音声認識技術を用いた会議録及び字幕の作成支援システム

河原 達也

（京都大学）

http://www.ar.media.kyoto-u.ac.jp/jimaku/

音声認識技術

- ディクテーションソフトウエア
  - 文法的な文の丁寧な読上げ音声
- カーナビ・電話応答(IVR)システム
  - 限定されたタスク・ドメイン

機械を意識した発声

- 人間どうしの自然な話し言葉音声…依然として困難
  - 会話・ミーティング

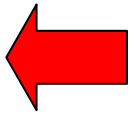
- 公の場で比較的明瞭に話される音声 [本研究]
  - 講演・講義
  - 議会

国会審議の音声認識システム

- 衆議院の次期会議録作成システム
  - 現在の速記制度に変わり、音声認識技術の導入
  - 速記者による認識結果の修正・整文作業

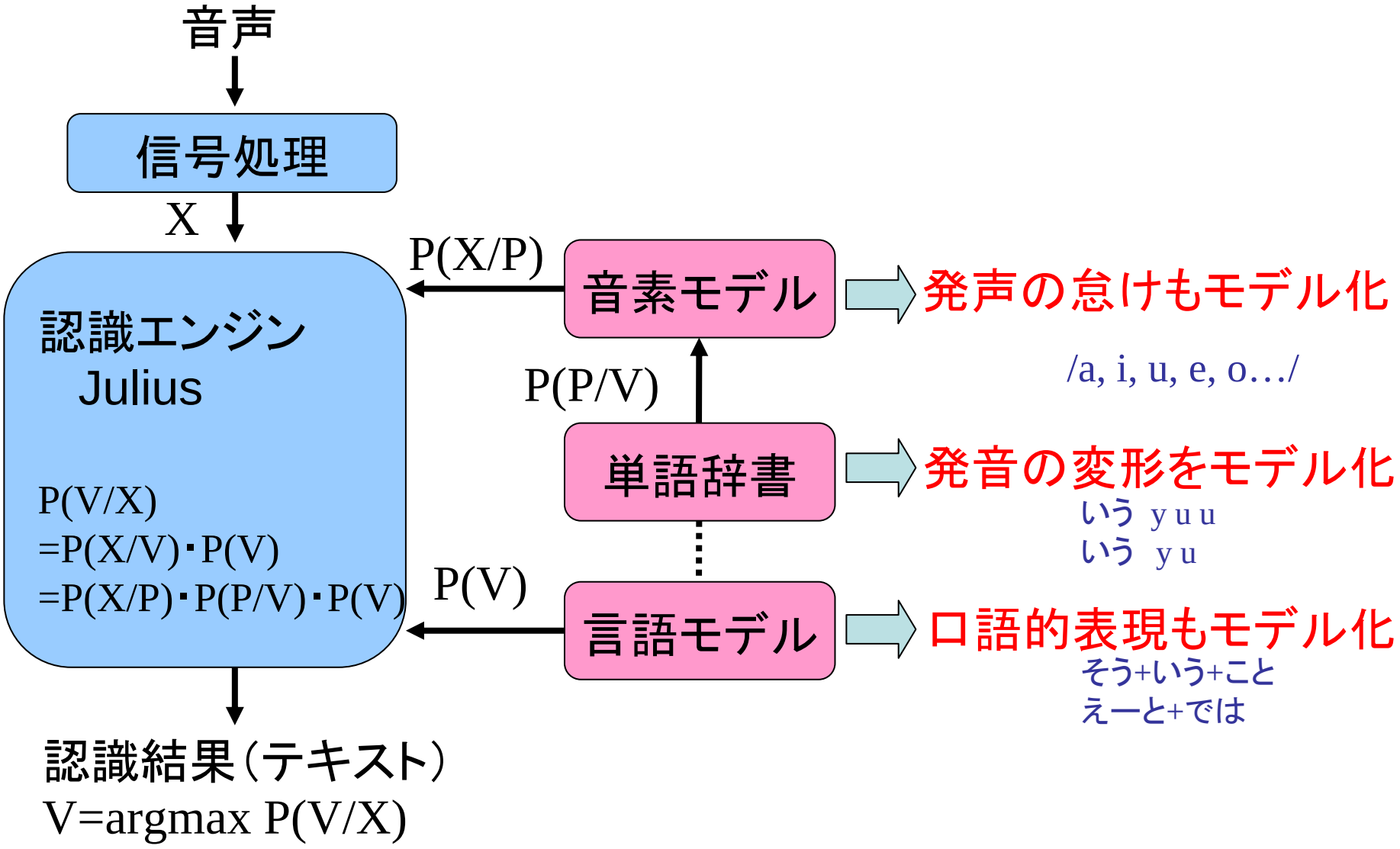
- 目標
  - 音声認識の利用が明らかに効率的と感じられること

- 認識性能
  - 文字正解精度 85%
  - 実時間の3倍以内
  - 時間遅れ10分以内



従来のシステム
- 自発性の高い音声（委員会審議）に対応できない
- 本研究開始前は80%弱

音声認識の話し言葉への対応

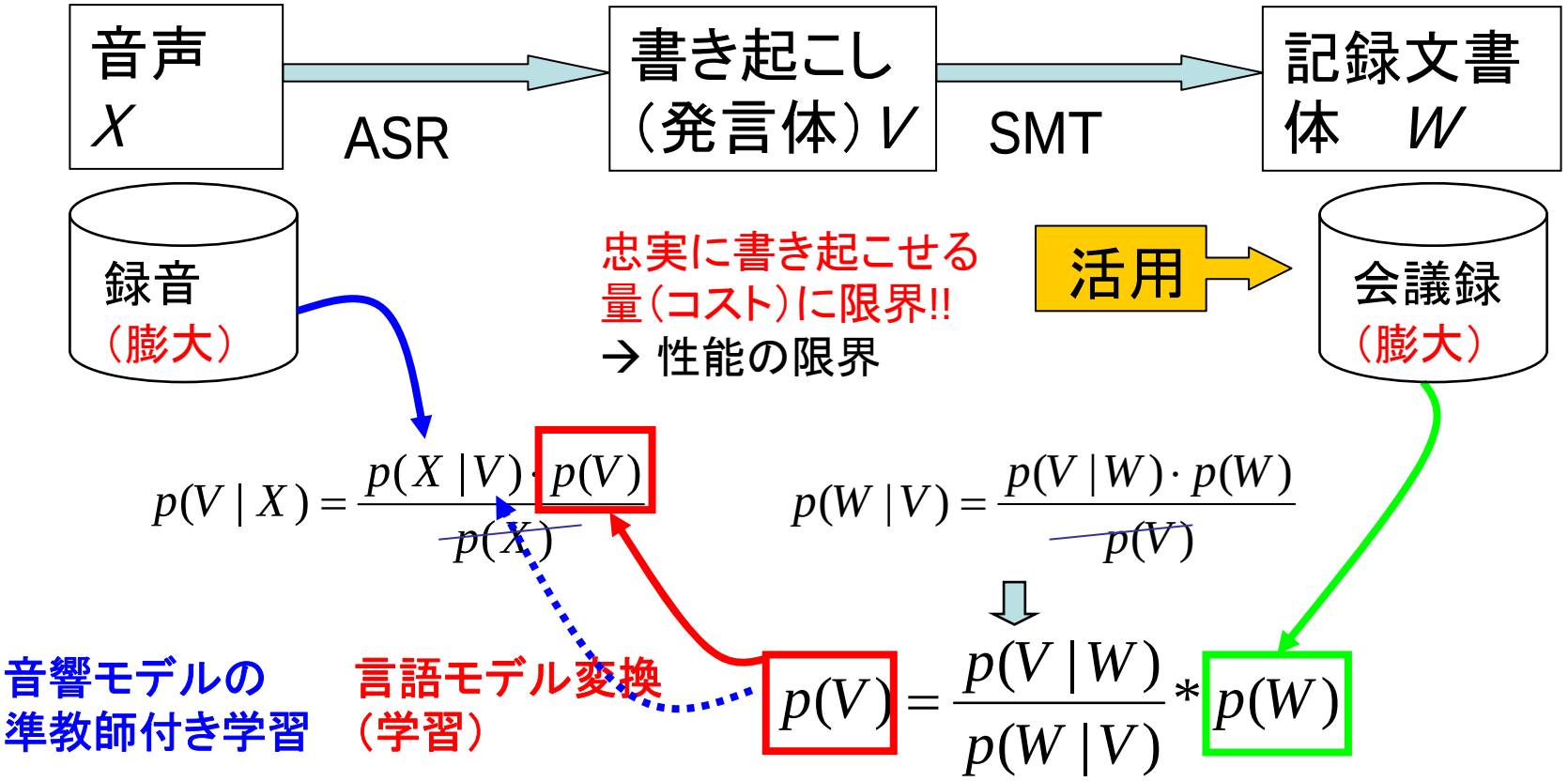


衆議院審議コーパス

- 本会議・主要委員会をすべてカバー
- 200時間+, 400万単語+　　→ 十分な量といえない
- 審議音声を忠実に書き起こし、会議録と対応付け

{えー}それでは少し、今{そのー}最初に大臣からも、{そのー}
貯蓄から投資へという流れの中に{ま}資するんじゃないだろう
かとかいうような話もありましたけれども、{だけど/だけれども}、
{まあ}あなたが言うと本当にうそらしくなる{んで/ので}{ですね、
えー}もう少し{ですね、あのー}これは{あー}財務大臣に{えー}
お尋ねをしたいんです{が}。
{ま}その{あの}見通しはどうかということでありますけれども、
これについては、{あのー}委員御承知の{その}「改革と展望」
の中で{ですね}、我々の今{あのー}予測可能な範囲で{えー}
見通せるものについてはかなりはっきりと書かせていただい
て(い)るつもりでございます。

話し言葉←→文書体の言語モデル変換



→　音声データ**X**と会議録テキスト**W**のみでモデル学習・更新が可能
　　（書き起こしを必要としない）

→　学習データ量の限界の打破！

衆議院審議の音声認識精度

- 2009年の総選挙・政権交代に伴い、10月召集国会のデータ（約100時間）で自動学習

	モデル更新前 (0910)	モデル更新後 (1001)
2010年 2月 8日　予算	85.4%	86.6%
2010年 2月24日　法務	86.1%	87.1%
2010年 2月24日　文部科学	85.4%	86.2%
上記の平均	85.7% →	86.6%

- 更新前のベースラインでも 85%（目標値）を実現
- モデル更新により、さらに1%の改善
  → 実運用可能なレベル

音声認識結果の後処理

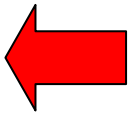
- 話し言葉の整形・区分化
  - 速記者・校閲者の整形過程を分析・モデル化
  - 忠実な発話の書き起こし*V*から会議録文書体*W*への統計的機械翻訳
  - 節・文単位への区分化（句点の挿入）、読点の挿入
- 音声アーカイブへのメタデータ付与
  - 話者区分化
  - 音響イベント（拍手・フィラー・笑い声・相づち等）の検出・アノテーション

大学講義の音声認識システム

- 聴覚障害者・外国人のための字幕付与
  - 手書きノートテイクに代わる
  - PC要約筆記を補完

- 目標
  - リアルタイムに編集可能な範囲で、ほぼすべての情報が伝えられる字幕が作成できること

- 認識性能の目標
  - キーワード検出率 80%
  - 実時間で認識



従来のシステム
•講師の音声を直接認識できるレベルにない
•復唱方式

音声認識システムの適応

- 講義スライド、及び当該講師の以前の講義データを用いて適応

	適応前 （ベースライン）	適応後
音響モデル	講演データベース	講師の声の特徴を反映
単語辞書	講演データベース	専門用語を反映
言語モデル	講演データベース	専門用語を反映 講師の話し方を反映
認識率	55-70%	60-75%
キーワードの認識率	65-80%	**80-90%**

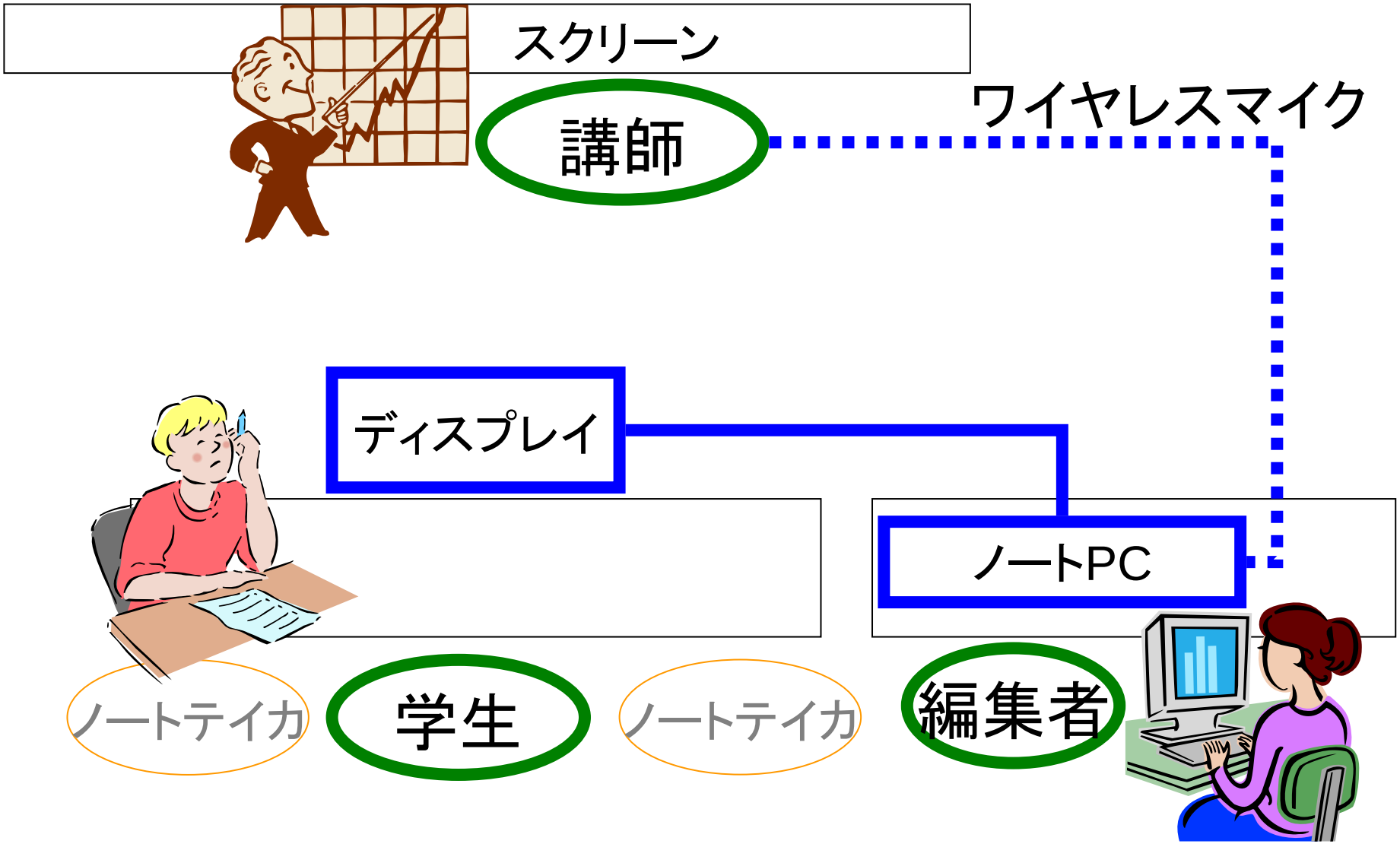
音声認識による字幕付与システム

- 修正インタフェース
  - IPtalkの確認修正パレットを利用
  - 長時間でも1人で作業可能

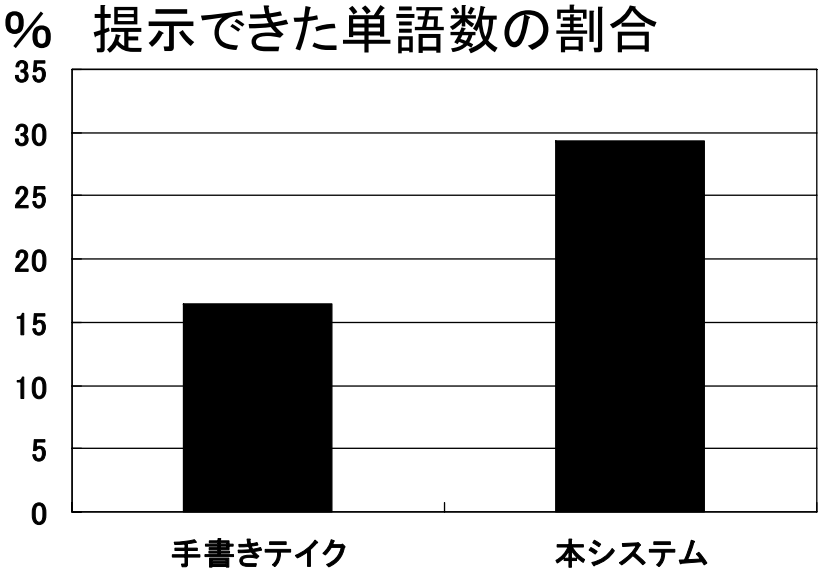
- 京都大学における講義で試験評価
  - 聴覚障害学生（工学部3回生）
  - 現行の手書きノートテイクより有意に多い情報量の提示
- 本プロジェクトのシンポジウムの講演で実演
  - ほぼすべての発話に字幕付与

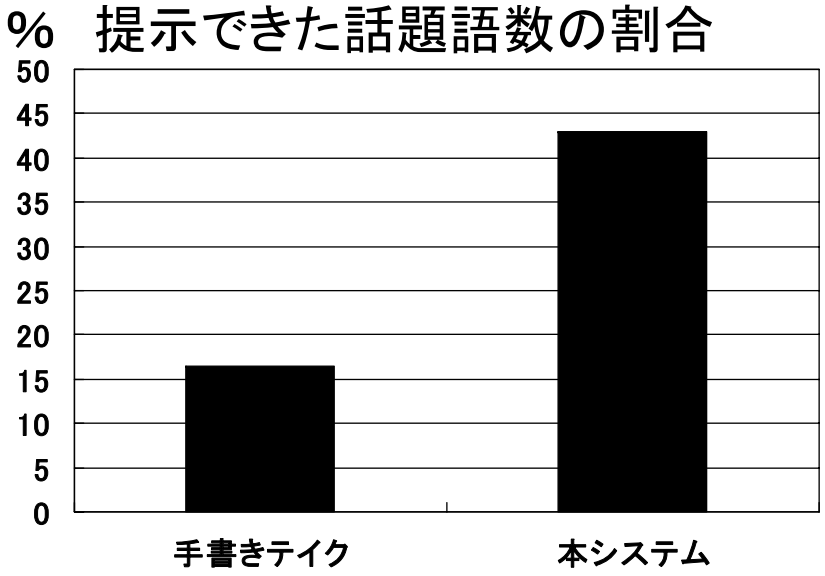
- NHK教育テレビ「ろうを生きる難聴を生きる」で放映（3年連続）

講義におけるシステムの試験運用



提示テキスト量に関する評価





- 手書きテイク(2人分)の1.8倍の分量(1825/3260単語)
- 話題語に関しては2.6倍(223/582単語)

有意に多くの情報量を提供

主要成果発表

- Y.Akita and T.Kawahara.
  Statistical transformation of language and pronunciation models for spontaneous speech recognition.
  IEEE Trans. Audio, Speech & Language Processing (採録決定).
- 秋田祐哉, 三村正人, 河原達也.
  会議録作成支援のための国会審議の音声認識システム.
  電子情報通信学会論文誌, (採録決定), 2010.
- 河原達也, 根本雄介, 勝丸徳浩, 秋田祐哉. スライド情報を用いた言語モデル適応による講義音声認識. 情報処理学会論文誌, Vol.50, No.2, pp.469--476, 2009.
- T.Kawahara, M.Mimura, and Y.Akita. Language model transformation applied to lightly supervised training of acoustic model for congress meetings. In Proc. IEEE-ICASSP, pp.3853--3856, 2009.
- NHK教育テレビ「ろうを生きる難聴を生きる」で3年連続取材・放映
- 特許出願
- プロジェクトWebサイト http://www.ar.media.kyoto-u.ac.jp/jimaku/