様式 C-19

科学研究費補助金研究成果報告書

平成 22 年 4 月 15 日現在

研究種目：基盤研究（B）
研究期間：2007～2009
課題番号：19300061
研究課題名（和文）
話し言葉音声コミュニケーションの構造の抽出と視覚化
研究課題名（英文）
Structure Extraction and Visualization of Spontaneous Speech Communication
研究代表者
河原　達也　（KAWAHARA TATSUYA）
京都大学・学術情報メディアセンター・教授
研究者番号：00234104

研究成果の概要（和文）：
講演･講義や会議･ミーティングなどの大規模な音声アーカイブの効果的な利活用を指向して、このような長時間の話し言葉音声を自動書き起こし（音声認識）するとともに、多層の言語的・談話的構造を抽出し､字幕化を含めて効果的に提示する方法について研究を行った｡学会講演、大学講義、及び国会審議の大規模なコーパスを用いて、音声認識・筆記録作成支援を行うシステムを構築した。

研究成果の概要（英文）：
For effective exploitation of large-scale audio archives such as lectures, conferences and meetings, we investigate automatic speech recognition of these kinds of spontaneous speech communication, as well as extraction of linguistic structures and effective presentation. Automatic transcription systems for academic lectures, classroom lectures and parliamentary meetings are implemented.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	5,400,000	1,620,000	7,020,000
2008 年度	4,200,000	1,260,000	5,460,000
2009 年度	4,200,000	1,260,000	5,460,000
総　計	13,800,000	4,140,000	17,940,000

研究分野：　音声言語処理
科研費の分科・細目：　知覚情報処理・知能ロボティクス
キーワード：　音声言語処理, 話し言葉, 音声認識, 言語解析, メタデータ付与, メディア検索

１．研究開始当初の背景

音声によるコミュニケーションは、太古より人間どうしの知識伝達・意見交換の根源的な手段であり、電子的媒体が発達した現在においても、新たな知の創造は、主にセミナーやミーティングなどの場で行われていると考えられる。研究代表者らは従来より、講演・討論などの話し言葉を対象として、音声認識の研究を行っている。しかしながら、このような話し言葉音声は、考えながら発話されるので、まとまりのない文や言い淀みが多く、たとえ 100%忠実に書き起こしても、かえって読みづらい反面、音声に含まれるニュアンスなどが失われてしまう。実際に、字幕や講演録・会議録を作成する際には文書体に整形（整文）し、ノートやメモをとる際には適度な要約を行う必要がある。また、多くのアプリケーションを考えると、音声アーカイ

ブを（書き起こしの情報を利用しながら）効率的に検索・ブラウジングする枠組みが重要であるが、その際には適切な境界・構造の抽出が必要である。

２．研究の目的

本研究では、講演・講義や会議・ミーティングなどの大規模な音声アーカイブの効果的な利活用を指向して、このような長時間の話し言葉音声から、多層の構造を抽出するとともに、視覚化を含めた効果的な提示を行うことを目指す。まず、音声言語処理に関する基礎的な研究として以下を行う。

(1) 実環境における発話区間検出
騒音や背景話者が存在する状況での長時間録音から、目的話者の発話区間を頑健に検出する方法を研究する。
(2) 話し言葉の音声認識
講演・講義や会議・ミーティングなどの話し言葉を対象とした音声認識の高精度化を図る。
(3) 話し言葉の構造のモデル化
話し言葉音声における節や文の構造、さらには談話の構造を分析･モデル化する。これにより、音声認識（Speech-To-Text）システムの機能的な改善を図る。また、ポーズやピッチなどの韻律との関係について調べる。
(4) 通常の文書レベルの構造抽出
以上の分析･モデル化に基づいて、句点・読点や改行・インデントに相当するパラグラフ境界の検出を行う。さらに、並列構造に基づいて箇条書きなどの構造の抽出や、引用節の抽出による引用符の付与などについても検討する。
(5) 発言体と文書体の対応のモデル化・変換
講演や会議の書き起こしと、それらから作成された講演録・会議録を比較することにより、発言体と文書体の対応付けを統計的に学習する。これは、講演録や会議録の作成支援（発言体→文書体）に必要であるのみならず、現存する膨大な講演録・会議録から音声認識用の発言体の言語モデル（文書体→発言体）を予測する上で有益である

学会講演等からなる『日本語話し言葉コーパス』、及び我々が構築を進めている『衆議院審議コーパス』に対して、上記のメタデータを人手で付与し、種々の統計的モデルの構築を検討する。これを、構造抽出や整形の観点から定量的な評価を行う。これにより、話し言葉音声の書き起こし（音声認識結果）から文書体の日本語へ「整文・フォーマット」するシステム(Speech-To-RichTextFormat)を実現する。

さらに、ミーティングやポスター発表などの多人数会話に対して、映像を含めたアーカイブを構築し、効率的なインデキシング・ブラウジングの方法を研究する。

３．研究の方法

(1) 2007 年度

基礎的な整備が終わっている２つの音声言語データベース（『日本語話し言葉コーパス』、『衆議院審議コーパス』）を用いて、音声言語処理に関する分析・モデル化を重点的に行うとともに、映像を含めた２つのデータベース（講義、ミーティング）の設計と収録を行う。

① コーパスの整備
② 話し言葉の構造のアノテーションと分析
③ 話し言葉コーパスの整形とその分析・モデル化
④ 話し言葉の音声認識のためのモデル化
⑤ 話し言葉の構造と韻律の関係の分析
⑥ 多人数会話コーパスの設計と収録
⑦ アーカイブの利用状況を想定した検索・ブラウジングの検討

(2) 2008 年度

コーパスの収集を引き続き行い、アノテーションの充実を図るとともに、分析・モデル化に基づいて、話し言葉から構造を抽出したり、講演録に整形・フォーマットを施す処理系を作成する。

① コーパスの整備
② 言語情報と非言語情報の分析
③ 話し言葉の音声認識のモデル適応
④ 話し言葉のチャンキング
⑤ 話し言葉の文書体への整文・フォーマットシステム
⑥ 実環境における頑健な発話区間検出
⑦ センシング・コンテクスト情報を利用したインデキシング

(3) 2009 年度

音声認識及び音声言語処理のさらなる高度化を図るとともに、話し言葉を自動的に書き起こして、講演録・会議録・字幕などに整形・フォーマットを施す処理系を作成する。

① 実環境における頑健な発話区間検出
② 実環境における頑健な音声認識
③ 音響モデルの準教師なし学習
④ 音響イベントの検出
⑤ 整文・フォーマットシステムの改善
⑥ 講演・講義のノートテイク支援システム
⑦ 多人数会話アーカイブへのインターフェース

４．研究成果

(1) オンライン変分ベイズ学習に基づくモデル比較を用いた音声区間検出

教師なし・オンラインの音声区間検出方法を提案した。オンライン EM は学習データのない未知の環境にも適用できる枠組みであるが、雑音のみの区間や音声のみの区間が連続すると、モデルの更新が適切に行われないという問題があった。これに対して、提案手法は変分ベイズ EM 学習に基づいており、その過程で得られる自由エネルギーをモデルの信頼度比較に利用するものである。VB-EM をオンライン学習に定式化し、モデルパラメータとモデル信頼度の推定を同時・逐次的に行う。CENSREC-1-C を用いた音声区間検出の評価実験により、提案手法が従来のオンライン EM よりも有意に効果的であることを確認した。

(2) スライド情報を用いた言語モデル適応による講義の音声認識

大学などの講義で使用されるスライドの情報を用いて、言語モデルを動的に適応することにより、音声認識の高精度化を実現する方法を提案した。まず、当該講義のスライド全体のテキストを用いて、PLSA (Probabilistic Latent Semantic Analysis) により N-gram モデルのスケーリングを行う。次に、発話に対応する個々のスライドの情報を用いて、キャッシュモデルによりスライドに現れる単語の確率を強化し、認識結果のリスコアリングを行う。京都大学で行われた技術講習会と正規の講義を対象とした音声認識において評価を行った結果、PLSA による大域的な適応とキャッシュモデルによる局所的な適応を組み合わせることにより、認識精度の有意な改善が得られた。特に、キーワードの検出で大きな改善が見られた。

(3) 言語モデルと発音辞書の統計的話し言葉変換に基づく国会音声認

国会討論などの音声認識のために、言語モデルと発音辞書を話し言葉スタイルに変換する手法を提案した。提案法では、発話の忠実な書き起こしとこれに対応する正書体のデータの相違点が統計的に抽出され、これをもとに確率的な変換パターンからなる変換モデルが言語モデルと発音辞書のそれぞれに対して構成される。このモデルに基づき、言語モデルに対しては話し言葉の N-gram の予測と統計頻度の推定を行う。一方発音辞書に対しては、話し言葉特有の発音変動の予測および発音確率の推定を行う。生成された話し言葉スタイルの言語モデルと発音辞書を衆議院の委員会審議音声で評価した結果、従来手法と比較して有意な改善が得られた。

(4) 局所的な係り受けと韻律の素性を用いた話し言葉の節・文境界推定

我々はこれまでに SVM を用いて節境界・文境界を自動的に推定する手法を提案しているが、本研究では、直後の文節への係り受け情報および一般的な韻律情報の導入による拡張を検討した。『日本語話し言葉コーパス』の講演音声を用いて評価実験を行った結果、隣接文節間の係り受け情報が文境界推定に対して有効であること、および韻律情報が音声認識結果における節・文境界推定に有効であることがわかった。

(5) 文脈を考慮した確率的モデルによる話し言葉の整形

自動音声認識の結果には認識誤りのみならず、言いよどみや口語的表現など、筆記録（講演録や会議録）にふさわしくない現象が多く含まれている。これらの現象を整形し、自然な筆記録を作成するために、認識結果（または忠実な書き起こし）と筆記録を異なる言語とみなし、統計的機械翻訳を用いて認識結果から筆記録へと翻訳する枠組みを検討している。本研究では、この枠組みの中で２つの手法を提案した。まず、文脈情報を考慮した翻訳モデルを導入し、システムのさらなる精度向上を目指す。また、翻訳モデルの条件付き確率と同時確率の対数線形補間を行うことで、高頻度の翻訳パターンを優先的に利用することを可能とする。有限状態トランスデューサー (WFST) による実装を行い、国会会議録と音声認識結果を用いた評価実験を行った。

(6) 多人数音声会話コンテンツを対象とした音リアクションイベント検出

ポッドキャストなどの多人数音声会話コンテンツ中の重要な箇所（ホットスポット）を抽出するための手掛かりとなる音響イベントの検出手法を提案した。本研究では、視聴者が興味を持ちそうな箇所と密接に関係すると思われる、発話者や対話参加者のリアクションに基づく笑い声やあいづちなどの音響イベント（音リアクションイベント）に着目し、ホットスポットの候補区間となる先行発話の区間とともに抽出することを考える。背景音楽が頻繁に混在するポッドキャストにおいて、頑健に区分化と分類を行うために、背景音に応じて分割重みを自動推定した BIC に基づく分割と GMM による識別を組み合わせた手法を提案する。評価実験において、大分類を行って分割重みを切り替える提案手法により、分類・識別の精度が改善され、笑い声やあいづちの検出精度も向上した。

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）

〔雑誌論文〕（計 14 件；すべて査読あり）

(1) D.Cournapeau, S.Watanabe, A.Nakamura, and T.Kawahara. Online unsupervised classification with model comparison in the Variational Bayes framework for voice activity detection. IEEE J. Selected Topics in Signal Processing, (accepted for publication), 2010.

(2) T.Shinozaki, S.Furui, and T.Kawahara. Gaussian mixture optimization based on efficient cross-validation. IEEE J. Selected Topics in Signal Processing, (accepted for publication), 2010.

(3) Y.Akita and T.Kawahara. Statistical transformation of language and pronunciation models for spontaneous speech recognition. IEEE Trans. Audio, Speech & Language Process., (accepted for publication), 2010.

(4) K.Ishizuka, S.Araki, and T.Kawahara. Speech activity detection for multi-party conversation analyses based on likelihood ratio test on spatial magnitude estimation. IEEE Trans. Audio, Speech & Language Process., Vol.18, No. (accepted for publication), 2010.

(5) T.Misu and T.Kawahara. Bayes risk-based dialogue management for document retrieval system with speech interface. Speech Communication, Vol.52, No.1, pp.61--71, 2010.

(6) H.Wang and T.Kawahara. Effective prediction of errors by non-native speakers using decision tree for speech recognition-based CALL system. IEICE Trans., Vol.E92-D, No.12, pp.2462--2468, 2009.

(7) H.Wang, C.J.Waple, and T.Kawahara. Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition. Speech Communication, Vol.51, No.10, pp.995--1005, 2009.

(8) 西光雅弘，秋田祐哉，高梨克也，尾嶋憲治，河原達也．局所的な係り受けの情報を用いた話し言葉の節・文境界の推定．情報処理学会論文誌，Vol.50，No.2，pp.544--552，2009.

(9) 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.

(10) 浜辺良二，内元清貴，河原達也，井佐原均．話し言葉における引用節・挿入節の自動認定および係り受け解析への応用．自然言語処理，Vol.16，No.1，pp.3--23，2009.

(11) D.Cournapeau and T.Kawahara. Voice activity detection based on high order statistics and online EM algorithm. IEICE Trans., Vol.E91-D, No.12, pp.2854--2861, 2008.

(12) 南條浩輝，河原達也，七里崇．音声理解を指向したベイズリスク最小化枠組みに基づく音声認識．電子情報通信学会論文誌，Vol.J91-D，No.5，pp.1314--1324，2008.

(13) 翠輝久，河原達也，正司哲朗，美濃導彦．質問応答・情報推薦機能を備えた音声による情報案内システム．情報処理学会論文誌，Vol.48，No.12，pp.3602--3611，2007.

(14) 翠輝久，河原達也．ドメインとスタイルを考慮した web テキストの選択による 音声対話システム用言語モデルの構築．電子情報通信学会論文誌，Vol.J90-D，No.11，pp.3024--3032，2007.

〔学会発表〕（計 31 件）

(1) G.Neubig, Y.Akita, S.Mori, and T.Kawahara. Improved statistical models for SMT-based speaking style transformation. In Proc. IEEE-ICASSP, pp.5206--5209, 2010 年 3 月 米国・ダラス.

(2) R.Gomez and T.Kawahara. Optimizing spectral subtraction and Wiener filtering for robust speech recognition in reverberant and noisy conditions. In Proc. IEEE-ICASSP, pp.4566--4599, 2010 年 3 月 米国・ダラス.

(3) D.Cournapeau, S.Watanabe, A.Nakamura, and T.Kawahara. Using online model comparison in the Variational Bayes framework for online unsupervised voice activity detection. In Proc. IEEE-ICASSP, pp.4462--4465, 2010 年 3 月 米国・ダラス.

(4) T.Kawahara. New perspectives on spoken language understanding: Does machine need to fully understand speech? In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (invited talk), pp.46--50, 2009 年 12 月 イタリア・メ

ラノ.
(5) R.Gomez and T.Kawahara. Tight integration of dereverberation and automatic speech recognition. In Proc. APSIPA ASC, pp.639--643, 2009 年 10 月 札幌.
(6) A.Lee and T.Kawahara. Recent development of open-source speech recognition engine Julius. In Proc. APSIPA ASC, pp.131--137, 2009 年 10 月 札幌.
(7) G.Neubig, S.Mori, and T.Kawahara. A WFST-based log-linear framework for speaking-style transformation. In Proc. INTERSPEECH, pp.1495--1498, 2009 年 9 月 英国・ブライトン.
(8) R.Gomez and T.Kawahara. Optimization of dereverberation parameters based on likelihood of speech recognizer. In Proc. INTERSPEECH, pp.1223--1226, 2009 年 9 月 英国・ブライトン.
(9) K.Sumi, T.Kawahara, J.Ogata, and M.Goto. Acoustic event detection for spotting hot spots in podcasts. In Proc. INTERSPEECH, pp.1143--1146, 2009 年 9 月 英国・ブライトン.
(10) Y.Akita, M.Mimura, and T.Kawahara. Automatic transcription system for meetings of the Japanese. In Proc. INTERSPEECH, pp.84--87, 2009 年 9 月 英国・ブライトン.
(11) 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 年 4 月 台北.
(12) T.Sasada, S.Mori, and T.Kawahara. Extracting word-pronunciation pairs from comparable set of text and speech. In Proc. INTERSPEECH, pp.1821--1824, 2008 年 9 月 豪州・ブリスベーン.
(13) H.Wang and T.Kawahara. A Japanese CALL system based on dynamic question generation and error prediction for ASR. In Proc. INTERSPEECH, pp.1737--1740, 2008 年 9 月 豪州・ブリスベーン.
(14) T.Kawahara, M.Toyokura, T.Misu, and C.Hori. Detection of feeling through back-channels in spoken dialogue. In Proc. INTERSPEECH, p. 1696, 2008 年 9 月 豪州・ブリスベーン.
(15) T.Kawahara, H.Setoguchi, K.Takanashi, K.Ishizuka, and S.Araki. Multi-modal recording, analysis and indexing of poster sessions. In Proc. INTERSPEECH, pp.1622--1625, 2008 年 9 月 豪州・ブリスベーン.
(16) K.Ishizuka, S.Araki, and T.Kawahara. Statistical speech activity detection based on spatial power distribution for analyses of poster presentations. In Proc. INTERSPEECH, pp.99--102, 2008 年 9 月 豪州・ブリスベーン.
(17) T.Misu and T.Kawahara. Bayes risk-based dialogue management for document retrieval system with speech interface. In Proc. COLING, Vol.Posters & Demo., pp.59--62, 2008 年 8 月 英国・マンチェスター.
(18) H.Wang and T.Kawahara. Effective error prediction using decision tree for ASR grammar network in CALL system. In Proc. IEEE-ICASSP, pp.5069--5072, 2008 年 3 月 米国・ラスベガス.
(19) T.Kawahara, Y.Nemoto, and Y.Akita. Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In Proc. IEEE-ICASSP, pp.4929--4932, 2008 年 3 月 米国・ラスベガス.
(20) D.Cournapeau and T.Kawahara. Using Variational Bayes Free Energy for unsupervised voice activity detection. In Proc. IEEE-ICASSP, pp.4429--4432, 2008 年 3 月 米国・ラスベガス.
(21) T.Shinozaki and T.Kawahara. GMM and HMM training by aggregated EM algorithm with increased ensemble sizes for robust parameter estimation. In Proc. IEEE-ICASSP, pp.4405--4408, 2008 年 3 月 米国・ラスベガス.
(22) T.Shinozaki and T.Kawahara. HMM training based on CV-EM and CV Gaussian mixture optimization. In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.318--322, 2007 年 12 月 京都.
(23) D.Cournapeau and T.Kawahara. Evaluation of real-time voice activity detection based on high order statistics. In Proc. INTERSPEECH, pp.2945--2948, 2007 年 9 月 ベルギー・ブリュッセル.
(24) T.Misu and T.Kawahara. Bayes risk-based optimization of dialogue management for document retrieval system with speech interface. In Proc. INTERSPEECH, pp.2705--2708, 2007 年 9

月 ベルギー・ブリュッセル.

(25) C.Waple, H.Wang, T.Kawahara Y.Tsubota, and M.Dantsuji. Evaluating and optimizing Japanese tutor system featuring dynamic question generation and interactive guidance. In Proc. INTERSPEECH, pp.2177--2180, 2007年9月 ベルギー・ブリュッセル.

(26) T.Shinozaki and T.Kawahara. Gaussian mixture optimization for HMM based on efficient cross-validation. In Proc. INTERSPEECH, pp.2061--2064, 2007年9月 ベルギー・ブリュッセル.

(27) Y.Akita, Y.Nemoto, and T.Kawahara. PLSA-based topic detection in meetings for adaptation of lexicon and language model. In Proc. INTERSPEECH, pp.602--605, 2007年9月 ベルギー・ブリュッセル.

(28) T.Misu and T.Kawahara. An interactive framework for document retrieval and presentation with question-answering function in restricted domain. In Proc. IEA/AIE, pp. 126--134, 2007年6月 京都.

(29) T.Misu and T.Kawahara. Speech-based interactive information guidance system using question-answering technique. In Proc. IEEE-ICASSP, Vol.4, pp.145--148, 2007年4月 米国・ホノルル.

(30) T.Kawahara, M.Saikou, and K.Takanashi. Automatic detection of sentence and clause units using local syntactic dependency. In Proc. IEEE-ICASSP, Vol.4, pp.125--128, 2007年4月 米国・ホノルル.

(31) Y.Akita and T.Kawahara. Topic-independent speaking-style transformation of language model for spontaneous speech recognition. In Proc. IEEE-ICASSP, Vol.4, pp.33--36, 2007年4月 米国・ホノルル.

〔図書〕(計 1件)

(1) S.Furui and T.Kawahara. Transcription and distillation of spontaneous speech. In J.Benesty, M.M.Sondhi, and Y.Huang, editors, Springer Handbook on Speech Processing and Speech Communication, pp.627--651. Springer, 2008.

〔産業財産権〕

○出願状況(計 1件)

名称：音響モデル学習装置、音声認識装置、及び音響モデル学習のためのコンピュータプログラム
発明者：三村正人, 河原達也
権利者：京都大学
種類：特許
番号：特願2009-094212
出願年月日：2009年4月8日
国内外の別：国内

○取得状況(計 0件)

〔その他〕

なし

６．研究組織

(1) 研究代表者

河原 達也 (KAWAHARA TATSUYA)
京都大学・学術情報メディアセンター・教授
研究者番号：00234104

(2) 研究分担者

中村 裕一 (NAKAMURA YUICHI)
京都大学・学術情報メディアセンター・教授
研究者番号：40227947
(２００７～２００８年度；２００９年度は連携研究者)

秋田 祐哉 (AKITA YUYA)
京都大学・学術情報メディアセンター・助教
研究者番号：90402742

内元 清貴 (UCHIMOTO KIYOTAKA)
情報通信研究機構・知識創成コミュニケーション研究センター・主任研究員
研究者番号：60358885
(２００７年度のみ；２００８～２００９年度は連携研究者)

森 信介 (MORI SHINSUKE)
京都大学・学術情報メディアセンター・准教授
研究者番号：90456773
(２００８～２００９年度)

(3) 連携研究者