Vol.2016-AAC-2 No.5
2016/12/2

自動音声認識を用いた 放送大学のオンライン授業に対する字幕付与

河原達也[†1]　秋田祐哉[†1]　広瀬洋子[†2]

概要：放送大学では様々な講義を配信しているが、アクセシビリティ向上のために字幕付与を進めている。京都大学では講演・講義を対象とした自動音声認識の研究開発を進めており、放送大学の講義に対しても音声認識を用いた字幕付与の実現可能性の検討を進めてきた。深層学習を用いた音声認識技術を用いることで、概ね 90%の認識率が得られ、その場合に人手で書き起こしを作成する場合に比べて有用性が確認された。本システムを用いて、今年度から主にオンライン授業を対象として大規模に字幕付与が行われている。

キーワード：字幕付与，音声認識，オンライン授業

1. はじめに

今年度から施行されている障害者差別解消法では、障害者の社会的障壁の除去について「必要かつ合理的な配慮」を行うことが義務づけられており、聴覚障害者に対しては情報保障を行うことがこれに該当する。情報保障には、手話通訳と要約筆記・字幕付与がある。後者は音声言語をテキスト化するもので、自動音声認識の適用が考えられる[1,2]。近年、音声認識技術は深層学習の導入などで大幅に性能が向上し、スマートフォン等での入力には実用的な水準に達しているが、自然な話し言葉の音声認識はまだ容易でなく、講演や講義の字幕付与のための研究開発が進められている[3,4]。

講演や講義への字幕付与の形態には、収録した映像に事後的（オフライン）に行う場合と、その場でリアルタイムに行う場合がある。後者はリアルタイム性が重視されるのに対して、前者では高い正確性が求められる。人手による修正を仮定しても、かなり高い音声認識精度でないと実用的でない。

本稿では、収録した講演・講義の映像に対する字幕付与について述べる。現在テレビ番組の大半では字幕が付与されるようになったが、ネット配信のコンテンツでは字幕はほとんど付与されていない。講演や講義のネット配信も一般的になってきたが、MOOC のように高コストで作成される短時間のコンテンツを除くと、字幕はほぼ皆無である。

放送大学は、我が国で最大のオンライン教育を行っている教育機関であり、約 300 の科目の講義がテレビもしくはラジオで配信されている。その大半がインターネットでも配信され、スマートフォンやタブレットなどでも視聴できるようになっている。現在、字幕が付与されているのはテレビ番組の半数程度であるが、近い将来 100%の字幕付与を目指している。今年度からは、すべての学習をインターネット上の講義や課題解答で行う「オンライン授業」も開設されており、これには原則字幕付与する方針である。

著者らは、音声認識技術を用いて、放送大学の講義の字幕付与を効率的に行う方法について検討を進めてきた。このような試みは以前にも行われたが[5,6]、十分な認識精度が得られていない。今回、深層学習などの最新の技術と教科書テキストを用いた適応を行うことで、90%程度の認識率を実現する。その上で、音声認識結果を編集することで、字幕テキストを効率的に作成できることを検証する。実際に、本システムを用いた字幕付与について報告する。

2. 音声認識・字幕生成サーバ

京都大学では、講演や国会審議[7]を対象とした音声認識の研究を進めてきたが、そのような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（URL は以下）を構築している[8]。

http://caption.ist.i.kyoto-u.ac.jp/

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。

ただし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（図 1 参照）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）

†1 京都大学
Kyoto University
†2 放送大学
Open University of Japan

音声認識システムはすべて、Julius と DNN-HMM 音響モデル（系列識別学習）及び単語 3-gram 言語モデルで構築している（学習データベースは上記参照）[3]。

また、コンテンツに関連するテキスト（予稿やスライドなど）を同時にアップロードすることで、音声認識の言語モデルを適応することも可能である。

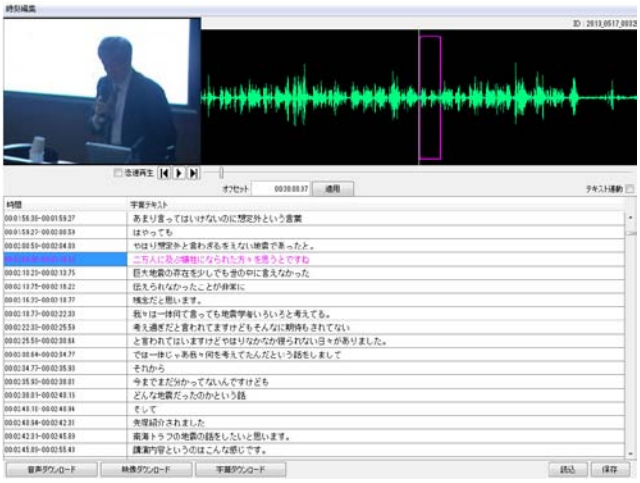

図 1　字幕編集用エディタ

3. 放送大学講義の音声認識

放送大学の講義は、基本的にスタジオで収録されていて、録音条件はよい。すべてに台本が用意されているわけではないが、一般の講演や講義に比べると、はるかに発声は明瞭である。ただし一部に、対談形式の部分やインタビューのビデオが挿入されている区間もある。言語的には、講義特有の専門用語が数多くあるのが問題となるが、教科書テキストが入手可能である。

音声認識の音響モデルは、『日本語話し言葉コーパス』(CSJ)の学会講演約 260 時間を用いて学習した DNN-HMM である（前節の講演カテゴリに相当）。放送大学の講義音声約 60 時間を用いて、半教師つき追加学習も試みた。

言語モデルは、同じく CSJ を用いて構築されているが、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。これにより専門用語の追加も自動的に行われるが、正しい読みが付与されているか確認が必要である。

次節に示すように、概ね 90%の認識率が得られている。

4. 音声認識を用いた字幕テキスト作成実験

いくつかの講義を対象に、音声認識を行ったものを編集することで字幕テキストを作成する実験を行った。これは、最終的に放送大学で字幕として配信するものでなく、テキストとして誤りがなく、適度に同期されて改行がされているレベルのものである。本実験に用いた講義は、2015 年度及び 2016 年度に開講された以下のラジオ講義である。

(1)　心理臨床の基礎
(2)　リスク社会のライフデザイン
(3)　ＣＧと画像処理の基礎

講義はいずれも各回 45 分で 15 回あるが、(2)については台本のある 3 回分を除いた。また、(3)については本実験では 7 回分のみを用いるが、これらには台本がある。

平均の音声認識率を表 1 に示す。これは、最終的に生成された字幕テキストに対する文字単位の正解率である。この編集に要した時間と実時間比も表 1 に示す。この編集作業は、１名の作業者が図 1 のエディタを用いて行ったものである。この後、実時間と同程度の確認作業を行っている。また講義(1)(2)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図 2・3 に示す。

表 1　講義の音声認識率と編集時間

	講義数	認識率	編集時間	実時間比
(1)	15	90.8%	3 時間 16 分	4.4
(2)	12	88.5%	3 時間 46 分	5.0
(3)	7	94.4%	2 時間 52 分	3.8

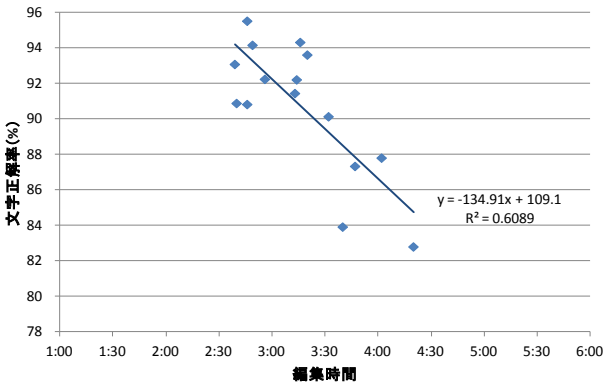

図 2　音声認識率と編集時間の相関（「心理臨床の基礎」）

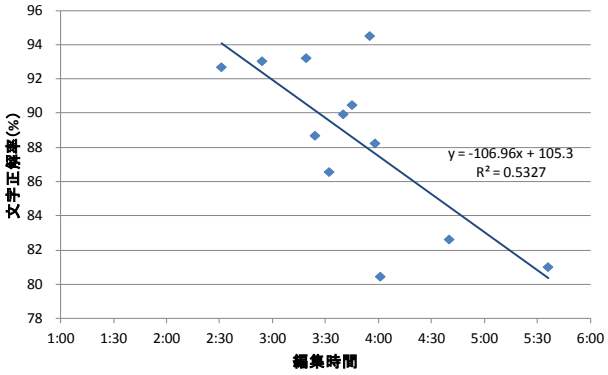

図 3　音声認識率と編集時間の相関(「リスク社会のライフデザイン」)

図 2・3 から音声認識率と編集時間の間には、かなり高い（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[6]。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. 音声認識を用いた字幕付与実施事例

前節の結果もふまえて、放送大学では実際に講義への字幕付与を実施している。

前記の 3 つのラジオ講義のインターネット配信しているコンテンツについては、2016 年度から字幕付与されている。その際に、講義で用いている図やグラフを静止画の形で貼り付けることで、通常のラジオ講義と比べて理解がしやすいようにしている（図 4 参照）。

また、2016 年度から開設されたオンライン授業の多くの番組で、音声認識を用いた枠組みで字幕が付与されている（図 5 参照）。ただしエディタは図 1 のものでなく、作業者がこれまで使い慣れたものを使用している。

現時点で、字幕付与が行われているオンライン授業の科目を以下に挙げる。

[2016 年度開講]
(1) がんを知る（15 回）
(2) 女性のキャリアデザイン入門（8 回）
(3) 感性工学入門（8 回）
(4) メディアと知的財産（15 回）
(5) 物理演習（8 回）
(6) 臨床推論（8 回）

[2017 年度開講]
(7) 学校と社会を考える（15 回）
(8) データの科学（15 回）
(9) フィールドワークと民族誌（15 回）
(10) 生涯学習を考える（15 回）
(11) 女性のキャリアデザインの展開（8 回）
(12) イランとアメリカ（15 回）

上記の合計は 115 時間に達しており、今後さらに増えていく予定である。

オンライン授業は以下のサイトで公開されており、体験版を視聴することもできる。

http://online-open.ouj.ac.jp/

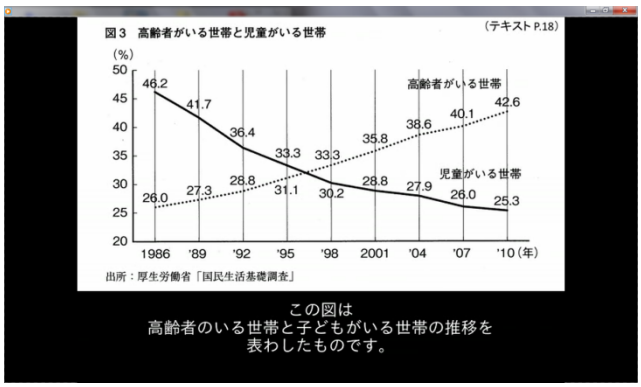

図 4 ネット配信の（ラジオ）講義への字幕付与例
（「リスク社会のライフデザイン」©放送大学）

図 5 オンライン授業への字幕付与例
（「物理演習」©放送大学）

6. おわりに

自動音声認識を用いることで効率よく字幕作成が可能になることを示した。同時に、そのためにはかなり高い認識率が必要であることもわかった。比較的容易に作業が可能であるので、他の様々なコンテンツに広がることを期待している。

字幕は情報保障として必要なものであるが、聴覚に障害がない人の学習においても有意義である。まず、電車内のように音を出せない環境でスマートフォンやタブレットで視聴する際に重宝する。しかしそれ以上に、字幕を併用することで理解が深まる効果が期待される。この効果は日本語のような表意文字で特に大きいように感じられ、字幕を表示する際にコマ切れでなく、意味的なまとまりのある文単位で表示する方が望ましい（図 4・5 参照）。このように、理解しやすい字幕の提示法についても検討が必要と考えられる。

参考文献

[1] 河原達也.
ICT・音声認識の活用による講演・講義の字幕付与.
情報処理, Vol.56, No.6, pp.543--546, 2015.
[2] 河原達也.
聴覚障害学生支援の最先端 --音声認識による字幕付与技術.
嶺重慎, 広瀬浩二郎（編）, 知のバリアフリー, 第 4 章,
pp.109--122. 京都大学学術出版会, 2014.
[3] 河原達也 編著.
音声認識システム （改訂 2 版）.
オーム社, 2016.
[4] 河原達也.
音声認識技術.
電子情報通信学会誌, Vol.98, No.8, pp.710--717, 2015.
[5] 西村雅史, 伊東伸泰.
講義コーパスを用いた自由発話の大語彙連続音声認識.
電子情報通信学会論文誌, VolJ83-DII, No.11, pp.2473-2480,
2000.
[6] 長妻令子, 福田健太郎, 柳沼良知, 広瀬洋子.
クラウドソーシングを活用した効率良い字幕作成手法.
電子情報通信学会技術報告, WIT2012-25, 2012.
[7] 河原達也.
議会の会議録作成のための音声認識－衆議院のシステムの概
要－.
情報処理学会研究報告, SLP-93-5, 2012.
[8] 秋田祐哉, 三村正人, 河原達也.
音声認識を用いた講義・講演の字幕作成・編集システム.
情報処理学会研究報告, SLP-108-2, 2015.