議会の会議録作成のための音声認識
-- 衆議院のシステムの概要 --

河原達也
（京都大学）
http://www.ar.media.kyoto-u.ac.jp/diet/

歴史的経緯

- 1772年　英国議会で手書き速記による発言録作成
- 1880年　イタリア議会上院で機械速記の採用
- 1890年（明治23年）
  - 日本において帝国議会の設立
  - 手書き速記による会議録作成
- 21世紀
  - 衆参両院で速記者の新規採用・養成の停止(2005年)
  - 代替手段の模索・調査
  - 衆議院において
    自動音声認識システムの導入(2010年)・運用(2011年)
    （審議音声を直接認識する方式は世界初）

システムの位置づけ

- すべての本会議・委員会の審議を対象
- 発言者のマイクから音声収録
  - （質問者）:（答弁者＋議長）の2チャネル
- 音声認識結果を元に反訳原稿作成
  - 一定の音声認識誤りは不可避
  - 話し言葉を忠実に書き起こしても会議録にならない
  - 速記者・校閲者の役割がなくなるわけではない

議会・裁判所における音声認識

- 米国の裁判所
  一部の速記者がディクテーションソフトに
  ボイスライティング（復唱入力）
- 日本の裁判所
  公判の検索（＋記録作成の支援）に音声認識 (by NEC)

- 諸外国の議会・参議院
  録音したものをタイプ入力で書き起こす（テープ起こし）
- イタリア議会
  録音したものを復唱してディクテーションソフトに入力
- 日本の地方議会
  北海道議会、東京都議会などで音声認識が導入
  本会議：80-90%, 委員会：70%

日本語固有の問題

- かな漢字変換の曖昧性
  (ex.) KAWAHARA → 河原 (not 川原)
  → リアルタイムにタイプ入力することがほぼ不可能
  (cf.) スピードワープロ
- 表記の揺れ
  (ex.) ×明日（あす）　×行なう
- 話し言葉と書き言葉の差異が大きい
  (ex.) じゃ、これいいですか → では、これはいいですか
  → 復唱入力が困難（高速タイプ入力も）

特に、国会の会議録では高い一貫性（品質）が要求

システムの基本的要件

- 高い音声認識精度
  - 80%以下では使いモノにならない；90%以上が望ましい
    ~~×「認識精度が少々低くても、一から入力するよりまし」~~
  - 文字認識精度 85%...大半の区間で80%を確保
- 速いターンアラウンド
  - 当該作業単位（5分）は10分後には編集開始
    ~~×「一晩かけて認識処理すればよい」~~
  - 実時間比(RTF) 1以下
- 厳格な表記・文字遣いの保証
  -「用字例」　×明日（あす）　×行なう
  - 新聞記事・Web・CSJ等の他のコーパスは利用不可

類似タスクとの比較

- TC-STARプロジェクト (2004-2007)
  - 欧州議会の本会議
  - 大半が原稿の朗読、流暢
    - フィラーの割合 2.0%
- 日本の国会（本システム）
  - 大半の審議が委員会
  - 原稿の読上げでない、丁々発止の議論、自発性高い
    - フィラーの割合 4.7%
- 「日本語話し言葉コーパス」(CSJ)
  - 原稿の読上げでない、自発性高い
    - フィラーの割合 5.5%（模擬講演）, 6.8%（学会講演）
  - 独話、ヘッドセットマイク、話題に偏り

基本的アプローチ：言語モデル変換

「大規模コーパス」の効率的・持続的な推定~~（×作成）~~

- 審議は毎日のように行われる
- 音声と会議録は大規模に集積（年間千時間規模）

  ⇩ But
- 会議録は実際の発言内容とかなりの差異
  - そのままでは音響・言語モデル学習に利用できない

  ⇩ Thus
- 忠実な書き起こし(ex. CSJ)が必要
  - 膨大なコストと時間
  - 持続的に用意できない
    - 音響環境や話者集合、語彙や話題の変化に対応できない

発言体と文書体の相違の分析

- 衆議院審議コーパス
  - 審議音声を忠実に書き起こし → 性能の確保のために重要
  - 会議録と対応付け
  - 225時間/270万単語（2008年度当時） → 十分な量といえない

{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。

全体の13%の形態素で編集・相違

発言体と文書体の相違の分析

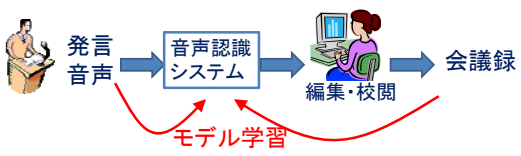
- 全体の13%の形態素で編集・相違
  - フィラー・言い淀み (ex.)「えー」「あの」
  - 冗長な文末表現 (ex.)「～ですが」「～ですね」
  - 話し言葉と書き言葉の差異 (ex.)「じゃ」→「では」
  - 文法的訂正 (ex.)「～してる」→「している」
- そのうちの93%は単純な編集（1-2単語の削除・置換・挿入）
- ただし、文脈依存のものが多い

  (ex.) これ は です ね 非常 に 大きな 問題 なん です ね 。

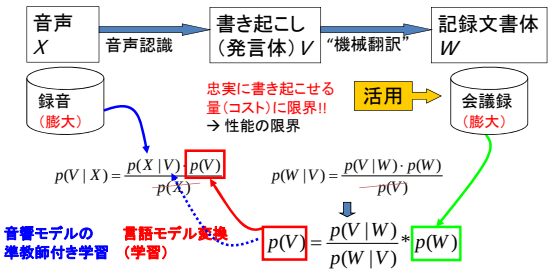
基本的アプローチ：言語モデル変換

「大規模コーパス」の効率的・持続的な推定~~（×作成）~~

- 会議録のテキストから発言内容を確率的に予測
  - テキスト自体を変換するのではない（事実上不可能）
  - 言語モデルの統計量を変換
- [言語モデル学習]：大規模な会議録のデータを変換
- [音響モデル学習]：個々の発言（ターン）毎に変換してできるモデルでラベル作成（準教師付き学習）



発言体と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
（学習データ量の限界の打破；持続的なシステム）

本アプローチの利点

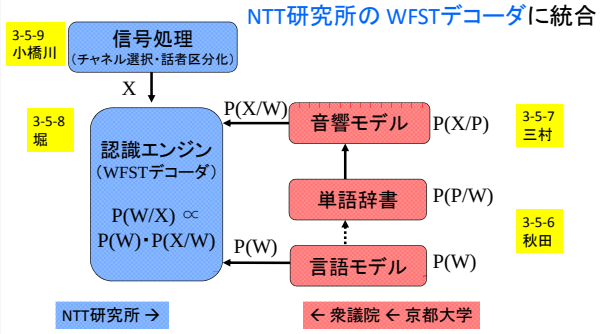
- 学習データをいくらでも増やせる
  - 音声と会議録のみあれば、忠実な書き起こしは不要
  - 学習データ量(→モデル性能)の限界の打破
- 言語・音響モデルの更新が容易
  - 総選挙・内閣改造に伴う議員・閣僚の交代
  - 話題・語彙の変化
  - 音響設備の変化
  - 持続的に進化するシステム
- 「用字例」に即した表記・文字遣いの保証
  - 会議録のみから単語辞書・言語モデル構築

本アプローチの利点
(コーパス混合の手法と比較して)

- 言語モデルが関連文書(会議録)のみで構成
- 言語モデルが統計モデルとして健全(完全)
  - すべての可能な単語の間にフィラーを予測可能
- コンパクトかつ強い制約
  - 音響モデルのラベルの推定が高い
- 単語辞書の品質が保証
  - 無関係の語彙が含まれない
  - 表記の揺れがない
- 混合重みなどのパラメータ推定が不要
  - モデルの更新作業が専門家でなくても容易

システムの基本構成

京都大学の(技術による)モデルを
NTT研究所の WFSTデコーダに統合



システムの基本構成

- 信号処理(NTT:小橋川ら)
  - チャネル選択 ← 室内拡声による回り込みが多い
  - 話者ターンへの区分化 → CMN, CVN, VTLN
- 音響モデル(京大:三村)
  - MPE学習
  - 225時間 → 約1000時間(忠実な書き起こしなし)
- 言語モデル(京大:秋田)
  - 語彙サイズ 64K
  - 1999年以降の会議録を変換; 約2億形態素
- 認識エンジン(NTT:堀ら)
  - 高速on-the-fly合成を用いるWFSTデコーダ
  - 処理速度 RTF 0.5

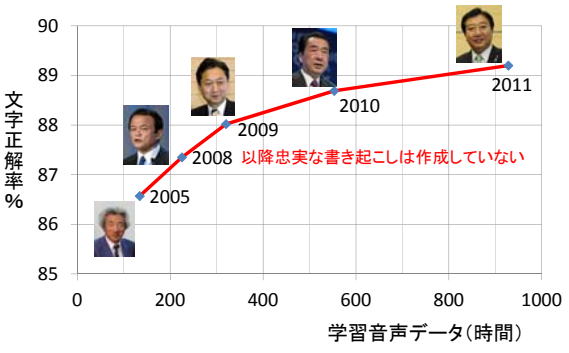
音声認識率の評価

- ほぼ全会議で会議録と照合し、文字正解率を推定
  - 実際の発言と異なるので、音声認識精度ではない
  - 最終目標に近い／書き起こしのコスト不要
  - 文字正解率(Cor.)は音声認識精度との差は1%未満
- 2010年度(試験評価)
  - 60会議で推定文字正解率(Cor.) 89.3%
  - 本会議に限ると95%
  - 5委員会で文字正解率(Cor.) 87.9%, 文字正解精度(Acc.) 85.7%
  - モデルの半自動更新により0.7%改善
- 2011年度(正式運用)
  - 118会議で推定文字正解率(Cor.) 89.8%
  - モデルの半自動更新により0.5%改善
- 2012年度
  - 推定文字正解率(Cor.) ほぼすべての会議で88%以上, 平均90%

政権交代

東日本大震災

音声認識率の改善
(評価セット:2011年の12会議)



システムのユーザビリティ

- 音声認識誤りの修正
  - 文字単位で10%程度
- 話し言葉の整形
  - フィラーの削除は自動化
  - 他の冗長語の削除・口語表現の修正が8%程度
    - 「ですね」などの冗長語の削除、「て(い)る」の挿入
- 会議録作成者(=速記者)の役割・負担は依然大きい
- 編集用の専用エディタが重要
  - ラインエディタでなく、ワープロ型のスクリーンエディタ
  - 音声に時刻・発言・文字単位でアクセス可 ←アライメント
  - 音声再生の速度を調節可能 ←話速変換

会議録作成の流れ

- 入力の作業単位への機械的な分割
  - 5分+前後の重複(各1分)
  - 分割後、約3分以内に認識結果が生成
  - 会議室では臨場者が会議の状況を把握
- 編集・校閲
  - 会議録作成者(=速記者)が作業室で修正・編集・確認(～1時間)
  - 校閲者が会議全体の会議録の確認(～1日)
  - 将来は速記者でない人が従事 → 養成が鍵

モデルの更新

- 音声認識率の継続的なモニタリング
- 言語モデルの更新(年1回めど)
  - 単語登録は、ワープロソフトと同様、いつでも可能
- 音響モデルの更新
  - 内閣改造・総選挙 → ここ毎年
- 用字例の改訂(常用漢字表の改定)への対応
  - 過去の会議録(学習データ)も修正
  - 数十年に1回?

まとめと今後の課題

- 最高水準の話し言葉音声認識システム
  - 言語モデル変換による準教師付き学習
  - 高速on-the-fly合成を用いたWFSTデコーダ
  - 持続的に、ほぼ自動で性能モニタ＆モデル更新
- 今後の課題
  - 会議録との文字正解精度(Acc.)は80%強
  - 持続的な認識性能の改善
  - 音声認識結果(発言体)から会議録(文書体)への自動変換
    - 現在はフィラーの削除のみ

謝辞

- 京都大学
  秋田祐哉, 三村正人
- NTT研究所
  政瀧浩和, 高橋敏, 小橋川哲, 堀貴明
  浅見太一, 山口義和, 阪内澄宇, 小川厚徳

- 参考資料
  http://www.ar.media.kyoto-u.ac.jp/diet/