様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２６　日現在

機関番号：　２１２０１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００２４１

研究課題名（和文）DNNを用いた音声による音声の検索の高精度・高速・低資源システムの実現

研究課題名（英文）Spoken term detection system with high retrieval accuracy, high speed and small resources using Deep Neural Network

研究代表者

伊藤　慶明（Yoshiaki, Itoh）

岩手県立大学・ソフトウェア情報学部・教授

研究者番号：９０３２５９２８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では、ビデオや音声データ中の音声に対して、検索したい言葉（検索語）を高精度・高速・低資源で検索を実現するシステムを目指すものである。本研究では、深層学習であるDNN(Deep Neural Network)を導入し、まず第一段階で従来手法を用いて有力な候補を抽出し、第二段階でその少数の有力候補に対してのみDNNを用いた詳細照合を行う方式を研究開発し、高精度化を実現しつつ計算時間の削減を果たした。さらに事前に音節バイグラムすべてを検索しておく方式を開発し、高速化・低資源化を実現した。また、検索語が音声与えられ際、テキストで与えられた場合同様に、高精度、高速、低資源の検索システムを実現した。

研究成果の概要（英文）：This research aims the realization of high retrieval accuracy, speed up and small resources for spoken term detection among video data or voice data. The research introduced deep learning so called DNN (Deep Neural Network). The developed method utilizes the conventional retrieval method for spoken term detection and extracts candidates in the first step. It realized the high retrieval accuracy and speed up by performing detailed matching between a query and the small number of extracted candidates in the second step. Furthermore, we realized the speed up and small resources by the method of pre-retrieval for all syllable bigrams.When a spoken query is given, we developed the spoken term detection system that realized high retrieval accuracy, speed up and small resources.

研究分野：　音声言語処理

キーワード：　音声言語処理　音声検索

１．研究開始当初の背景

ビデオ機器の大容量化に伴い、ここ一週間あるいは一ヶ月で放送された全ての番組の中から、見たい番組や見たい部分だけを鑑賞できれば，テレビの番組予約などしないですみ、テレビ鑑賞が新しいスタイルに変わっていくと予想できる。その際、録画した長大なビデオを簡便に検索できる機能が望まれる。

携帯電話のようにテレビのリモコンで検索キーワードを入力せざるを得ないのが現状であるが、検索キーワードを音声で入力できれば非常に簡便となる。一週間の電子的なテレビプログラムを音声で検索する従来の「音声でテキストを検索する」だけでなく、一週間のビデオ中の音声の中から音声で検索する「音声で音声データを検索する」新しい機能の実現が求められる。この機能を実現する上では、実際にビデオ機器に搭載することを考慮すると，検索精度が高いことは言うまでもなく、長大なデータの中から高速でかつメモリなどの資源を要さないで検索できる機能を実現する必要がある。

この「音声でデータを検索する」技術が確立できれば音声データを膨大にコントロールする業務、例えばコールセンターにおける法令違反の返答の検索等に適用できる他、セキュリティ/治安面から長大な電話データの検索・分析等の応用も期待できる。

２．研究の目的

ビデオ機器の大容量化に伴い、前述の通り、所望の区間のみを鑑賞するライフスタイルに変わっていくと予想している。その際、ビデオを簡便に検索できる機能が望まれる。例えば一週間分のテレビプログラムを録画した中から、ビデオ中の音声を「音声で」検索する機能を、高精度・高速・低資源で実現するシステムの開発を目指す。

これまで音声でテキストを検索する技術の研究は多くなされてきたが、音声で音声データを検索する精度は低くその技術は確立されていない。また、高速にすれば精度低下や必要資源の増加に繋がり、精度・速度・資源のいずれかが損なわれてしまった。音声認識では近年 Deep Neural Network（DNN）により認識精度が飛躍的に向上しており、本研究では大型な DNN を導入しながら、高精度・高速・低資源で音声による音声データ検索を可能にする技術を新たに開発する。

当研究課題の開始時、30 時間の音声データに対し、6～7 割程度の音声検索精度，数秒の検索時間であった。実用的な音声検索機能を目指す上では、音声検索技術の更なる高精度化・高速化、低資源化が求められる。また、より簡便な入力方式として音声によるクエリの入力が期待される。そこで本研究テーマでは「テキスト及び音声による音声検索」の高精度・高速・低資源システムの確立を目指した。

３．研究の方法

本研究テーマでは、平成 26 年度までの科研費テーマをさらに発展させるために、音声検索性能のさらなる高度化を図るもので、音声検索技術の(1)高精度化，(2)高速化，(3)低資源化，(4)未知語音声クエリの実現の 4 件のサブテーマで研究開発を推進し、最終的には未知語のテキスト・音声クエリに対し、(1)～(3)の狙いを両立する「音声で音声を検索する」システムを実現するものである。(1)では DNN による音声データの認識，DNN によるリランキング，複数のモデルの検索結果の統合等により，検索精度を向上させた。(2)では音節バイグラム等で事前に検索しておく高速化方式を研究開発する。(3)では事前検索結果をコンパクトにインデックス化する方式を、(4)では未知語音声クエリに対して、DNN による音素列の抽出方式，様々な認識器の認識結果の統合，WEB 上の知識の利用により高精度な音声クエリ検索方式を研究開発した。

本研究では、上記の４件のサブテーマについて、以下に示す体制で研究を推進した。(1)における DNN による音声認識および候補区間のリランキング方式について研究実績のある伊藤慶明(岩手県立大) を、複数モデルの利用による精度向上は，複数の音響モデルを用いた研究を進めている李時旭博士（産総研）を主担当とした。(2) (3)については事前検索方式の研究を推進中でビデオの高速照合研究に実績のある伊藤慶明を主担当とした。(4)は実環境音声認識や音声データ検索を中心として実績のある李時旭博士を主担当とした。音声処理研究全般に経験がある研究協力者の田中和世教授(筑波大)は(1)(2)の副担当とし，各研究機関と緊密に連絡をとりながら研究開発を推進した。

４．研究成果

大量のビデオや大量の音声データ中からユーザが所望する区間を簡便に検索する機能が望まれている。本研究課題では、このようなニーズに対し、ビデオや音声データ中の音声に対して、検索したい単語や句（検索語）を「テキスト」あるいは「音声」で与え、高精度かつ高速かつ低資源で検索を実現するシステムを目指した。

検索語が音声認識システムの辞書に含まれていない「未知語」の場合には検索が困難である。検索語は人名、地名などの固有名詞

になることが多く、それらが未知語に該当することが多い。このため未知語の検索機能は必要不可欠である。本研究課題では未知語検索の高精度化、高速化、低資源化を目的として、まず高精度化を目指すために言語情報の利用と深層学習である DNN(Deep Neural Network)の導入を行い、2 つの高精度化を実現する新しい検索方式を研究開発した。(1)言語情報の利用では、高順位候補を利用し、高順位候補と同じドキュメント内の候補を有利にすることにより高精度化を実現した(雑誌論文②)。一方、DNN を用いると計算時間を要し、高速化する必要があるため、(2)2 段階検索方式を研究開発した。まず、第一段階で従来手法を用いて有力な候補を抽出し、第二段階でその少数の有力候補に対してのみ DNN を用いた詳細照合を行う方式を研究開発した(雑誌論文③,④)。これにより、計算時間の削減を実現した。さらに(3)事前に音節バイグラム（音節の２連続）をすべて検索しておく方式を開発し、高速化・低資源化を実現した(雑誌論文①)。

また、「検索語が音声」で与えられた際において、テキストで与えられた場合同様に、高精度、高速、低資源の検索システムの研究開発を行った。検索語が音声の場合、１秒間の特徴量系列は 100 フレームとなり、音声データの同様の特徴量系列との照合を行うと、データ量が大量となり、メモリ上には載らず、長時間の照合が必要であった。そこで(4)ビット列照合/スパースベクトル照合方式を導入し、メモリ上での検索方式を研究開発し検索の高速化を実現した（学会発表④)。さらに(5)音声クエリに対し、各フレームの約 3,000 次元の事後確率を最大の確率（最尤）の状態に置き換えることにより高速かつ低資源で照合を実現する方式を研究開発した（学会発表⑧）

以下、上記の成果(1)～(5)について詳しく述べる。

(1) 音声中の検索語検出における同文書内の高順位候補を利用したリスコアリング方式

あるクエリに対し、検索結果が得られる。得られた候補のうち、高順位候補は高い正解精度を示し信頼できる。またクエリは特定のドキュメントに複数回出現しやすいという性質を利用し、高順位候補が含まれるドキュメントには複数の正解が含まれると仮定し、そのドキュメント内の候補を有利にするという検索方式を研究開発し、検索精度の向上を実現した。

(2) 音声中の検索語検出における Deep Neural Network の出力確率を用いたリスコアリング手法

DNN でのフレームレベルで照合することにより高精度な照合が可能である一方、１秒間に 100 フレームを必要とするフレームレベルでの照合は時間を要する。このため、2 段階検索方式を研究開発した。従来手法で高速にクエリと音声データとの照合を行った後，検索結果上位候補を DNN の出力確率を用いてリスコアリングする手法を研究開発した．クエリの音素系列をフレーム系列に変換して照合する手法と，状態系列に変換して照合する手法の 2 種類を開発した．NTCIR-9，10 の Formal run，Dry run 計 4 種のテストセットを用いた評価実験の結果，検索精度を表す MAP（Mean Average Precision）が 5.80pt～11.55pt 向上し，処理時間はフレーム単位照合で約 0.18 秒，状態単位照合で約 0.10 秒となり、検索精度の向上と検索の高速化を実現した．

(3) 音声中の未知語の検索語検出における音節バイグラムのインデックス化方式

音素や音節などのサブワード認識結果を利用し、未知語の検索を行う音声中の検索語検出システムにおいて，検索時間，検索精度，インデックスの空間計算量の観点から，類似音節バイグラムリストを用いた方式を研究開発した。音節バイグラム同士を直接比較することにより類似バイグラムリストを構築し、音節バイグラムすべてで事前全照合結果のシミュレートを実現した．オープンな評価セットを用いた検証実験の結果，日本語話し言葉コーパス(CSJ)の 600 時間以上の音声ドキュメントの検索に対して，空間計算量 200MB 以下，MAP は全照合と同レベルを 1 秒以内の検索時間で実現できることを示した．また，音節バイグラム間照合により構築した音節バイグラムリストでは，事前に音節バイグラムの全てで連続 DP 照合する必要がなく，CSJ とは異なる音声ドキュメントに適用した場合でも有効に動作することを確認し，その汎用性を検証した．

(4) ビット列照合/スパースベクトル照合方式

検索語が音声で与えられ際において、前述の通り、音声検索語・音声データとも１秒間の特徴量系列は 100 フレームとなり、高精度な検索を実現するために、DNN の出力である約 3000 次元の事後確率（１フレーム当たり）であるポステリオグラムを用いると、30 時間程度の比較的小規模の音声データにおいてもデータ量が大量となり、メモリ上には載らず、1 音声クエリとのポステリオグラムレベルの照合に 45 秒を要してしまう。そこで 3,000 次元の事後確率ベクトルをビット列化して照合する方式と、スパースベクトル化した上で照合する方式を研究開発し、メモリ上での検索方式を研究開発し検索の高速化を実現した。

(5) 音声クエリの最尤状態系列化方式

前述のポステリオグラムレベルでの照合は約 3000 次元のベクトルの内積計算を局所距離として各フレームで行う必要があり、この内積計算に多大の計算時間を要していた。そこで、音声クエリの各フレームにおいて 3,000 個の最大事後確率のうち最大の確率値を示す最尤の状態番号に事後確率ベクトル

を置き換え、音声クエリを最尤状態番号系列にする。音声データのあるフレームとの局所距離を、音声クエリの最尤状態番号の事後確率を参照するだけで、求められるようした。これにより、3,000 次元同士の内積計算をせずに局所距離を求めることで、大幅な検索時間削減を実現した。

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　４件）

① 伊藤慶明, 鳴海司朗, 大内一揮, 菅原翔太, 李時旭, “音声中の未知語の検索語検出における音節バイグラムのインデックス化方式, ”電子情報通信学会論文, D Vol. J99-D, No. 2, pp. 178-187 (2016-2), DOI: 10.14923/transinfj.2015JDP7049.
② 小嶋和徳, 紺野和磨, 田中和世, 李時旭, 伊藤慶明, “音声中の検索語検出における同文書内の高順位候補を利用したリスコアリング方式, ”電子情報通信学会論文, Vol.J100-D, No.1, pp.70-80 (2017-1), DOI: 10.14923/transinfj.2016JDP7067.
③ 紺野良太, 小嶋和徳, 李時旭, 伊藤慶明, “音声中の検索語検出における Deep Neural Network の出力確率を用いたリスコアリング手法の提案, ” DOI: 10.14923/transinfj.2016JDP7103, 電子情報通信学会論文, Vol.J100-D,No.5, pp.595-604 (2017-5), DOI: 10.14923/transinfj.2016JDP7103.
④ 紺野良太, 小嶋和徳, 李時旭, 伊藤慶明, “音声中の検索語検出における Deep Neural Network の出力確率を用いた音響距離構築方式, ”電子情報通信学会論文, Vol.J100-D,No.8,pp. 798-807 (2017-8), DOI: 10.14923/transinfj.2016JDP7122.

〔学会発表〕（計 28 件）

① Kazuki Oouchi, Ryota Kon'no, Takahiro Akyu, Kazuma Konno, Kazunori Kojima, Kazuyo Tanaka, Shi-wook Lee, Yoshiaki Itoh, “Evaluation of re-ranking by prioritizing highly ranked documents in spoken term detection,” INTERSPEECH, pp. 3675-3679, 2015-9, Dresden, Germany.
② Shi-wook Lee, Kazuyo Tanaka and Yoshiaki Itoh, “Combination of diverse subword units in spoken term detection,” INTERSPEECH, pp. 3685-3689, 2015-9, Dresden, Germany.
③ Ryota Konno, Kazunori Kojima, Lee Shi-Wook, Kazuyo Tanaka, Yoshiaki Itoh, “Rescoring by a Deep Neural Network for Spoken Term Detection,” 4 pages, Asia-Pacific Signal and Information Processing Association APSIPA, pp. 1207-1211, 2015-12, Hong Kong.
④ Masato Obara, Kazunori Kojima, Kazuyo Tanaka, Shi-wook Lee and Yoshiaki Itoh, “Rescoring by Combination of Posteriorgram Score and Subword-Matching Score for Use in Query-by-Example,” INTERSPEECH, pp. 1918-1922, 2016-9, San Francisco, U.S.A.
⑤ Yoshino Shimizu, Yoshiaki Itoh, “Score priority integration method using multiple search results for Spoken Term Detection,” 5th Joint Meeting Acoustical Society of America and Acoustical Society of Japan, 2016-12, Hawaii, U.S.A..
⑥ Shi-wook Lee, Kazuyo Tanaka and Yoshiaki Itoh, “Generating Complementary Acoustic Model Spaces in DNN-Based Sequence-to-Frame DTW Scheme for Out-of-Vocabulary Spoken Term Detection” INTERSPEECH, pp. 755-759, 2016-9, San Francisco, U.S.A.
⑦ Daisuke Kaneko, Kazunori Kojima, Kazuyo Tanaka, Shi-wook Lee, Yoshiaki Itoh, “Constructing Acoustic Distances between Subwords and States Obtained from a Deep Neural Network for Spoken Term Detection,” INTERSPEECH, pp. , 2017-9, Stockholm, Sweden.
⑧ Masato Obara, Kazunori Kojima, Shi-wook Lee and Yoshiaki Itoh, “Acceleration for Query-by-Example Using Posteriorgram of Deep Neural Network,” 4 pages, Asia-Pacific Signal and Information Processing Association APSIPA, pp. , 2017-12, Kuala Lumpur, Malaysia.
⑨ 紺野良太, 小嶋和徳, 李時旭, 田中和世, 伊藤慶明, “音声中の検索語検出におけるフレームレベル状態系列間照合方式”, 電子情報通信学会技術研究報告, IEICE-SP2015-37, pp.7-12 (2015-07).
⑩ 伊藤慶明, 紺野良太, 小原真人, 李時旭, 田中和世, “音声中の検索語検出の研究動向と DNN の導入事例”, 【招待講演】, 電子情報通信学会技術研究報告, IEICE-SP2016-24, pp.21-26 (2016-07).
⑪ 紺野良太, 李時旭, 田中和世, 小嶋和徳, 伊藤慶明, “音声中の検索語検出における音響距離構築方式の検討”, 電子情報通信学会技術研究報告, IEICE-SP2016-25, pp.27-32 (2016-07).
⑫ 紺野良太, 李時旭, 田中和世, 小嶋和徳, 伊藤慶明, “STD におけるフレームレベル状態系列間照合による検索精度向上”, 日本音響学会秋季研究発表会, 1-Q-21, 4 pages (2015-9).
⑬ 李時旭, 田中和世, 伊藤慶明, “確率分布間の距離近似と異種性に基づく音声検索語検出システムの統合*”, 日本音響学会秋季研究発表会, 1-Q-19, 4 pages (2015-9).
⑭ 橋本拓観, 小嶋和徳, 伊藤慶明, “運転中等に用いる音声対話継続システム”, 第 14 回情報科学技術フォーラム FIT, E-026

(2015-9).

⑮ 紺野良太，小嶋和徳，伊藤慶明，“DNN 分布間距離より構築したサブワード/状態間音響距離の STD への適用”，日本音響学会春季研究発表会，1-R-10，4 pages (2016-3).

⑯ 小原真人, 小嶋和徳, 伊藤慶明,“DNN 出力確率系列 Posteriorgram との併用による STD 検索精度の向上”，日本音響学会春季研究発表会, 1-R-11, 2 pages (2016-3).

⑰ 橋本拓観，小嶋和徳，伊藤慶明,“興味推定による自動車内音声対話情報提供システム”，情報処理学会第 78 回全国大会, 6Q-04 (2016-3).

⑱ 清水嘉乃，岩﨑瑛太郎，李時旭， 田中和世，小嶋和徳，伊藤慶明,“STD における複数検索結果のスコア優先統合方式”, 日本音響学会秋季研究発表会，2-Q-12，69-72 (2016-9).

⑲ 紺野良太, 李時旭，田中和世, 小嶋和徳, 伊藤慶明,“サブワード/状態/フレーム照合スコアの統合による SQ-STD 検索精度向上”，日本音響学会秋季研究発表会, 2-Q-13，73-76 (2016-9).

⑳ 李時旭， 田中和世，伊藤慶明,“音声検索語検出システムのスコアリングに関する実験的検討”, 日本音響学会春季研究発表会，2-P-17，181-182 (2017-3).

21 紺野良太, 小嶋和徳, 李時旭，田中和世, 伊藤慶明,“SQ-STD における DNN 及び CTC 導入方式の検討”，日本音響学会春季研究発表会，2-P-18，183-186 (2017-3).

22 関恒平，小嶋和徳，李時旭，伊藤慶明,“音声中の検索語検出における拗音及び長母音モデルの検討”, 日本音響学会春季研究発表会，2-P-19，187-188 (2017-3).

23 大内一揮, 小原真人, 小嶋和徳, 李時旭, 伊藤慶明,“音声中の検索語検出の上位候補に対する SVM を用いたリランキング”, 電子情報通信学会総合大会，ISS-SP-209, 209 (2017-3).

24 清水 嘉乃, 李 時旭, 小嶋 和徳, 伊藤 慶明，“音声中の検索語検出における Paragraph Vector を用いたリスコアリング手法,” 日本音響学会秋季研究発表会, 2-Q-9，pp.145-148 (2017-9).

25 小原 真人，小嶋 和徳，伊藤 慶明，田中和世，李 時旭，“音声中の検索語検出における深層学習を用いた検索時間削減方式,”日本音響学会春季研究発表会, 1-Q-8, pp.83-86 (2018-3).

26 丹治 遥，小嶋 和徳，李 時旭，南條 浩輝，伊藤 慶明，“音声中の検索語検出における最上位候補を含む講演及びその類似講演優先方式,” 日本音響学会春季研究発表会，2-Q-17，pp.185-186 (2018-3).

27 李 時旭，田中 和世，伊藤 慶明音声検索語検出の距離値における事後確率の統合,” 日本音響学会春季研究発表会, 2-Q-21，pp.197-198 (2018-3).

28 清水 嘉乃, 李 時旭, 小嶋 和徳, 伊藤 慶明，“音声中の検索語検出におけるドキュメント間類似度を利用したリスコアリング方式,” 情報処理学会第 80 回全国大会， 5Q-08，pp.2-393--394 (2018-3).

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://p-www.iwate-pu.ac.jp/~y-itoh

６．研究組織

(1)研究代表者
伊藤 慶明（Yoshiaki Itoh）
岩手県立大学・ソフトウェア情報学部・教授
研究者番号：90325928

(2)研究分担者
李 時旭 （Lee Shi-wook）
国立研究開発法人産業技術総合研究所・知能システム研究部門・主任研究員
研究者番号： 50415642

(3)連携研究者
小倉 加奈代（Ogura Kanayo）
岩手県立大学・ソフトウェア情報学部・講師
研究者番号： 10432139