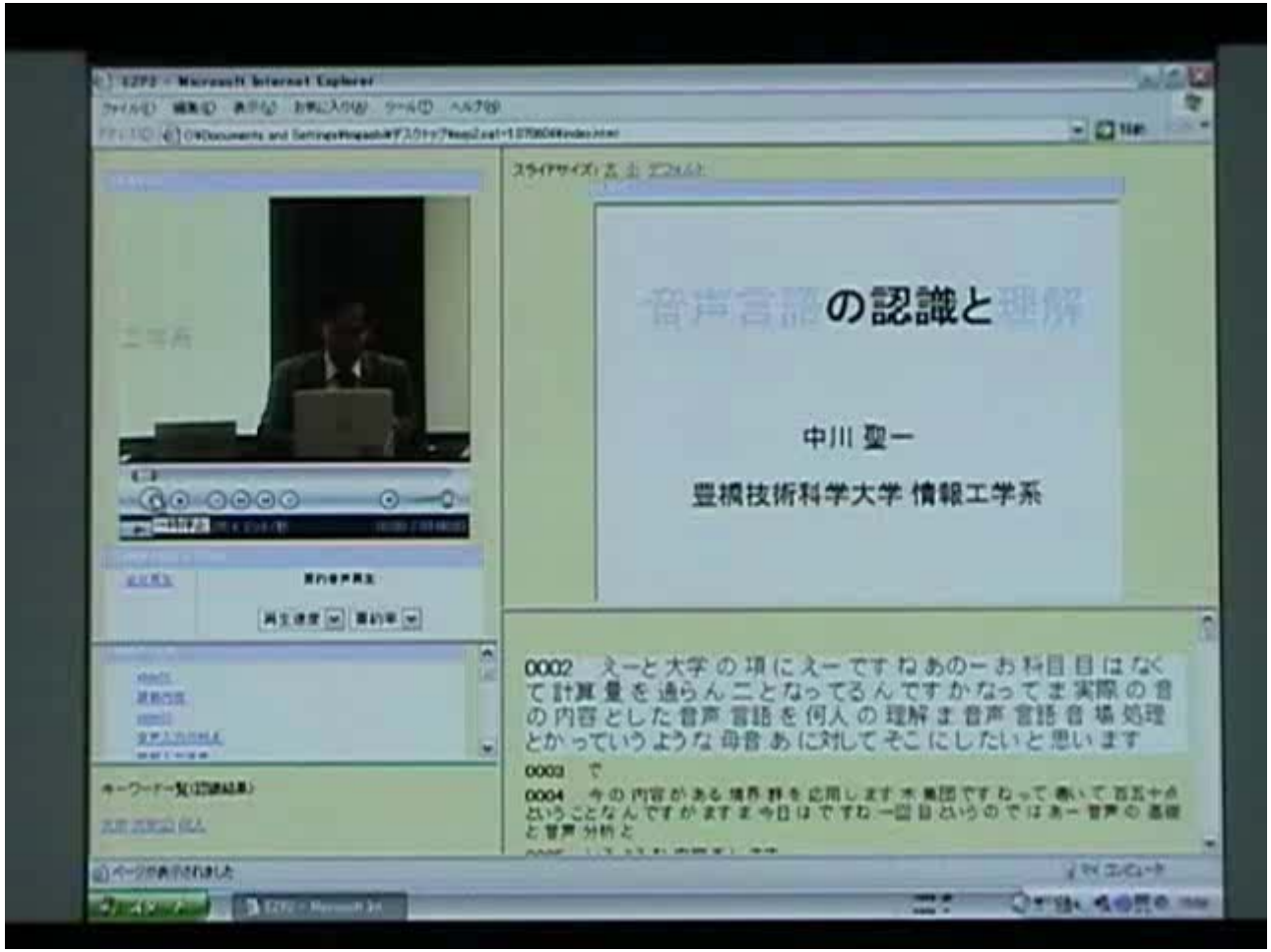
- １名詞だけではキーワードらしいとは言えない
  - 例;「音声認識」⇒「音声」、「認識」
  - 複数のキーワードとして独立に処理してしまう
- ”TermExtract”（専門用語自動抽出Perlモジュール）を使用し複合語を抽出

抽出例
（2形態素以上の
複合語を使用）

音声認識	2 morphs
音声言語	2 morphs
音声分析	2 morphs
言語音	2 morphs
音声合成器	3 morphs
音声入力	2 morphs
言語モデル	2 morphs
音声波形	2 morphs
オリジナル音声	2 morphs
機械音	2 morphs

……

音声認識を高度に利用した教材の例（デモ）

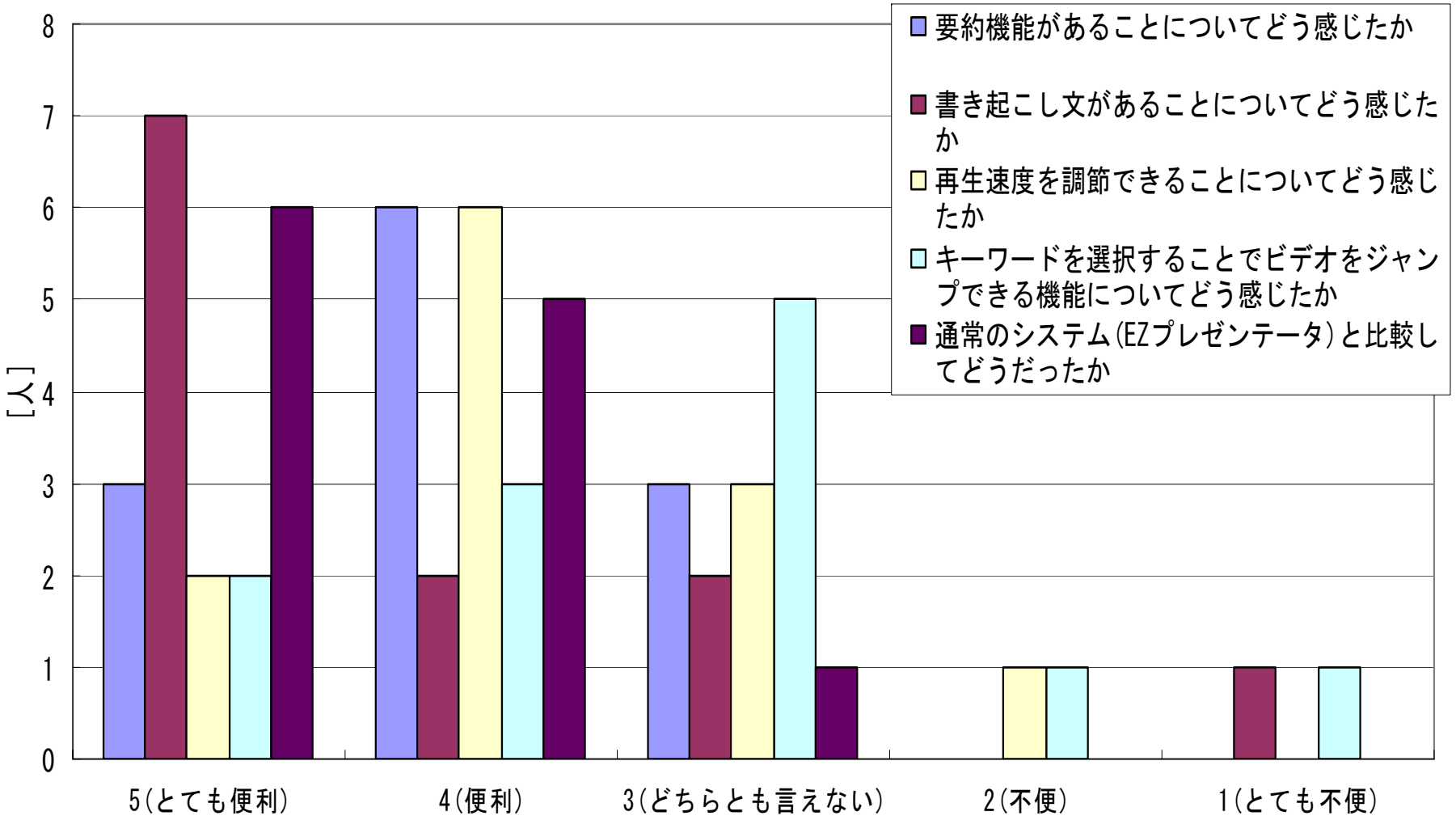


システム被験者実験

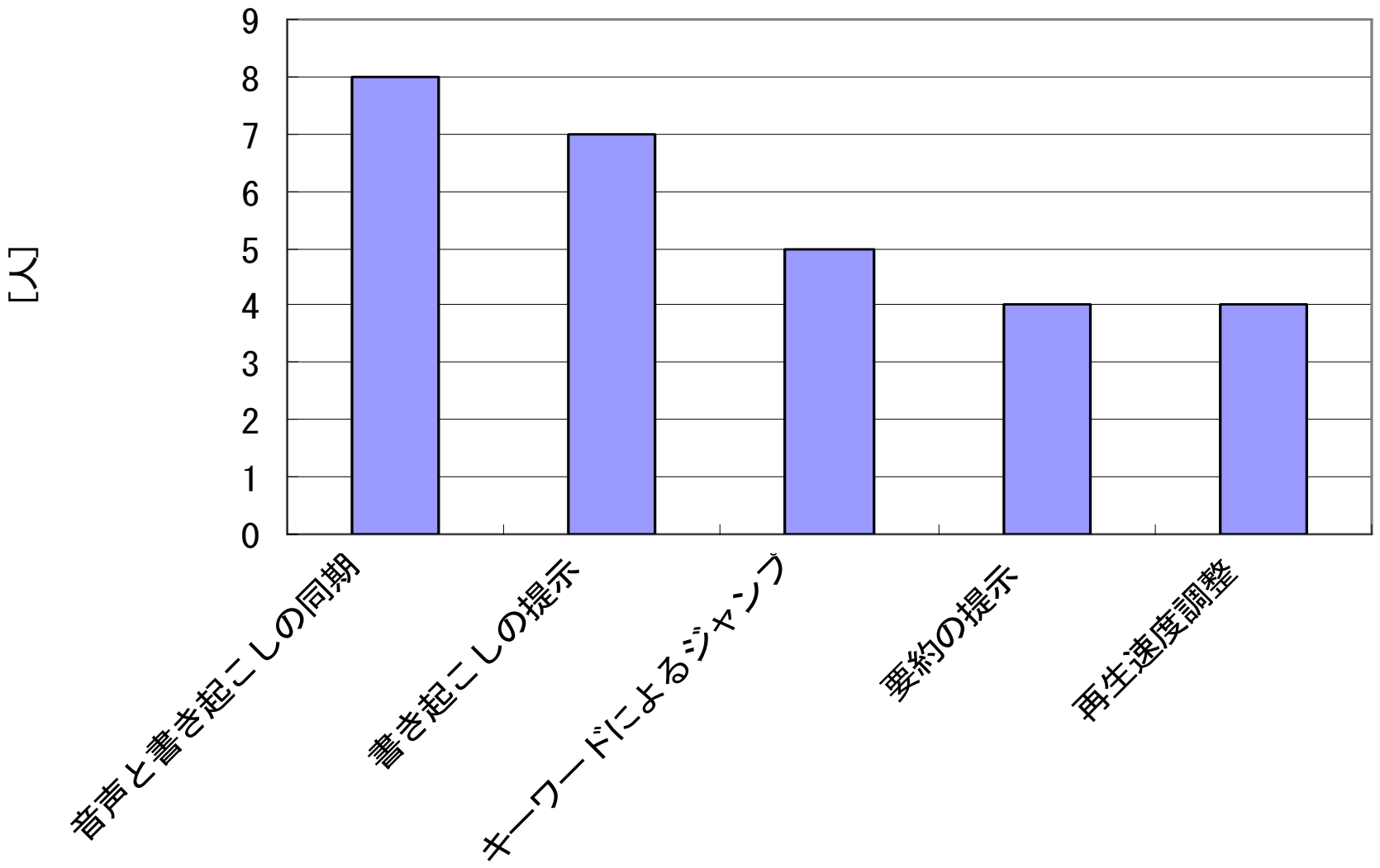
評価方法

- 便利さの評価
  - 講義1と講義3をそれぞれ2区間に分割し、4つの区間を設定 （1区間≒30分）
  - 各区間について通常機能のコンテンツと、機能拡張コンテンツを作成
  - 12人の被験者のうち、6人ごとに異なる区間を視聴してもらう。また、通常版、拡張版を視聴する順序は変える
  - 通常版と拡張版を比較して、5段階評価

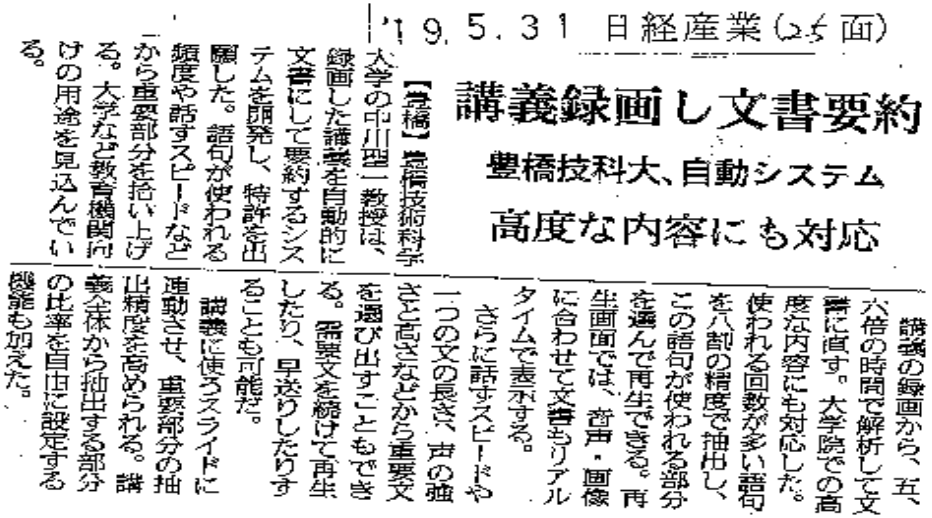
便利さのアンケート結果



どの機能が便利でしたか？



新聞報道



'19.5.31 日経産業(25面)

講義録画し文書要約

豊橋技科大、自動システム

高度な内容にも対応

【豊橋】豊橋技術科学大学の中川聖一教授は、録画した講義を自動的に文書にして要約するシステムを開発し、特許を出願した。語句が使われる頻度や話すスピードなどから重要部分を拾い上げる。大学など教育機関向けの用途を見込んでいる。

講義の録画から、五、六倍の時間で解析して文書に直す。大学院での高度な内容にも対応した。使われる回数が多い語句を八割の精度で抽出し、この語句が使われる部分を選んで再生できる。再生画面では、音声・画像に合わせて文書もリアルタイムで表示する。

さらに話すスピードや一つの文の長さ、声の強さと高さなどから重要文を選び出すこともできる。需要文を続けて再生したり、早送りしたりすることも可能だ。

講義に使うスライドに連動させ、重要部分の抽出精度を高められる。講義全体から抽出する部分の比率を自由に設定する機能も加えた。

特許　中川、北岡、富樫、山口：
プレゼンテーション解析装置およびプレゼンテーション視聴システム
特願2007－61123（H19．3．10）

ビジネスショウで展示：　2007．7．11－13（中川、富樫、野中・知財）

サブテーマ

音声対話による音声ドキュメントの検索・QAシステム

新田研・秋葉研のデモビデオ紹介

3年間のまとめ

- **講義コンテンツデータベースの公開と分析・評価**

（小暮、西崎、土屋）: http://www.slp.ics.tut.ac.jp/CJLC/

- **講義コンテンツのブラウジング（音声認識、要約、インデキシング、視聴システム）**（中川、北岡）

・音声ドキュメントの音声対話による検索・QA（新田・秋葉）

- **話し言葉テキストコーパスのコンテンツ化(要約、マイニング、文・トピックのセグメンテーション）**

（増山、酒井、山本）

----外部発表のまとめ------

誌上発表論文数	45件
口頭発表数	129件
申請特許数	3件
報道発表	1件