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　２１　日現在

機関番号：　１３８０１
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５３３０１２８

研究課題名（和文）音声ドキュメント内の検索とフィードバックに基づく高度なインデキシング機能の実現

研究課題名（英文）Advanced indexing based on spoken document retrieval and its feedback

研究代表者
甲斐　充彦（KAI, ATSUHIKO）

静岡大学・工学部・准教授

研究者番号：６０２８３４９６

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：音声ドキュメント検索システムの開発を目的として、インデキシングや検索精度の改善に関わる要素技術の開発を進めた。音声コンテンツに含まれる多様な収録環境や話者の違いを考慮するために、近年のパターン認識分野で高い精度を示している深層ニューラルネットワークモデル（DNN）を用いて音声区間検出モデルや残響除去モデルを構築し、話者区間分類や音声認識の前処理として用いる方法を提案し、性能を改善した。また、音声ドキュメント検索の性能に大きく影響を与える自動音声認識システムの未知語に起因する検出漏れを軽減するため、DNNによる音声特徴量変換に基づく再照合手法を提案し、検索性能を改善した。

研究成果の概要（英文）：We investigated and developed elemental technologies for indexing and other related processes which are designed to permit efficient and sustainable development of spoken document retrieval systems. For dealing with a possible change in speech features regarding to the recording conditions and speakers, we proposed DNN-based voice activity detection (VAD) and dereverberation models as a frontend of speaker diarization and speech recognition systems and improved accuracy for those systems. Also, we proposed DNN-based feature transformation as a rescoring step of spoken term detection (STD) system for coping with out-of-vocabulary words and the STD performance has been significantly improved.

研究分野：　音声情報処理

キーワード：　音声ドキュメント検索　音声検索語検出　STD　音声クエリ　DNN　音声認識信頼度　スコア正規化

１．研究開始当初の背景

今日ではスマートフォンのような身近な情報機器を使って低コストに大量の動画や音声データを保存できる。しかしこのような大規模かつ多様な収録対象の音声データに対して発話内容による検索技術はまだ十分に実用化されていない。特に、発話内容の検索技術の鍵となり音声からテキストへの変換を実現する自動音声認識システムでは、収録音声に含まれる語彙や話し言葉の多様な言語現象への対応がまだ不十分といえる。その一方で、近年では音声ドキュメントを対象とした検索システムの研究が盛んになってきた。例えば国内では、国立情報学研究所（NII）を中心として毎年開催されている情報アクセス技術の評価のためのワークショップ（NTCIR)において音声ドキュメント検索タスク「SpokenDoc」が設定され、国内外の研究機関が参加している。そのサブタスクとして、世界最大規模の話し言葉コーパスを対象として、検索語（検索クエリ）の出現箇所を推定する音声検索語検出（Spoken Term Detection: STD）が定義され、我々の研究グループも参加してきた。このタスクではテキスト入力による検索キーワード（クエリ）を仮定して、音声ドキュメントを完全に言語情報のみでテキスト表現した場合の一致部分を正解と定義し、一致箇所検出を目的としている。しかし、これまでの多くの研究では収録音声から自動的に推定されたテキスト情報だけを手掛かりとして検索システムの仕組みを実現しており、自動音声認識で十分な精度が保障されない場合を考慮した検索システムの枠組みについての検討はまだ不十分であった。

２．研究の目的

本研究では、音声データのまとまり（音声ドキュメント）に対する検索可能なコンテンツ化支援を目的として、インデキシングや検索の精度の改善に関わる要素技術を開発する。特に、様々な音声ドキュメントに対して基盤となる検索システムの要素技術の高精度化を中心として、音声ドキュメントから自動抽出する音声特有の付加的情報を自動書き起こし情報と併せてインデキシングや事例検索に利用する仕組みの開発と、インデキシングの精度を維持するため必要となる修正の候補提示の自動化に関する技術開発を行う。このシステムにより、少ないコストでも効果的にインデキシングの精度を高めたコンテンツ化が可能で、コストに応じて精度の高いインデキシングまで漸進的に適用可能なコンテンツ化支援技術の実現を目指す。

３．研究の方法

（１）音声収録時の話者やマイクロフォン、部屋などの収録環境の違いによって、自動音声認識システムの認識精度は大きく影響を受ける。特に雑音や残響による音声の変動は一般的な自動音声認識システムにおいて前段として抽出する音声特徴量に大きな影響を与える。また、長時間の収録音声には、複数の話者の音声が含まれることが多い。そこで、長時間の収録音声からこのような環境の違いや話者の混在に頑健な自動音声認識や検索を実現するために、残響下音声に頑健な音声区間検出（voice activity detection: VAD）や話者区間分類（話者ダイアライゼーション）のための環境の違いに頑健な話者特徴抽出に焦点を当てる。これらの問題に対して、本研究では近年のパターン認識関連の研究で大きな効果が実証されつつある深層学習の適用によって既存のシステム性能の改善を図る。具体的には、VAD のための特徴抽出と識別器設計に Deep Belief Network(DBN)の学習手法を適用し、Deep Neural Network（DNN）に基づく VAD システムを実装する。残響下音声に対して頑健な DNN 学習のために、複数の研究機関の会議室で遠隔マイクや接話マイクを同時に使って収録された会議音声コーパス（RT-05S コーパス）を利用する。また、話者区間分類のための特徴量の改善として、雑音を含む入力からクリーン音声信号を推定する方法として Denoising Autoencoder (DAE)を前処理として利用する。前述の会議音声コーパスにおいて遠隔マイクと接話マイクのペアで収録された音声を利用し、それらをそれぞれ DAE の入力と出力と対応させて雑音残響除去をする特徴量変換モデルを学習する。この結果を従来の話者区間分類（Speaker diarization）の前処理に適用し、分類精度の改善を図る。

（２）検索のためのインデックス信頼度推定やそれを利用した検索精度の改善のために、自動音声認識システムから得られる信頼度や発話速度などの発話固有の特徴を抽出し、話者固有の認識精度を推定するモデルの構築を行う。また、大語彙音声認識システムの単語グラフ（複数候補）出力情報を用いて検出区間の信頼度を推定し、検索候補の絞込みや順位付け、スコア正規化に利用する方法を新たに開発する。具体的には、これまでに我々が開発してきた音声検索語検出(STD)システムでは認識結果を音節単位やそれよりも小さな音節状態単位で音響モデルから計算される音声類似度によって類似区間の検出（スポッティング）をする方法を開発した。このシステムの処理段階の前処理または後処理として候補の絞込みやスコア統合によって改善を図る。また、音声ドキュメントや検索入力（テキストまたは音声によるクエリー）の音声特徴量や前処理から得られる音声認識結果以外の補助的な情報を、上述の信頼度評価と併せて検索スコア評価に組み入れて、全体のチューニングを可能とする仕組みを開発する。

４．研究成果

（１）残響下音声に頑健な音声区間検出（VAD）システムの構築のため、音声特徴抽出と識別器設計において Deep Belief Network(DBN)の学習手法を用いた DNN に基づく VAD システムを構築した。同様に DBN を使用した VAD の先行研究では人工的に残響特性を畳み込んだ人工残響音声を使用していたが、我々は実際に複数の会議室環境で収録された音声コーパスを利用したモデル構築を行った。更に、未知の収録環境に対する適応化によってより高い精度を得るため、DNN による VAD を他の VAD モデル(GMM や SVM)との組み合わせによって教師なし適応を行うクロス適応の方法を提案した。提案した方法の学習・評価のために、遠隔マイクと近接マイクで同時収録されている会議音声のコーパス（RT-05S）を利用し、学習データの含まれない会議室の遠隔マイク収録の音声データを評価用として利用した。複数の VAD システムで、フレーム単位での音声と非音声の識別誤りを比較評価した結果を図 1 に示す。DNN 単独で環境適応なしの場合（下図の DNN0-base0 システム）と比べて、GMM による VAD モデルを介在したクロス適応を適用した提案システム（DNN3-cross6 および DNN2-cross3b、両者は教師なし適応の適応回数が 3 または 6 回）では、識別誤りが大きい会議室において大きな改善が得られた（学会発表⑤、⑬）。

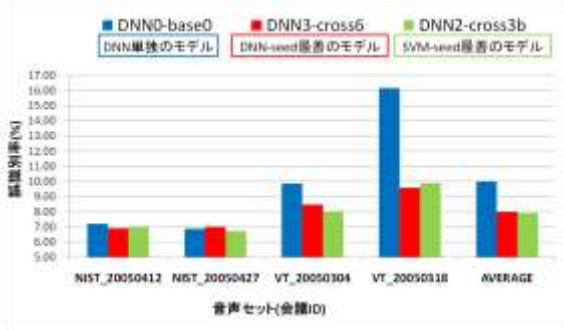

図 1. 音声区間検出性能（誤り率）の比較

また、話者区間分類のための DAE による音声特徴量変換においては、前述と同じ複数の会議室環境で収録された遠隔マイクと近接マイクの対応データを学習データとして利用した。遠隔マイク音声を DAE の入力として、近接マイク音声を DAE 出力の教師信号として利用することで、雑音残響除去を意図した特徴量変換モデルを DAE として構築した。その DAE による特徴量変換モデルを、図 2 に示すような話者ダイアライゼーションシステムにおける音声セグメントの話者クラスタリングの前段として適用する方法を提案した。

前述の RT-05S コーパスを使った話者ダイアライゼーションの評価実験結果を図 3 に示す。5 種類の会議室（Set1～Set5）の会議室のデータを用いて DAE を学習し、会議室毎の学習データとは別の日の会議音声データを評価用データとした用いた。提案方法（図の“DAE-frontend-based system”）は、全ての会議室環境に対して話者ダイアライゼーションの誤り率（DER）が改善された。

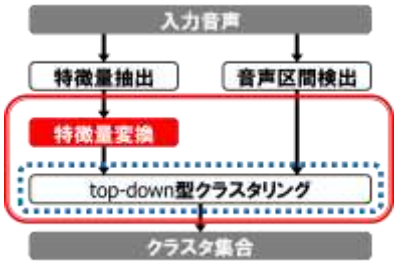

図 2. 雑音残響除去モデル（DAE）による特徴量変換を含む話者区間分類（ダイアライゼーション）処理の構成

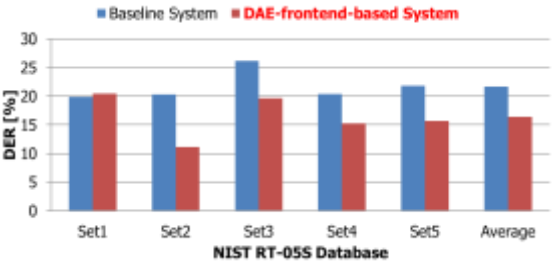

図 3. 話者区間分類（ダイアライゼーション）の性能比較

（２）自動音声認識システムから得られる信頼度や発話速度などの発話固有の特徴を抽出し、話者固有の認識精度を推定するモデルの構築を行った。具体的には、約 130 名の音声データを用いて実際に連続音節認識を行い、音節単位での正誤クラスを予測するモデルをロジスティック回帰モデルとして構築し、話者単位での認識誤り率を推定するモデルを構築する方法を提案した。そして、約 130 名の音声データの一部の話者の音声を学習データから取り除いて、その話者の 10 単語の音声サンプルから認識精度を予測する評価実験を行った。その結果、認識精度が話者間で最大 50%の違いがある状況において、話者単位で予測した精度と実際の精度との差が 5%以内に収まる割合が約 70%となる結果が得られた。

また、大語彙音声認識システムの単語グラフ（複数候補）出力情報を用いて検出区間の信頼度を推定し、検索候補の絞込みや順位付け、スコア正規化に利用する方法を新たに開発した。特に、検索対象の音声ドキュメントに含まれる未知語（前処理に用いる自動音声認識ステムの単語辞書に含まれない単語）による認識誤りの影響を軽減するため、認識結果だけを利用するのではなく音声特徴量の照合を併用する方法を開発した。信頼度を併用する方法としては、単語グラフに含まれる音素レベルの事後確率をもとに、クエリ中に含まれる音素 n-gram の事後確率を組み合わせて照合スコアとする方法を用いて、従来の

サブワード単位の音響的類似度によるスポッティングと併用する方法を提案した（雑誌論文③）。

更に、2014～2015 年度にかけては NTCIR の SpokenQuery&Doc のサブタスクと同様に、音声クエリによる STD システム性能の改善を図った。このタスクは、本研究でドキュメント内の類似箇所の頑健な検出を実現するための要素技術として改善を試みた。具体的には、前記（１）の研究において大きな効果を示した DNN を、話者や環境の変動に影響を受けにくい音声特徴量への変換のモデルとして用いるため 2 通りの方法で構築し、併用する方法を提案した。一つ目の方法では、フレーム単位（前後 4 フレームを含む）の入力音声の MFCC 特徴量に対して音素（triphone）状態のクラスを出力ノードとする DNN 音響モデルを、42 個のノードからなるボトルネック層を含めて学習し、評価時にはボトルネック層の出力を変換された 42 次元の特徴量として用いた。2 つ目の方法では、同様にフレーム単位の MFCC 特徴量（前後 4 フレームを含む）の入力に対して、145 種類の音素（monophone）を出力ラベルとして DNN 音響モデルを学習し、評価時には出力ノードの値を変換された 145 次元の音声特徴量として用いた。そして、従来の STD システムが出力する複数の検出候補区間に対して、これらの音声特徴量を用いて時間伸縮照合（DTW）で照合スコアを求め、他の照合スコアや自動音声認識システムから得られる特徴量（例えばクエリーの長さ）と併せてロジスティック回帰モデルとして正解検出確率を推定する方法により、スコア正規化を行う方法を提案した。従来の検出方法との比較実験を行った結果、前段にて DNN 音響モデルによる自動音声認識結果を行い、音節状態単位の音響的な距離を GMM-HMM 音響モデルで算出するベースラインシステムに対して、2 パス目での DNN ベースの音声特徴量での照合スコアなどを統合した方法によって STD 精度（F 値や MAP 値）の大きな改善が得られた（学会発表①）。一例として、研究会のべ 7 回分で収録された音声ドキュメントを対象とした音声検索語検出(STD)タスクにおいて、未知語を含む音声クエリによる検索で約 72%の MAP 値を得た。

（３）雑音残響下で収録された音声において、深層学習による残響除去モデルや音響モデルの有効性が極めて高いことがこの数年間に関連研究や我々の前述の研究進捗によって明らかになってきた。一方、我々の先行研究では単一マイクロフォンの入力信号からの残響除去法を提案しており、そこで用いていた残響成分の推定手法である Multi-step Linear Prediction (MSLP)法を前記（２）で提案した DNN による残響除去法と併用する方法が有効と考えた。そこで、この MSLP の残響推定の結果を前述の深層学習による残響除去モデル(DAE)に補助的な入力として加えて残響除去後の自動音声認識精度を改善する方法を提案した。複数の残響環境を想定した音声データを含む REVERB challenge の評価タスクにおいて従来の残響除去法と比較実験を行った結果、大語彙音声認識システムの性能において約 9～12%の単語誤り率の削減が得られた（雑誌論文②）。

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　６件）

① Bo Ren, Longbiao Wang, Liang Lu, Yuma Ueda, and Atsuhiko Kai, “Combination of bottleneck feature extraction and dereverberation for distant-talking speech recognition”, Multimedia Tools and Applications, 査読あり, Volume 75, Issue 9, pp 5093-5108, May 2016.
(doi: 10.1007/s11042-015-2849-1)

② Yuma Ueda, Longbiao Wang, Atsuhiko Kai and Bo Ren, “Environment-dependent denoising autoencoder for distant-talking speech recognition”, EURASIP Journal on Advances in Signal Processing, 査読あり, vol.2015, no.1, pp.1-11, November 2015.
(doi: 10.1186/s13634-015-0278-y)

③ Mitsuaki Makino, Naoki Yamamoto, Atsuhiko Kai, “Utilizing State-level Distance Vector Representation for Improved Spoken Term Detection by Text and Spoken Queries”, Proc. INTERSPEECH 2014, 査読あり, pp.1732-1736, September 2014.

④ Zhaofeng Zhang, Longbiao Wang and Atsuhiko Kai, “Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation”, EURASIP Journal on Audio, Speech, and Music Processing, 査読あり, 2014:15, April 2014.
(doi:10.1186/1687-4722-2014-15)

⑤ Takanori Yamada, Longbiao Wang and Atsuhiko Kai, “Improvement of distant-talking speaker identification using bottleneck features of DNN”, Proc. of INTERSPEECH 2013, 査読あり, pp.3661-3664, August 2013.

⑥ Naoki Yamamoto, Atsuhiko Kai, “Using Acoustic Dissimilarity Measures Based on State-Level Distance Vector Representation for Improved Spoken Term Detection”, Proc. of APSIPA Annual Summit and Conference 2013, 査読あり, October 2013.

http://dx.doi.org/10.1109/APSIPA.2013.6694151

〔学会発表〕（計１４件）

① Shuji Oishi, Tatsuya Matsuba, Mitsuaki Makino, Atsuhiko Kai, “Combining State-level and DNN-based Acoustic Matches for Efficient Spoken Term Detection in NTCIR-12 SpokenQuery&Doc-2 Task”, Proc. of the 12th NTCIR Conference on Evaluation of Information Access Technologies, 2016/6/10, 学術総合センター（東京都千代田区）

② 上田 雄磨, 王 龍標, 甲斐 充彦, Cepstral domain denoising autoencoder および DNN-HMM による雑音･残響下音声認識, 日本音響学会 2015 年春季研究発表会講演論文集, 2-1-2, 2015/3/17, 中央大学後楽園キャンパス（東京都文京区）

③ Bo Ren, Longbiao Wang and Atsuhiko Kai, “Speech selection and environmental adaptation for asynchronous speech recording based on deep neural network”, 電子情報通信学会音声研究会（第 16 回音声言語シンポジウム）, SP2014-121, 6 pages, 2014/12/16, 東京工業大学すずかけ台キャンパス（神奈川県横浜市）

④ 張 兆峰、王 龍標、甲斐充彦、李 衛鋒、岩橋政宏、DNN に基づく特徴変換による残響環境話者認識, 電子情報通信学会音声研究会（第 16 回音声言語シンポジウム）, SP2014-119, 6 pages, 2014/12/16, 東京工業大学すずかけ台キャンパス（神奈川県横浜市）

⑤ 中谷彰宏、王 龍標、甲斐充彦、会議音声における音声区間検出のための Deep Neural Network とクロス適応の検討, 電子情報通信学会音声研究会（第 16 回音声言語シンポジウム）, SP2014-107, 6 pages, 2014/12/15, 東京工業大学すずかけ台キャンパス（神奈川県横浜市）

⑥ Longbiao Wang, Bo Ren, Yuma Ueda, Atsuhiko Kai, Shunta Teraoka and Taku Fukushima, “Denoising autoencoder and environment adaptation for distant-talking speech recognition with asynchronous speech recording”, Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1-5 (5 pages), 2014/12/12, Cambodia.

⑦ Mitsuaki Makino and Atsuhiko Kai, “Combining Subword and State-level Dissimilarity Measures for Improved Spoken Term Detection”, Proc. of the 11th NTCIR Conference on Evaluation of Information Access Technologies, pp. 413-418, 2014/12/12, 学術総合センター（東京都千代田区）

⑧ Yuma Ueda, Longbiao Wang, Atsuhiko Kai, Xiong Xiao, EngSiong Chng and Haizhou Li, “Single-channel dereverberation for distant-talking speech recognition by combining denoising autoencoder and temporal structure normalization”, Proc. ISCSLP 2014, pp. 379-383, 2014/9/12, Singapore.

⑨ Ikuya Hirano, Kong Aik Lee, Zhaofeng Zhang, Longbiao Wang and Atsuhiko Kai, “Single-sided Approach to Discriminative PLDA Training for Text-Independent Speaker Verification without Using Expanded I-vector”, Proc. ISCSLP 2014, pp. 59-63, 2014/9/12, Singapore.

⑩ 寺岡俊汰、上田雄磨、王 龍標、甲斐充彦、福島 拓、非同期音声収録を用いた遠隔発話音声認識, 電子情報通信学会音声研究会（音学シンポジウム 2014）, SP2014-16, pp. 153-157, 2014/5/24, 日本大学文理学部キャンパス（東京都世田谷区）

⑪ 牧野光晃、山本直樹、甲斐充彦、分布間距離ベクトル表現による音響的類似度を利用したテキスト及び音声クエリからの音声検索語検出の改善, 第 8 回音声ドキュメント処理ワークショップ, 2014/3/15, 豊橋市民センター（愛知県豊橋市）.

⑫ 山本 直樹、甲斐 充彦、分布間距離ベクトルに基づく音響的類似度とサブワード事後確率の併用による音声検索語検出の改善, 情報処理学会音声言語情報処理研究会, Vol. 2013-SLP-99, 2013/12/19, 筑波大学文京キャンパス（東京都文京区）.

⑬ 中谷 彰宏、王 龍標、甲斐 充彦、“雑音に頑健な音声区間検出のための Deep Belief Network の適用, 日本音響学会 2013 年秋季研究発表会, 2013/9/25, 豊橋技術科学大学（愛知県豊橋市）

⑭ Naoki Yamamoto and Atsuhiko Kai: Spoken Term Detection Using Distance-Vector based Dissimilarity Measures and Its Evaluation on the NTCIR-10 SpokenDoc-2 Task, Proc. of the 10th NTCIR Conference, pp. 648-653, 2013/6/20, 学術総合センター（東京都千代田区）.

６．研究組織

(1)研究代表者

甲斐 充彦（KAI, Atsuhiko）

静岡大学・工学部・准教授

研究者番号： 60283496

(2)研究分担者

王 龍標（WANG, Longbiao）

長岡技術科学大学・技学研究院・准教授

研究者番号： 30510458

(3)連携研究者
小暮　悟 (KOGURE, Satoru)
静岡大学・情報学部・講師
研究者番号： 40359758