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　２７　年　　６　月　２２　日現在

機関番号：　１３９０４
研究種目：　若手研究(B)
研究期間：　2011～2014
課題番号：　２３７００１１５

研究課題名（和文）スライドと音声を組み合わせた講義コンテンツの構造化と要約に関する研究

研究課題名（英文）Construction and Summarization of Lecture Contents Using Both Slides and Lecture Speech

研究代表者

土屋　雅稔（Tsuchiya, Masatoshi）

豊橋技術科学大学・情報メディア基盤センター・准教授

研究者番号：７０３７８２５６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：講義音声には，多くの話し言葉的現象(フィラー，ポーズなど)が含まれる．講義音声を要約するには，それらの話し言葉的現象に対して頑健な自動音声認識を実現する必要がある．日本語話し言葉コーパスに収録された音声情報からポーズ出現位置に関するモデル(ポーズ挿入モデル)を学習し，ポーズ情報を含まないコーパス(国会会議録)を組み合わせることによって，ポーズに対応した言語モデルを構築する方法を提案し，その有効性を示した．
また，講義音声に頻出する講義内容に特有の固有的な事物を，検出漏れをできるだけ少なく検出する方法について検討した．講義スライドと講義音声書き起こしとの人手対応付けの作業手順の作成を行った．

研究成果の概要（英文）：Because lecture speech contains spoken phenomena such as filled pauses and silent pauses, a robust automatic speech recoginition method is necessary in order to realize automatic summarization of lecture speech. Our method consists of two steps: 1st step is to predict filler insertion locations and pause insertion locations against loosely transcribed corpora which has no pause information using filler insertion model and pause insertion model learned from precisely transcribed corpora including filler information and pause information, and 2nd step is to construct a language model based on both loosely transcribed corpora and predicted information.
And more, a method to detect lecture specific named entities was developed. The human annotation scheme to map lecture slides and lecture speech transcriptions was also established.

研究分野：　自然言語処理

キーワード：　自動要約

１．研究開始当初の背景

ネットワーク環境の進歩とストレージの低価格化にともない，各種教育機関において，講義音声・動画および講義スライドからなる大規模講義アーカイブを公開する動きが広がっている．このようなアーカイブを高度に利活用するためには，目次や索引などのメタデータの付与を行ってコンテンツを構造化し，同時に要約を付与することが不可欠である．例えば，講義音声は，重要な部分だけを選択的に視聴することが困難なので，効率的に視聴できる環境を実現するには，講義内容の索引と講義音声の自動要約が必要である．

２．研究の目的

本課題では，スライドと音声を組み合わせて講義コンテンツを構造化・要約する方法を研究する．講義音声の自動要約を実現するために，講義音声に含まれる話し言葉的現象に対して頑健な音声認識手法について検討する．さらに，講義スライドの構造情報と講義音声の対応付けを行って，講義スライドの構造情報に基づいて講義音声を要約する手法を検討する．

３．研究の方法

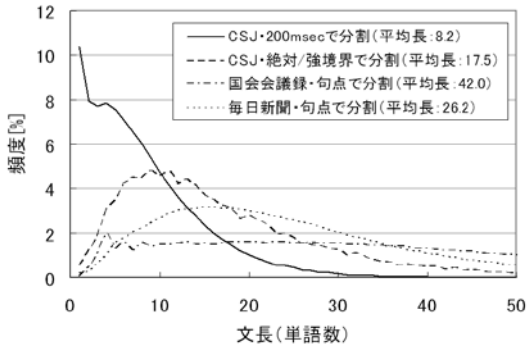
講義音声には,多くの話し言葉的現象(フィラー,ポーズ,言い淀み,言い直しなど)が含まれる．最初に，これらの話し言葉的現象に対して頑健な自動音声認識技術について研究し，講義音声の自動音声認識を実現する．特に，話し言葉的現象の中でも出現頻度の高いフィラーとポーズに注目して研究を行う．

並行して，講義スライドと講義音声の対応付けについての研究を行う．具体的には，最初に，講義スライドからスライド構造情報を抽出する．典型的な講義スライドでは，重要度は文字の大きさや色・飾りによって，内容間の依存関係はインデントによってグラフィカルに表現されている．このような情報を，重要度を根からの距離，依存関係を枝として表した木構造に抽象化して抽出する．得られたスライド要素と発話を，含意関係に基づいて対応付ける．この対応付けの研究を行うため，実際の講義音声とスライド要素の人手対応付けを行って正解データを作成する．また，講義音声および講義スライドの両方の言語表現において，重要な役割を果たす機能表現についても研究する．

講義音声には，講義内容に特有の固有的な事物を表す固有表現も頻出する．これらの固有表現は，講義内容を理解するために重要であるから，再現率良く検出する手法が必要である．

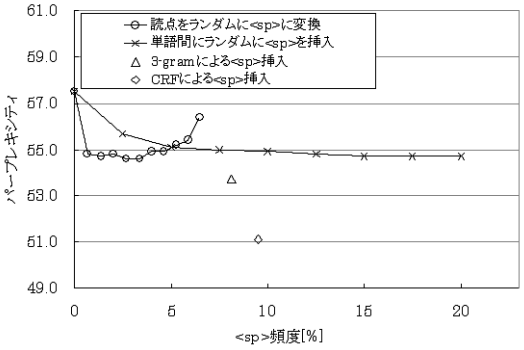
４．研究成果

講義音声には,多くの話し言葉的現象(フィラー,ポーズ,言い淀み,言い直しなど)が含まれる．講義音声の自動要約を実現するためには，これらの話し言葉的現象に対して頑健な自動音声認識技術が必要である．最初に，話し言葉的現象の中でも，ポーズについての研究を行った．自発的に発声される話し言葉音声には,言語的な区切りとは無関係な位置に多数のポーズが出現する．例として，日本語話し言葉コーパス(CSJ)における転記基本単位と節単位，および国会会議録と毎日新聞における文単位の長さの分布を下図に示す．



図より，CSJ の転記基本単位は，国会会議録および毎日新聞の文単位に比べて，非常に短い単位が多いことが分かる．このように，ポーズに基づいて得られた処理単位と，コーパス作成者または新聞記事の筆者によって設定された文単位との間には明白な不整合が存在する．よって，コーパス中の句読点をポーズとみなして言語モデルを学習すると，ポーズのモデル化は不十分になってしまう．

この問題に対応するため,日本語話し言葉コーパスに収録された音声情報からポーズ出現位置に関するモデル(ポーズ挿入モデル)を学習する．具体的には，形態素列を対象とし，個々の形態素に対して，その直後にショートポーズを挿入するべきかどうかという二値のラベルを付与する，系列ラベリング問題として定式化し，CRF を用いて学習を行った．各種のポーズ挿入モデルを用いて，ポーズを挿入・比較した結果を下図に示す．

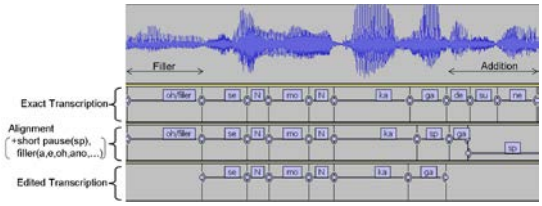


図より，句読点をポーズとして用いる従来手法や，直前の 3-gram のみを用いる手法，単語間にランダムにポーズを挿入する手法に比べて，提案手法である CRF を用いるポーズ挿入モデルの性能が良いことが分かる．

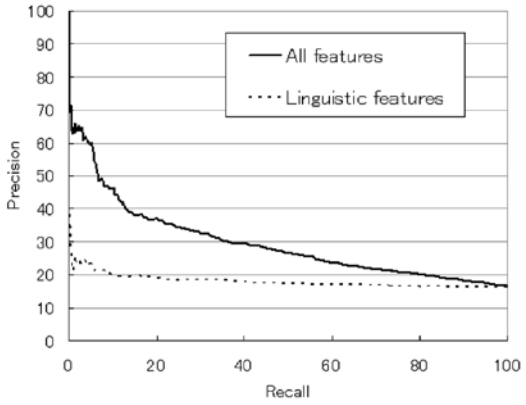
さらに，得られたポーズ挿入モデルと，ポ

ーズ情報を含まないコーパス(国会会議録)を組み合わせることによって,ポーズに対応した言語モデルを構築し，音声認識実験を行った．全ての句読点をポーズとして用いた場合には，音声認識精度(Acc)は 61.8 だったのに対して,提案手法では 64.4 と改善された．

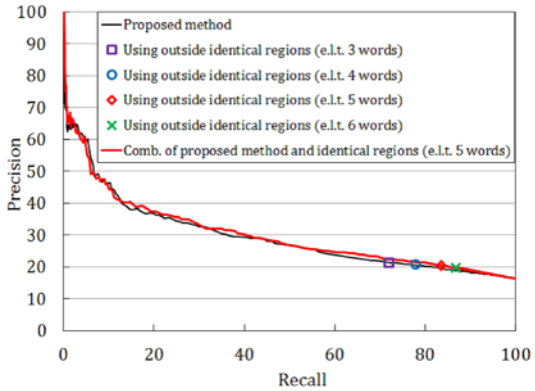
次に，言い淀み・言い直しについての研究を行った．言い淀み・言い直しについては，言い淀み・言い直しに対応した言語モデルを構築するために,言い淀み・言い直し情報が整形されて失われている国会会議録テキストと元の音声情報とを対応付ける方法について検討した．下図に，言い淀み・言い直し情報を含む正確な書き起こしと，整形された国会会議録テキスト，および音声波形の例を示す．



提案手法は，音響的素性に基づく会議録テキストと音声情報の強制アラインメントと,強制アラインメントによって得られた素性に基づいて整形個所を検出する識別器からなる．音響的素性と言語的素性を組み合わせて検出を行った場合と，言語的素性のみで検出を行った場合の比較結果を，下図に示す．



図より，音響的素性は整形箇所の検出に有効であることが分かる．また，従来手法である連続音声認識結果を用いた整形箇所の検出手法との比較結果を下図に示す．



図より，連続音声認識結果を用いる従来手法と，音響素性と言語素性を用いて識別する提案手法とを，組み合わせて用いる方法が，もっとも精度よく整形箇所を検出していることが分かる．

この組み合わせ手法によって検出された整形箇所以外の音声区間を用いて音響モデルの適応を行ったところ，適応前の音響モデルでは音声認識精度(Acc)が 67.2 だったのに対して，適応後は 71.0 に達した．正確な書き起こしを人手で準備して，全ての音声区間を用いて音響モデルの適応を行った場合の音声認識精度(Acc)は 71.6 である．したがって，本提案手法によって検出された整形箇所以外の音声区間を用いる音響モデル適応は，全区間を用いる場合にかなり近い性能を達成していることが分かる．

また,日本語には,複数の語がひとかたまりの表現として非構成的な意味を持ち,機能的関係を表すようになった機能表現が多数存在する.含意関係抽出を行うには,意味的に類似している内容語だけでなく,意味的に類似している機能語と機能表現(例えば,「について」と「に関して」)の差異を吸収する必要がある.その前段階として,多種多様な機能表現の用法を判定する方法について検討した.

講義音声には,講義内容に特有の固有的な事物を表すための固有表現が頻出する.これらを再現率良く検出するための方法について検討した.

講義スライドと講義音声書き起こしとの人手対応付け(正解データの作成)の作業を行った.小規模データを対象とし,複数の作業者(研究代表者を含む)が対応付けを試みたところ,作業者間の一致度がかなり低いことが判明した.この問題に対応するため,人手対応作業手順書の整備を行った.その上で,複数講義の講義スライドと講義音声書き起こしの人手対応付けを行い,その上で,対応付けの自動推定を試みた.

５．主な発表論文等

〔雑誌論文〕（計 1 件）

太田健吾，土屋雅稔，中川聖一．ポーズを考慮した話し言葉言語モデルの構築．情報処理学会論文誌，Vol.53, pp.889-900, 2012.

〔学会発表〕（計 7 件）

太田健吾，土屋雅稔，中川聖一．整形された書き起こしからの整形・非整形部分の自動検出.第 6 回音声ドキュメントワークショップ，2012.

Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa. Detection of Precisely Transcribed Parts from Inexact Transcribed Corpus. Automatic Speech Recognition and Understanding Workshop (ASRU2012), 2012.

鈴木敬文，阿部佑亮，宇津呂武仁，松吉俊，土屋雅稔．代表・派生関係を利用した日本語機能表現の解析方式の評価．言語処理学会第18回年次大会，2012.
鈴木敬文，阿部佑亮，宇津呂武仁，松吉俊，土屋雅稔．『現代日本語書き言葉均衡コーパス』における複合辞の検出と評価．コーパス日本語学ワークショップ，2012.
Kengo Ohta, Masatoshi Tsuchiya, Seiichi Nakagawa. Developing Partially-Transcribed Speech Corpus from Edited Transcriptions. The 8th International Conference on Language Resources and Evaluation (LREC2012), 2012.
川口亮，土屋雅稔，中川聖一．音声ドキュメント中の人名抽出．日本音響学会2013年春季研究発表会，2013.
川口亮，土屋雅稔，中川聖一．プライバシ保護のための音声からの人名除去とその評価．日本音響学会2013年秋季研究発表会，2013.

６．研究組織

(1)研究代表者

土屋雅稔（TSUCHIYA, Masatoshi）
豊橋技術科学大学・情報メディア基盤センター・准教授
研究者番号：70378256