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

令和　元 年　 ６ 月 ２４ 日現在

機関番号：１４３０１
研究種目：基盤研究(B)（一般）
研究期間：2016～2018
課題番号：１６Ｈ０２８４７

研究課題名（和文）半自律的な音声認識による講演・講義への字幕付与

研究課題名（英文）Automatic speech recognition based on semi-autonomous learning for captioning lectures

研究代表者
河原　達也（Kawahara, Tatsuya）

京都大学・情報学研究科・教授

研究者番号：００２３４１０４

交付決定額（研究期間全体）：（直接経費）　　12,500,000円

研究成果の概要（和文）：入力音声から単語系列に直接写像するEnd-to-Endの枠組みに基づく音声認識を提案し、従来の音声認識手法と比較して、処理時間を大幅に（1/30以下）に削減しながら、高い認識精度を実現できることを示した。また、講演・講義を対象として字幕を付与するシステム(http://caption.ist.i.kyoto-u.ac.jp/)を構築・試験運用した。さらに、パソコンでも動作する音声認識パッケージを構成し、聴覚障害者の情報保障のためにリアルタイムで字幕を付与するソフトIPtalk(http://www.s-kurita.net/)に統合して一般に公開した。

研究成果の学術的意義や社会的意義
障害者差別解消法の施行に伴い、講義や講演において聴覚障害者に対する情報保障、すなわち字幕付与が求められているが、現状では量と質の両方において十分でない。これを支援するための音声認識技術の研究開発を行った。新たな深層学習に基づくモデルを導入することで、認識精度と速度の両方で大きな改善が得られた。サーバベースで音声ファイルに字幕を付与するシステム(http://caption.ist.i.kyoto-u.ac.jp/)に加えて、パソコン要約筆記で一般的に用いられているIPtalkにも音声認識の組込みを行い、一般公開した。また、『聴覚障害者のための字幕付与技術』シンポジウムを開催した。

研究成果の概要（英文）：We have proposed a new end-to-end framework of speech recognition that directly converts speech signal to a word sequence. It is demonstrated to achieve higher accuracy with a drastically faster speed compared with the conventional systems. We have also developed a captioning system based on the server-based speech recognition system, and also a speech recognition package for PC which is integrated with the captioning software IPtalk widely used in Japan. The software is freely open to the public.

研究分野：音声情報処理

キーワード：音声認識　コンテンツ・アーカイブ　機械学習　字幕付与　情報保障

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

音声・映像コンテンツに対する字幕は、全国約６百万人の難聴者のみならず、さらに多数の高齢者や外国人にとって必要であり、特に講演・講義では不可欠である。テレビ放送では字幕付与された番組は着実に増えているものの、インターネットで配信されているコンテンツ（MOOC など）については皆無に近い。また、障害者差別解消法の施行に伴い、講義や講演の場において聴覚障害者に対する情報保障、すなわち字幕付与が求められているが、現状では量と質の両方において十分でない。そこで、字幕を効率的に付与するための音声認識技術が求められている。

音声認識技術は近年、スマートフォンの音声検索等に代表されるように大きな進展を遂げているが、機械を意識して、単純な文を明瞭に発声することが前提となっており、自然な話し言葉音声への対応は依然として難しい。

２．研究の目的

講演・講義への字幕付与を目的として、話し言葉音声認識の高度化を行う。長時間の話し言葉音声に対して、教師付き学習のための書き起こしデータの構築を大規模に行うのは限界があるので、音声データに対して完全な書き起こしがない条件、あるいはテキストデータのみがある条件で、半自律的にモデル学習を行う枠組みを考える。また、近年著しい進展を遂げている深層学習に基づく新たな枠組みを検討する。放送大学の講義や学会等における講演を対象として、この枠組みの実装・評価を行い、一般にも利用できるシステムの実現を目指す。

３．研究の方法

新たな音声認識のモデル・アルゴリズムを研究するとともに、放送大学の講義や学会の講演を対象とした字幕付与システムの開発を行う。

(1) End-to-End 音声認識の導入

ニューラルネットワークに基づいて音響モデルと言語モデルを一体的にモデル化し、入力音声から認識結果の単語列を直接求める End-to-End 音声認識を導入・実装し、従来の方法と比較・評価する。

(2) End-to-End 音声認識の適応

上記の音声認識システムを新しいドメインに適応したり、専門用語を追加する方法について検討する。

(3) 字幕付与システムの構築と評価

講演・講義の音声ファイルに字幕を付与するシステムを構築・試験運用し、評価を行う。

(4) リアルタイム情報保障の実証実験

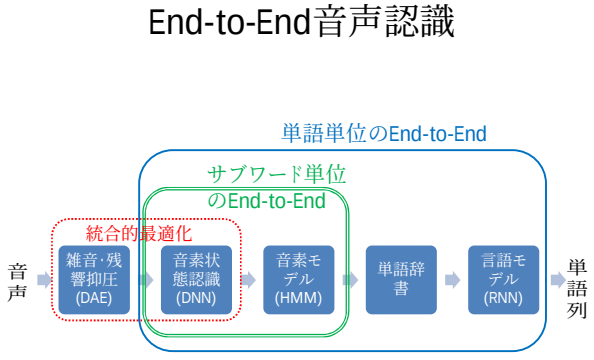
聴覚障害者の情報保障のためにリアルタイムで字幕を付与するシステムを構築し、情報処理学会の講演などで試験運用する。本ソフトウェアを一般の方が利用できるように公開するとともに、『聴覚障害者のための字幕付与技術』シンポジウムを開催する。

４．研究成果

(1) 単語を単位とした End-to-End 音声認識の実現

音声認識は、大規模データベース（ビッグデータ）の蓄積と深層学習（ディープラーニング）の導入により、大幅な性能改善が実現されている。ただし従来のシステムは、音響モデル・単語辞書・言語モデルなどを精緻にモデル化する必要があり、複雑な処理構成となっていた。これに対して、End-to-End の枠組みが検討されるようになった。音声認識では、音響特徴量系列から文字／単語列への系列間(sequence-to-sequence)写像学習として定式化される。しかし、これまでの End-to-End 音声認識に関する研究の大半は、音素や文字を単位として行われており、語彙の制約や単語単位の言語モデルは、後処理として適用する必要がある。また、認識精度の点において、従来のシステムを上回る結果はあまり得られていない。

これに対して、単語を単位とした End-to-End (Acoustic-to-Word) モデルを検討した。これは、語彙や言語モデルをすべて包含して、一気に音声から単語系列に変換するもので、ニューラルネットワークのみで音声認識処理が完結し、発音辞書や複雑なプログラムである音声認識デコーダを一切必要としない。これを右図に示す。



ただし、音素や文字に比べて、単語のエントリ数は圧倒的に多く、また出現頻度の偏りも大きいため、学習が容易でない。実際にかなり大規模な音声データベースがないと構

築が困難である。さらに、サブワード単位の認識と比較して、音声データベース中に出現しない未知語を認識できない（単語辞書に後で追加することもできない）という実用的に重要な問題もある。これらの問題の解決に取り組んだ。

まず、文字単位のモデルと併用することで解決を図った。単語単位のモデルについては、注意機構モデルの方が言語モデルを直接的に包含するので望ましい。この注意機構モデルの正則化の効果を期待して、文字単位のモデルは CTC を採用し、エンコーダ部分を共有する構成とした。その上で、両者のマルチタスク学習を行った。また、単語単位のモデルで未知語を検出した際には、文字単位のモデルの認識結果を用いて復元する。『日本語話し言葉コーパス』(CSJ)を用いて、約２万語彙のシステムを実装・評価したところ、学会講演・模擬講演ともに、DNN-HMM ハイブリッドモデルを認識率で上回ることができた。また、認識に要する時間の実時間比は約 0.03 で DNN-HMM に比べて 30 分の１以下となった。発話終了後でないと処理を開始できないが、この処理速度であれば問題ないと考えられる。

(2) End-to-End 音声認識の適応

上記の語彙の問題を含めて、End-to-End 音声認識は学習データベースに特化する傾向があり、しかも、音声と書き起こしのパラレルデータからしか学習できず、テキストのみのデータを活用できないという問題もある。新たなドメインの学習データを効率的に増強するために、音声合成を用いる方法を検討した。最新の音声合成は End-to-End モデルに基づいており、構築が比較的容易な上に、品質も高い。音声認識の学習データには音響特徴量で十分で、音声波形を生成する必要はない。単一話者の音声合成でも、エンコーダ転移学習と併用することにより、テキストデータのみから語彙の追加を含めたドメイン適応が可能になった。ただし音声認識では、多様な話者のパターンをモデル化する必要があるため、音声合成において話者情報を含めることで多数話者の音声を出力できるように拡張している。これにより、テキストデータのみから音声認識学習用のパラレルデータをいくらでも生成することが原理的に可能になる。

(3) 字幕付与システムの構築と公開

講演や講義などの音声コンテンツに対して音声認識と字幕付与を行うサーバを構築した。

http://caption.ist.i.kyoto-u.ac.jp/　（字幕付与サーバ）

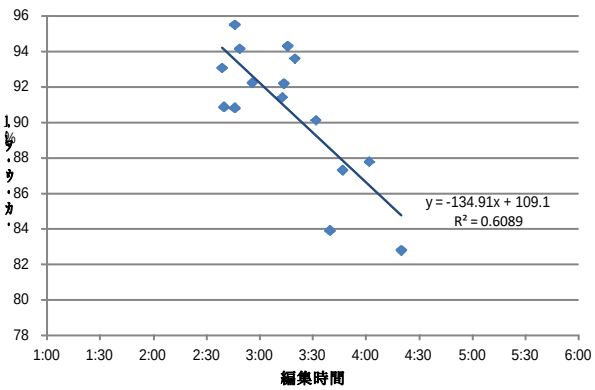
利用者は、音声ファイルや映像ファイルを当該サーバにアップロードし、所定の手続きをすると、音声認識による書き起こしにタイムスタンプが付与されたファイル（SAMI や SRT など複数のフォーマット）が生成される。これらは、一般的な再生ソフトで字幕ファイルとして利用可能である。音声認識には誤りが含まれる上に、話し言葉には言い淀みなども多いため、字幕として提示するには編集が必要である。また適当な位置での改行や句読点挿入も必要である。そのためのエディタも上記サイトで提供している。

現在、想定しているコンテンツは以下の３種類であり、各々について音声認識のモデルが用意されている。

- 講演：学会や講義など大教室で１人で行う学術講演（CSJ の学会講演データで構築）
- スピーチ：一般的な話題に関してゆっくり話すもの（CSJ の模擬講演データで構築）
- 討論：議会審議など公共の場で複数人で行う討論（国会審議のデータで構築）

(4) 放送大学の講義への字幕付与

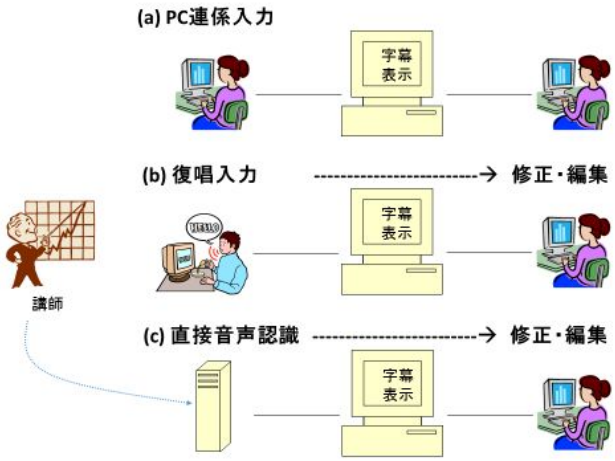
放送大学で配信されている講義では、受講生の要望に応じて一部に字幕が付与されているが、テレビ番組の半数程度にとどまっている。そこで、インターネットで配信されている講義に対して、音声認識を用いた字幕付与に取り組んだ。音声認識には、上記(3)の字幕付与サーバの「講演モデル」を用いるが、科目毎に教科書テキストを追加混合することで適応を行う。



平均の音声認識率（文字正解率）は約 90%となった。これを元に字幕の編集に要した時間と実時間比は約５程度となった。またいくつかの講義について、各回の認識率と編集時間の関係をプロットしたものを上図に示す。音声認識率と編集時間の間に高い（0.5～0.6）相関があることがわかる。以前放送大学で行った実験では、音声認識を用いずに放送大学の講義を書き起こした場合、実時間の平均 5.3 倍（約４時間）程度と報告されている。これをグラフ中の直線と重ねると、87%程度の認識率の場合に相当する。これから、音声認識率が 87%以上の場合にその効果があることが示唆される。また、93%になると 1/3 以上の時間短縮効果が示され、かなり優位性があるといえる。

(5) 講演のリアルタイム情報保障

学会やシンポジウムの講演に対してリアルタイムに字幕付与を行うシステムも構築・試験運用を行った。このような字幕では、一定の正確性とともに時間遅れも重要になる。情報保障の品質の観点からは、認識誤りの修正や不要な箇所の削除などを行う編集が必要となる。しかし、パソコン連係入力に対する効率性・優位性を目指して、１名で編集する枠組みを採用している（右図(c)参照)。



音声認識には、上記(3)の 字幕付与サーバの「講演モデル」を使用し、言語モデルと単語辞書は講演予稿やスライドのテキストを用いて話題への適応を行う。これらのテキストが利用可能となるのは一般的に講演開催の直前で、適応や調整の作業に時間的な余裕がないため、適応には単純なテキストの線形補間手法を用いている。

講師の音声は、会場の拡声機器（PA）から分配して編集端末（PC）に入力することを想定しているが、これが困難な場合は独自にマイクを設置して入力する。入力された音声は、会場の端末では音声認識を行わず、音声区間が切り出されるたびにそのデータを、上記(3)の字幕付与サーバにインターネット経由で送信する。サーバ側で音声認識を行い、その結果を逐次的に取得し、端末で編集と出力を行う。字幕の提示には、パソコン要約筆記で一般的に用いられているソフトウェア IPtalk を使用する。

(6) リアルタイム情報保障のための音声認識ソフトウエアの公開

上記(3)のシステムを一般の方が利用するのは困難であるので、パソコンでも動作する「話し言葉音声認識キット」及び「講演音声認識キット」を作成・公開した。

http://julius.osdn.jp/index.php?q=dictation-kit.html（音声認識キット）

また、IPtalk にも、これらのキットを使用して音声認識を行い、字幕を作成できる機能が搭載された。

http://www.s-kurita.net/ （IPtalk）

本プロジェクト及びこのソフトの紹介を兼ねて、『聴覚障害者のための字幕付与技術』シンポジウムを 2016 年と 2018 年に開催した。聴覚障害者や要約筆記者などを含めて、２回とも約 150 名の参加者があり、当該技術の展望について様々な意見交換を行った。

http://www.sap.ist.i.kyoto-u.ac.jp/jimaku/ （シンポジウムの詳細）

５．主な発表論文等

〔雑誌論文〕（計　１２件）

[1] K.Shimada, Y.Bando, M.Mimura, K.Itoyama, K.Yoshii, and T.Kawahara.
Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition.
IEEE/ACM Trans. Audio, Speech ¥& Language Process., Vol.27, p. accepted for publication, 2019.

[2] T.Zhao and T.Kawahara.
Joint dialog act segmentation and recognition in human conversations using attention to dialog context.
Computer Speech and Language, Vol.50, pp.108--127, 2019.

[3] K.Inoue, D.Lala, K.Takanashi, and T.Kawahara.
Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue.
APSIPA Trans. Signal ¥& Information Process., Vol.7, No.e9, pp.1--16, 2018.

[4] 山本賢太, 井上昂治, 中村静, 高梨克也, 河原達也.
人間型ロボットのキャラクタ表現のための対話の振る舞い制御モデル.
人工知能学会論文誌, Vol.33, No.5, pp.C--I37¥_1--9, 2018.

[5] M.Mirzaei, K.Meshgi, and T.Kawahara.
Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening.
Computer Speech and Language, Vol.49, pp.17--36, 2018.

[6] 井上昂治, Divesh Lala, 吉井和佳, 高梨克也, 河原達也.
潜在キャラクタモデルによる聞き手のふるまいに基づく対話エンゲージメントの推定.

人工知能学会論文誌, Vol.33, No.1, pp.DSH--F¥_1--12, 2018.

[7] R.Duan, T.Kawahara, M.Dantsuji, and J.Zhang.
Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning.
IEICE Trans., Vol.E100-D, No.9, pp.2174--2182, 2017.

[8] M.Mirzaei, K.Meshgi, Y.Akita, and T.Kawahara.
Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill.
ReCALL Journal, Vol.29, No.2, pp.178--199, 2017.

[9] S.Li, Y.Akita, and T.Kawahara.
Semi-supervised acoustic model training by discriminative data selection from multiple ASR systems' hypotheses.
IEEE/ACM Trans. Audio, Speech ¥& Language Process., Vol.24, No.9, pp.1524--1534, 2016.

[10] **河原達也**.
音声認識技術の変遷と最先端－深層学習による End-to-End **モデル－**.
日本音響学会誌, Vol.74, No.7, pp.381--386, 2018.

[11] **河原達也**, **荒木章子**.
会議録作成を支援するＩＣＴ.
電子情報通信学会誌, Vol.101, No.5, pp.486--491, 2018.

[12] **河原達也**, **秋田祐哉**.
聴覚障害者のための講演・講義の音声認識による字幕付与.
日本音響学会誌, Vol.74, No.3, pp.156--162, 2018.

〔学会発表〕（計　１６件）

[1] M.Mimura, S.Ueno, H.Inaguma, S.Sakai, and T.Kawahara.
Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition.
In Proc. IEEE Spoken Language Technology Workshop (SLT), pp. 477--484, 2018.

[2] H.Inaguma, M.Mimura, S.Sakai, and T.Kawahara.
Improving OOV detection and resolution with external language models in acoustic-to-word ASR.
In Proc. IEEE Spoken Language Technology Workshop (SLT), pp. 212--218, 2018.

[3] S.Ueno, T.Moriya, M.Mimura, S.Sakai, Y.Yamaguchi, Y.Aono, and T.Kawahara.
Encoder transfer for attention-based acoustic-to-word speech recognition.
In Proc. INTERSPEECH, pp.2424--2428, 2018.

[4] M.Mimura, S.Sakai, and T.Kawahara.
Forward-backward attention decoder.
In Proc. INTERSPEECH, pp.2232--2236, 2018.

[5] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara.
Acoustic-to-word attention-based model complemented with character-level CTC-based model.
In Proc. IEEE-ICASSP, pp.5804--5808, 2018.

[6] K.Shimada, Y.Bando, M.Mimura, K.Itoyama, K.Yoshii, and T.Kawahara.
Unsupervised beamforming based on multichannel nonnegative matrix factorization for noisy speech recognition.
In Proc. IEEE-ICASSP, pp.5734--5738, 2018.

[7] R.Duan, T.Kawahara, M.Dantsuji, and H.Nanjo.
Efficient learning of articulatory models based on multi-label training and label correction for pronunciation learning.
In Proc. IEEE-ICASSP, pp.6239--6243, 2018.

[8] H.Inaguma, M.Mimura, K.Inoue, K.Yoshii, and T.Kawahara.
An end-to-end approach to joint social signal detection and automatic speech recognition.
In Proc. IEEE-ICASSP, pp.6214--6218, 2018.

[9] M.Mimura, S.Sakai, and T.Kawahara.
Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks.
In Proc. IEEE Workshop Automatic Speech Recognition ¥& Understanding (ASRU), pp.134--140, 2017.

[10] T.Kawahara.
Automatic meeting transcription system for the Japanese Parliament (Diet).
In Proc. APSIPA ASC, (overview talk), 2017.

[11] T.Kawahara.
Modeling difficulties of second language learners using speech technology.
In Proc. Seoul International Conference on Speech Sciences (SICSS), p. 11 (keynote speech), 2017.
[12] M.Mimura, Y.Bando, K.Shimada, S.Sakai, K.Yoshii, and T.Kawahara.
Combined multi-channel NMF-based robust beamforming for noisy speech recognition.
In Proc. INTERSPEECH, pp.2451--2455, 2017.
[13] H.Inaguma, K.Inoue, M.Mimura, and T.Kawahara.
Social signal detection in spontaneous dialogue using bidirectional LSTM-CTC.
In Proc. INTERSPEECH, pp.1691--1695, 2017.
[14] R.Duan, T.Kawahara, M.Dantsuji, and J.Zhang.
Effective articulatory modeling for pronunciation error detection of L2 learner without non-native training data.
In Proc. IEEE-ICASSP, pp.5815--5819, 2017.
[15] S.Li, X.Lu, S.Sakai, M.Mimura, and T.Kawahara.
Semi-supervised ensemble DNN acoustic model training.
In Proc. IEEE-ICASSP, pp.5270--5274, 2017.
[16] R.Duan, T.Kawahara, M.Dantsuji, and J.Zhang.
Multi-lingual and multi-task DNN learning for articulatory error detection.
In Proc. APSIPA ASC, 2016.

〔図書〕（計　1件）
[1] 河原達也 編著.
音声認識システム （改訂 2 版）.
オーム社, 2016.

〔産業財産権〕
○出願状況（計　0件）
○取得状況（計　0件）

〔その他〕
ホームページ等
- 音声認識技術を用いた字幕付与支援プロジェクト
  http://www.sap.ist.i.kyoto-u.ac.jp/jimaku/
- 音声認識を用いた自動字幕作成システム
  http://caption.ist.i.kyoto-u.ac.jp/

６．研究組織

(1)研究分担者
研究分担者氏名：秋田　祐哉
ローマ字氏名：Akita, Yuya
所属研究機関名：京都大学
部局名：経済学研究科
職名：准教授
研究者番号（8 桁）：90402742

(2)研究協力者
研究協力者氏名：広瀬　洋子
ローマ字氏名：Hirose, Youko