様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　６　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５７３０１１２

研究課題名（和文）音声認識と自動整形の統合的なモデル化に基づく字幕生成の研究

研究課題名（英文）Study of automatic captioning based on unified modeling of spontaneous speech recognition and automatic editing

研究代表者

秋田　祐哉（Akita, Yuya）

京都大学・経済学研究科・講師

研究者番号：９０４０２７４２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：講義や講演などの話し言葉には冗長な表現や口語表現が含まれるため，音声認識を字幕などに活用する際は，まず音声認識器が話し言葉特有の表現をカバーした上で認識を行い，その結果に含まれる冗長表現・口語表現を読みやすく整形するというアプローチが取られる．本研究ではこれらを話し言葉への変換とその逆変換としてとらえ，話し言葉の特徴のモデルを構築し，このモデルに基づき音声認識（変換）と自動整形（逆変換）を行う．本研究では講義・講演の字幕をターゲットとし，これによる字幕の生成・配信システムを構築した．

研究成果の概要（英文）：A variety of redundant and colloquial expressions are observed in spontaneous speech such as classroom lectures and academic talks. Thus, automatic speech recognition (ASR) systems need to cover these kinds of spontaneous expressions to make accurate transcripts, while these expressions in the transcripts are then edited and/or removed for better captions. In this research, models of ASR and automatic editing are considered as transformation to/from the style of spontaneous speech. Characteristics of spontaneous speech are modeled, then the model is applied to ASR and automatic editing. We have developed an automatic captioning system based on this framework for lectures.

研究分野：　音声情報処理

キーワード：　音声認識　自動整形　話し言葉　字幕

１．研究開始当初の背景

人間どうしのコミュニケーションにおいて音声は重要な手段であり，特に講演や講義のような場面では音声は情報伝達の中心となるものである．したがって音声を自動認識してテキスト化することは，情報の記録や伝達の支援といった面で大きな意義がある．研究代表者は講演や講義，会議などのいわゆる話し言葉音声を対象とした音声認識の研究を行っており，たとえば国会審議を対象とした音声認識では 80%を越える精度を達成している．また，話し言葉の音声認識結果を読みやすくするために表現の修正や削除，句読点挿入を行う自動整形の研究にも取り組んできている．ただし，研究代表者も含めて，これまで音声認識と自動整形は別個の取り組みで，それぞれの処理は独立に構成されていることがほとんどである．

２．研究の目的

本研究では，音声認識と自動整形に共通する，話し言葉の特徴のモデルを確立し，このモデルに基づいてそれぞれの処理を行うことを目指す．本研究では講演や講義の音声のために音声認識器を学習し，音声認識と自動整形を経て字幕を生成するタスクを想定する．音声認識器（具体的には言語モデル）の学習の際に話し言葉へのスタイルの変換を行い，認識文を字幕とする際に書き言葉への逆変換を行う．ここで字幕をターゲットとしているのは，字幕は読みやすさ・分かりやすさの大きく問われるテキストであることから整形の目的として適当であり，また実際に字幕を必要とする人々が存在するからである．

３．研究の方法

(1)音声と字幕データの収集・分析
音声認識・自動整形の処理内容を検討するため，実際の講義・講演とその書き起こしを収集する．研究代表者は大学に所属しているため，大学での講義や講演を収集の対象として想定している．収集したデータに対して実際に字幕を作成し，これをもとにどのような整形が適当か分析を行う．

(2)モデル化の枠組みの検討と精緻化
音声認識と自動整形を実現するための話し言葉のモデルについて，各種の音声・テキストデータベースや(1)の分析結果を用いて，種々の枠組みの比較検討を行う．具体的には機械学習や統計翻訳などの手法を想定している．また，収集した音声・テキストデータを使用して話し言葉のモデルの洗練を図る．具体的にはモデルの詳細化（パラメータ等の追加）や学習データの追加である．音声認識と自動整形の接続（入出力）形式の検討や，音声認識誤りの影響への対処も行う．さらに，音声認識結果における精度と自動整形結果における読みやすさを基準として話し言葉のモデルの評価を行う．特に後者については，実用上の観点から，実際に字幕として提示した場合の読みやすさに関する評価も実施する．

(3)配信システムの構築
音声認識・自動整形の各モジュールを構成し，講義・講演音声から字幕テキストを自動生成するシステムを構築する．また，字幕を提示するインターフェースについて設計を行い，配信システムのプロトタイプを構築する．プロトタイプの字幕提示インターフェース・配信システムについて，実際の講義・講演データを用いてテストを行い，速度や効率などの観点から性能を検証し，本システムの設計に反映する．本システムを実際の講義・講演の場面に導入し，リアルタイムの字幕作成と配信を行うことで，音声認識・自動整形による字幕の評価も行う．

４．研究成果

本研究の成果物である，自動字幕作成システムについて述べる．本研究では，専門家・技術者でない利用者でも字幕の作成・編集が行えるよう，入力された音声に対して自動的に音声認識をセットアップ・実行して字幕草稿を作成するサーバと，字幕草稿を編集するために設計されたエディタからなる，字幕作成・編集システムを構築した．本システムでは音声に対して具体的な資料を与えて言語モデルや単語辞書の適応処理を行うことができる．これと関連して，本システムではリアルタイムに字幕を作成するための音声認識サーバとして利用することも可能である．

(1)字幕作成・編集システムの構成
本システムでは，ユーザにより収録された講義・講演や討論などの音声・映像に対して，事後的に字幕を付与することを想定している．まずユーザがこれらのコンテンツを字幕サーバにアップロードする．音声・映像に加えて，言語モデルを話題に適応させるために，コンテンツの話題と関連するテキスト（たとえば講演予稿やスライド）もアップロードすることができる．字幕サーバではコンテンツからの音声の抽出および検査が行われ，ユーザの指定に応じて自動的に音声認識システムが構成された上で認識処理が実行される．認識処理はおおむね 1 日以内に終了し，音声と同期した字幕ファイルがサーバ上に出力されるとともに，これらにアクセスするためのアドレスがユーザに通知される．なお，ここで音声認識結果を用いる代わりに，あらかじめ人手で書き起こされた正しいテキストを与えて，音声への同期処理のみを行うこともできる．
ユーザは通知されたアドレスから字幕をダウンロードできる．また，Web ブラウザ上で字幕エディタをダウンロード・起動して，サー

バ上の音声を聴取しながら字幕のテキストや時刻を編集することもできる．編集した結果はサーバ上に保存され，更新された字幕ファイルとして取得可能である．

(2)字幕サーバ

字幕サーバは，ユーザからのコンテンツを処理するフロントエンド，音声認識，後処理・字幕生成の 3 つの機能で構成されている．

字幕サーバには，コンテンツとして PCM（Microsoft WAV）や MP3 形式の音声のほか，MPEG 等の映像ファイルを入力することができる．入力がいずれの形式であっても，フロントエンドで 16kHz・16bit のモノラル音声に変換されて処理される．なお，音声に対してはビットレートや周波数の分布，SN 比のチェックが行われ，これらが一定の品質条件を満たさない場合は音声認識に進まず処理を中止する．コンテンツに関連する文書としては，プレーンテキストのほか PDF や Microsoft Word/PowerPoint などの文書も受け付ける．フロントエンドによりこれらの文書から自動的にテキスト部分が抽出され，次節で述べる音声認識のための言語モデル適応に用いられる．

本システムでは，講演や討論など，コンテンツの種類に応じていくつかの音響モデル・言語モデルの組み合わせをプロファイルとして用意しており，ユーザは実際に音声認識に利用するプロファイルをコンテンツのアップロードの際に選択することができる．現時点でのプロファイルとして，たとえば，「講演」には『日本語話し言葉コーパス』(CSJ) の学会講演データから学習したモデルを，「討論」には国会音声・会議録から学習したモデルを用意している．

サーバで行われる処理の流れは次の通りである．前述した音声品質チェックの後，コンテンツとともに関連文書が与えられている場合は，選択されたプロファイルの言語モデルに対してテキスト混合に基づく適応が行われる．次に，音響モデルが GMM-HMM の場合は，音声の話者区間の推定・分割を行う．話者区間に分割するのは後段の声道長正規化 (VTLN) のためであり，DNN-HMM モデルの場合は VTLN を行わないので，セグメンテーションは省略される．1 回目の音声認識は，GMM-HMM モデルの場合は VTLN のワープ係数を推定することが目的のため，簡易なモデル・パラメータで行われる．この結果を用いて，音声の区間ごとに VTLN を適用し，2 回目の音声認識を行う．話者が 1 名の場合は，この認識結果を教師なし音素ラベルとして，さらに MLLR による話者適応を音響モデルに適用したのち 3 回目の音声認識を行う．DNN-HMM モデルでは VTLN および MLLR 話者適応を行わないため，1 回目の音声認識のみ行う．本システムで用いる音声認識エンジンは Julius である．

音声認識結果には各単語の推定時刻が付与されているので，これを表示のタイミングとして，認識文からなる字幕ファイルを作成する．この際，音声認識結果には文や節の境界は与えられていないため，句読点の自動推定を用いて字幕の行に分割する．また，フィラーや口語表現，文末表現などの冗長部分を削除・修正するため，自動整形手法を適用する．一方，あらかじめ人手による字幕テキストが与えられた場合は，このテキストと音声認識結果との間で文字単位のアライメントを行って時刻を字幕テキストに付与し，字幕ファイルを作成する．字幕ファイルは SAMI・SRT など，複数の形式で出力される．これらの字幕はサーバに保存され，インターネットからアクセスが可能である．

(3)字幕エディタ

本システムでは生成された字幕を編集するためのエディタを提供する．音声認識結果の編集システムには様々な商用ソフトウェアがあり，またサーバ上で音声認識結果を編集するアプリケーションもあるが，これらに対して，字幕に特化して行単位でテキストと時刻を編集するエディタであること，サーバ上のデータをオンラインで聴取・編集できることが本エディタの特徴といえる．オンラインの編集により，任意の場所や多様な環境で作業が可能である．エディタは Java アプリケーションとして実装されている．字幕の編集は，語句の修正や文の長さ・タイミングの調整を主に行うフェーズと，最終的な字幕の出力を確認するプレビューのフェーズに大別できることから，本システムではそれぞれに合わせた 2 つのエディタを持つ．

(4)放送講義における字幕作成

本研究では，提案システムによる字幕作成の効率を測定するため，実際の講義音声を用いて字幕の作成を行った．ここで利用したのは放送大学で実施されたラジオ講義 2 科目，計 27 講義である．講義の長さは 1 件あたり 45 分となっている．これらの講義音声を字幕サーバに入力して，得られた音声認識結果を字幕エディタにより編集して字幕を作成した．言語モデルの適応用テキストとしては，講義の教科書に加えて，台本がある場合はこれも使用した．ただし台本は必ずしも全ての発話を網羅しているわけではなく，また正確とは限らない．字幕作成にあたった作業者は 1 名で，計算機の一般的な操作スキルはあるが，字幕作成の専門家ではなく，またこれらの講義内容の専門家でもない．作業にあたり，講義ごとに所要時間を計測し，あわせて文字正解率を算出した．所要時間は編集時間（主に認識誤りの修正や文の長さ・タイミングの調整に要した時間）と確認時間（プレビュー用のエディタ上で行った出力チェックに要した時間）に分けて計測した．実時間比として，合計の作業時間を講義の長さ（45 分）で除して計算したところ，平均して実時間の 5.2〜6.1 倍の作業時間となっており，このうちほぼ実時間分は確認作業に費やされた．

科目ごとに各講義の文字正解率と編集時間から相関の程度を測ったところ，これらに強い相関があることがわかった．同様に放送大学の講義を対象とした字幕作成の報告では，45 分の講義に対して約 4 時間の書き起こし時間を必要としていることから，86%～87%以上の精度が得られれば，はじめから人手で書き起こすよりも効率的であるといえる．

(5)リアルタイムの字幕作成システム

これまでに述べた字幕作成は配信を前提とした事後的な処理である．一方，講義・講演の会場で情報保障のためにリアルタイムに字幕を作成・表示するシステムの開発にも取り組みを行った．

講義や講演において，文字情報による情報保障の手段としては，手書きのノートテイクや PC を用いた要約筆記などが一般的に用いられている．手書きにより書き起こせる分量は限られており，遅延をともなう．また，いずれの方法であっても作業者が長時間作業し続けることができないため，複数の作業者を用意する必要があり，情報保障を提供する上での支障となっている．これに対して，音声認識により精度よく書き起こすことができれば作業の負担が軽減でき，より少ない人数（たとえば 1 人）で作業できることから，情報保障の機会を拡大することができる．

我々が検討した，音声認識を用いたリアルタイム字幕作成システムは，講義・講演会場で作業者が入出力・編集に使用する PC と，音声認識を行うサーバから構成される．PC とサーバはネットワークを通じて通信するため，十分な通信帯域が確保できれば，サーバは遠隔地に設置することができる．

講師の音声は PC に入力され，Julius 付属の音声入力ツールである Adintool によって発話検出・セグメンテーションを行ったのち，サーバ側の Julius にネットワーク経由で送信される．サーバではあらかじめ対象の講義・講演用に構成された音響モデル・言語モデルを使用して音声認識を行い，この結果をネットワーク経由で作業者の PC に送信する．なお，本研究では音響モデル・言語モデルとして講演プロファイルに相当するものを主に使用している．GPU を用いてデコードすることにより，DNN-HMM 音響モデルでも実時間以下の処理時間で音声認識を行っている．

作業者の PC では，PC 要約筆記で一般的に用いられている IPtalk を字幕の編集ツールとして使用する．IPtalk では複数の要約筆記者がネットワーク経由で連携して入力することができるが，本システムでは Julius の出力をブリッジするツール Julius2IPtalk を使用して，IPtalk の「確認・編集パレット」に認識結果を入力する．作業者が認識結果をチェック・修正した上で送出すると，字幕として表示される．字幕は，PC に接続したプロジェクタや，ネットワークを通じた Web 配信などで表示可能である．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計７件）

(1) Y.Akita, N.Kuwahara, and T.Kawahara. “Automatic Classification of Usability of ASR Result for Real-time Captioning of Lectures.” APSIPA ASC, 2015 年 12 月 16 日～19 日，香港（中国）.

(2) 秋田祐哉，三村正人，河原達也.「音声認識を用いた講義・講演の字幕作成・編集システム」. 情報処理学会音声言語情報処理研究会，2015 年 10 月 30 日，早稲田大学（東京都新宿区）.

(3) Y.Akita, Y.Tong, and T.Kawahara. “Language Model Adaptation for Academic Lectures using Character Recognition Result of Presentation Slides.” IEEE-ICASSP,2015 年 4 月 19 日～24 日,ブリスベン(オーストラリア).

(4) 大田健翔，秋田祐哉，河原達也.「講演音声認識結果の誤り箇所の復唱入力を用いたノートテイクシステム」. 情報処理学会全国大会，2015 年 3 月 17 日～19 日，京都大学（京都府京都市）.

(5) 童弋正，秋田祐哉，河原達也.「講演スライドの文字認識結果を用いた音声認識の改善」. 情報処理学会音声言語情報処理研究会，2014 年 7 月 24 日～26 日，ホテル花巻（岩手県花巻市）.

(6) 桑原暢弘，秋田祐哉，河原達也.「音声認識結果の有用性の自動判定に基づく講義のリアルタイム字幕付与システム 」. 日本音響学会春季研究発表会，2014 年 3 月 10 日～12 日，日本大学（東京都千代田区）.

(7) 秋田祐哉，河原達也.「音声認識を用いたオンライン自動字幕作成・編集システム」.日本音響学会秋季研究発表会,2013 年 9 月 25 日～27 日，豊橋技術科学大学（愛知県豊橋市）.

〔その他〕

ホームページ等
http://caption.ist.i.kyoto-u.ac.jp/

６．研究組織

(1)研究代表者

秋田　祐哉（AKITA, Yuya）
京都大学・大学院経済学研究科・講師
研究者番号：９０４０２７４２