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（071707004）

Automatic Speech Transcription System for Generating Meeting Records and Captions

研究代表者

河原達也　京都大学

Tatsuya Kawahara　Kyoto University

研究分担者

秋田祐哉[†]　高梨克也[†]　坂井信輔[††]　清水徹[††]　山田篤[†††]

Yuya Akita[†]　Katsuya Takanashi[†]　Shinsuke Sakai[††]　Tohru Shimizu[††]　Atsushi Yamada[†††]

[†]京都大学　[††]国際電気通信基礎技術研究所　[†††]京都高度技術研究所

[†]Kyoto University　[††]ATR　[†††]ASTEM

研究期間　平成 19 年度～平成 21 年度

概要

音声認識技術に基づいて、国会討論の会議録作成支援と大学講義の情報保障（字幕化）支援を行うシステムを研究開発した。国会討論に関しては、会議固有の話し言葉の精緻なモデル化を行って、85～90%の認識精度を実現し、衆議院の次期会議録作成システムに供することができた。大学の講義に関しては、専門分野や講師の話し方に効率的に適応する方法を研究し、要約筆記の補完・代替となるシステムを開発し、聴覚障害のある学生向けに試験評価を行った。また、音声認識結果から可読性の高い字幕を生成するための自然言語処理についても研究を行った。

Abstract

We have developed an automatic transcription system for meetings of the Diet (National Congress) and classroom lectures of universities. We have proposed an elaborate modeling technique for spontaneous speech, and achieved accuracy of 85-90% for meetings of the Diet. The automatic speech recognition (ASR) modules will be used in the House of Representatives. We have also studied adaptation techniques for topics and speakers of lectures, and developed an ASR-based note-taking system to assist hearing-impaired students in classrooms. Furthermore, we have investigated natural language processing for cleaning transcripts to improve their readability.

１．まえがき

本研究開発は、音声認識技術を発展させることにより、会議録や字幕の作成支援を行うシステムの構築をめざして行われた。具体的には、国会討論と大学の講義を主要なターゲットとして、内容が十分に把握できるレベル（85～90%程度の精度）の書き起こしを自動的に作成し、人手による修正と組み合わせて運用できることを目標とした。

衆議院では、速記制度に代わる新たな会議録作成の方式として、音声認識技術の導入を検討してきたが、従来の音声認識システムでは、自発性が高く、丁々発止の議論がなされる委員会の審議には対応できなかった。これに対して、本研究開発では、話し言葉の精緻なモデル化を行うことで、既存のものとは一線を画したシステムを構築する。

大学の講義においては、聴覚障害のある学生に対して情報保障を提供することがきわめて重要である。現在、手書きのノートテイクやパソコン要約筆記が主に用いられているが、大学の講義を専門分野の異なる学生が聞き取るのは困難で、ボランティアの確保が難しい。これに対して、本研究開発では、講師の音声を直接音声認識し、１人の作業者でも字幕付与できるシステムを構築する。

２．研究内容及び成果

２．１　要素技術に関する研究

(1)　音声・言語コーパスのアノテーション

『日本語話し言葉コーパス』(CSJ)に対して、整形文や句読点･段落などの境界のアノテーションを行った。また、衆議院の本会議・各種委員会の審議音声・忠実な書き起こしと会議録を対応付けた『衆議院審議コーパス』を合計 300 時間の規模に集積した。

(2)　音声認識のための音響モデル・言語モデルの構築

『日本語話し言葉コーパス』(CSJ)を用いて講演や講義の音声認識用のモデルを、『衆議院審議コーパス』を用いて国会討論の音声認識用のモデルを、それぞれ構築した。

言語モデルについては、本研究者が提案している統計的言語モデル変換法を実装し、効果を確認した。この方法は会議録 W から、元の忠実な発話 V のパターンを推定するものである。従来の音声認識システムでは、言語モデル確率 p(V)を推定するのに、忠実な書き起こし V のコーパスを必要としていたが、話し言葉を大規模に書き起こすのは（コスト面で）困難なため、性能の限界があった。これに対して提案手法では、既存の膨大な会議録・講演録 W を話し言葉に（統計量レベルで）変換することでこの限界を打破している。

さらに、本手法を用いて音声の書き起こしを推定することで、音響モデルの効率的な学習（準教師つき学習）を行う方法も考案・実装した。従来の音響モデル学習では、音響モデル確率 p(X|V)の分布を推定するために、音声データ X に対応する忠実な書き起こし V を必要としていたが、上述のようにこの量には限界があった。これに対して、会議録 W から V を予測する（発話毎に p(V)を推定して音声認識を行う）ことにより、人手による書き起こしをしなくてもモデル学習ができるようになり、言語モデルと同様に学習データ量の限界を打破している。

このように、音声データと会議録テキストのみで音響・言語モデルの学習・更新が可能になったので、将来（総選挙や内閣改造などに伴う話者集合の大きな変更）におけるモデル更新も容易にしている。

(3)　話者・話題に応じた音声認識システムの適応

大学の講義では専門用語が多用されるため、ベースラインのモデルでは十分に対応できない。そこで、講義で使用されるスライドからキーワードを抽出し、確率的潜在意味解析(PLSA)や関連Webテキスト収集の枠組みで言語モデルを話題に適応する方法を検討した。また、当該講師の以前の講義データを用いて、音響モデル・言語モデルともに適応したところ、認識精度の大きな改善が得られた。

(4)　話し言葉の整形・区分化に関する研究

会議録や講演録を作成する際に、速記者・校閲者が書き起こしに対してどのような修正・整文を行っているかモデル化を行った。具体的には、編集の過程を統計的機械翻訳の枠組みで定式化・学習を行った。また、節・文の単位への区分化や句読点の挿入についても研究を行った。

２．２　国会討論向けシステムの開発と評価

前項(2)で述べた音響・言語モデルを、研究代表者らが開発を進めている音声認識エンジン Julius に統合することで、国会討論用の音声認識システムを構成した。2009 年 8 月の総選挙に伴う政権交代及び議員・閣僚の大規模な入れ替わりに対応するために、2009 年 10 月召集の第 173 回国会のデータを収集して、音響モデル及び言語モデルの更新も行った。

その結果、2010 年の委員会審議から抽出した評価セットに対して、文字正解率 88.7%、文字正解精度 86.6%を達成した。このレベル（おおむね 80%以上）の正解精度であれば、会議録作成支援に利用できることは、プロトタイプシステムで検証されている。

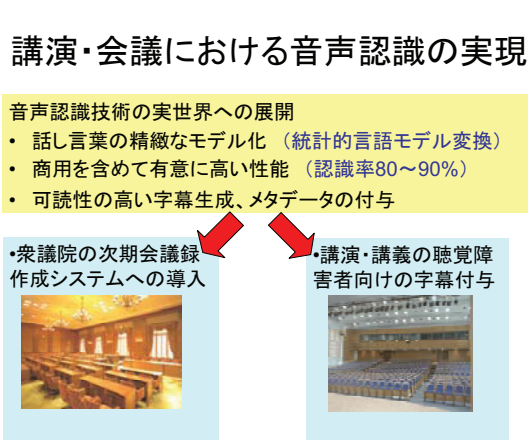
２．３　大学講義向けシステムの開発と評価

大学講義向けの音声認識システムは、前項(3)で述べたように、『日本語話し言葉コーパス』(CSJ)で学習した音響モデル・言語モデルを講師・話題に適応した上で、Julius に統合することにより構成した。音声認識結果の修正・編集には、パソコン要約筆記で一般的に用いられている IPtalk の「確認修正パレット」の機能を利用することにした。パソコン要約筆記者が慣れ親しんでいるソフトウエアの方が受け入れやすいと考えたためである。Julius と IPtalk のインタフェースソフトウエア(Julius2IPtalk)は、研究分担機関の ASTEM で作成した。このソフトウエアはフリーで公開している（本稿末尾の URL 参照）。

京都大学において、聴覚障害学生が受講している講義で評価を行ったところ、単語正解精度で約 60%、キーワード抽出精度で約 80%を実現した。

同じ講義で情報保障を試みたところ、90 分の講義に対して 1 名で通して作業できた。この講義では、2 名による手書きノートテイクも行われていたので、本システムの出力とテキスト量において比較を行った。その結果、本システムは、全体のテキスト量において 1.8 倍、キーワードに限定すると 2.6 倍の分量を提示できていた。

平成 21 年 11 月 28 日に、本プロジェクト主催で、京都大学で開催した『聴覚障害者のための字幕付与技術』シンポジウムにおいて、研究代表者が行った講演に対して、音声認識による字幕付与を行った。読上げ原稿に沿って言語モデルを構築しておいたので、音声認識精度も高く、ほぼすべての発話に対して字幕が付与された。約 30 分の講演に対して 1 名の作業者で対応できた。聴衆からは高く評価され、この模様は NHK 教育テレビ「ろうを生きる難聴を生きる」でも放映された。



３．むすび

本研究開発では、国会討論や講演・講義などの公的な場で比較的明瞭に発話される音声を対象として、内容が把握できるレベルの書き起こしの自動生成を行った。国会討論の音声認識に関しては実用レベルに到達したと考えており、このような公的な会議には比較的容易に展開できると考えている。

大学の講義の音声認識・字幕付与に関しては、今後さらなる発展と試験評価を進めていく必要があるが、講義・講演は毎日、全国で膨大な数が行われており、本技術が適用できる潜在的な場は非常に大きい。聴覚障害者に対する情報保障の必要性は改めて指摘するまでもないが、近年国際化の進展に伴い、外国人留学生が増大しており、字幕の付与は非常に有用と考えられる。音声認識技術によって人手の要約筆記をすべて置き換えられるとは考えていないが、大学の講義は専門性が高く、人間でも聞き取り・筆記が容易でない。これに対して、音声認識システムは大量の専門用語や論文などの情報を瞬時に記憶して、的確に表記することができるので、特定のドメインにおいては人間のレベルを上回る可能性を十分に有するものと考えている。

【誌上発表リスト】

[1] 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.

[2] 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.

[3] Y.Akita, M.Mimura, and T.Kawahara. Automatic transcription system for meetings of the Japanese National Congress. In Proc. INTERSPEECH, pp.84--87, 2009.

【報道発表リスト】

[1] ここまで来たリアルタイム字幕, NHK 教育テレビ「ろうを生きる難聴を生きる」, 2007 年 10 月 13 日.

[2] 今ノートテイクを考える～「２００８ノートテイクシンポジウム」～, NHK 教育テレビ「ろうを生きる難聴を生きる」, 2008 年 10 月 19 日.

[3] 情報保障の可能性を広げよう ～「字幕付与技術シンポジウム２００９」から～, NHK 教育テレビ「ろうを生きる難聴を生きる」, 2009 年 12 月 13 日.

【本研究開発課題を掲載したホームページ】

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