

音声認識を用いた生放送番組への字幕付与

佐藤 庄衛¹⁾

テレビ番組の音声で伝えられる情報を文字で伝える字幕放送は、聴覚障害者やテレビの音の聞き取りが容易でない高齢者にとって、テレビ番組を楽しむための有用なメディアである。しかし、1985年に字幕放送が開始されて以降2000年までの間、この字幕放送は事前に収録した番組に限定されていた。生放送番組にリアルタイムに字幕を付与できる方法が確立されていなかったためである。NHKでは、他の研究機関とも連携してニュース音声認識の研究開発を進め、2000年に世界に先駆けて音声認識を用いて生放送のニュース番組に字幕を付与した。本稿では、当時の音声認識を用いた字幕付与技術から、その後に、スポーツ番組の実況アナウンスや情報番組などの各種の生放送番組に字幕を付与するために研究開発された音声認識技術を紹介し、さらに効率的な字幕制作を目指して、現在検討中のシステムを紹介する。また、東日本大震災時の報道への字幕付与の実例とともに、残された音声認識技術の課題を考察する。

キーワード

音声認識、字幕放送、情報保障、認識誤り修正、リスピーク方式

1. はじめに

NHK（日本放送協会）は、放送法第15条に記載されるとおり、「公共の福祉のために、あまねく日本全国において受信できるように豊かで、かつ、良い放送番組による国内基幹放送を行うとともに、放送及びその受信の進歩発達に必要な業務を行う」ことを目的としている。また、放送法第20条（三）に記載のNHKの業務「放送及びその受信の進歩発達に必要な調査研究を行うこと。」を放送技術の側面から実施する放送技術研究所が組織されている。NHK放送技術研究所は、これまで衛星放送や高精細放送ハイビジョンを世に送り出し、将来の新しい放送メディアやサービスの研究開発を行なっている。そして、この研究所の成果は、国内外の標準化、技術協力、共同研究、特許の実施許諾などにより、放送受信機器、番組制作機器などを通じて社会に還元されている。

「人にやさしい放送」は、障害者や高齢者を含む、すべての人が好みの手段で放送を楽しみ、利用できる放送を目指すものであり、当研究所の重要な研究テーマの一つである（伊藤，2008）。この「人にやさしい放送」の一つに字幕放送が挙げられる。字幕放送は、ナレーションやせりふなどのテレビ番組の音声で伝えられる情報を、画面上に対応する文字を表示して伝えるものであり、聴覚に障害のある方や、テレビの音声の聞き取りが容易ではないと感じる高齢者などへの重要な情報保障手段で

ある。放送法の第4条の2では「放送事業者は（中略）音声その他の音響を聴覚障害者に対して説明するための文字または図形を見ることのできる放送番組をできる限り多く設けるようにしなければならない。」とその重要性が明記されている。

ここで、その歴史を振り返ると、日本の字幕放送は、1983年のNHK連続テレビ小説「おしん」での実験放送の後、1985年に本放送が開始されており、30年近くの歴史がある。しかし、字幕放送の開始から長い間、放送前に番組が完成していて、字幕も事前に作成することが可能なドラマやドキュメンタリーなどの事前収録番組だけで字幕放送が行われ、ニュースなどの生放送の番組には字幕を付与することができなかった。また、一般的に、字幕放送を受信するためには、テレビ受信機とは別に専用の文字放送受信機が必要であった。一方、アメリカでは、テレビ受信機は字幕放送受信機能（字幕デコーダー）を内蔵することが義務付けられ、裁判所の速記入力用に開発された高速に入力できる特殊なキーボードを用いて生放送にも字幕を付与できたために、字幕放送の普及は早かった（総務省・三菱総研，2006）。1990年代に入ると、日本でもニュースなどの生放送番組での字幕放送を要望する聴覚障害者からの声が高まった。特に、1999年9月の東海村臨界事故の際には、字幕のないテレビ報道の内容を聴覚障害者が十分に理解することができなかったため、不安が大きかったとして、災害報道における字幕の重要性が指摘された。

一方、1990年代のアメリカでは、Defense Advanced

¹⁾ NHK放送技術研究所

Research Projects Agency (米国防総省高等研究計画局)を中心として、英語ニュースの音声認識するための研究が盛んに行われていた。このプロジェクトは放送番組への字幕付与を目的としたものではなかったが、音声認識を用いた字幕制作の可能性が示された。1996年、放送技術研究所は、早稲田大学、豊橋技術科学大学、電気通信大学、東京工業大学、NTTと連携し、日本語のニュース番組にリアルタイムに字幕を付与することを目標として「ニュース音声認識プロジェクト」を発足させた。

その後の音声認識技術の研究開発の進展により、ニュース番組のアナウンサーがスタジオで原稿を読み上げている部分で、リアルタイムに字幕を付与するために必要とされる95%以上の認識率が得られるようになり、この成果を用いて、2000年3月末から世界に先駆けて生放送のニュース番組での音声認識によるリアルタイム字幕放送が開始された。本稿では、当時の音声認識技術から、2000年以降、スポーツの実況番組や情報番組など、各種の生放送番組に字幕を付与するために研究開発された音声認識技術を紹介し、さらに効率的な字幕制作を目指して検討を進めているシステムを紹介する。そして最後に、東日本大震災時の災害報道への字幕付与の実例とともに、残された音声認識技術の課題を考察する。

2. 字幕放送拡充のロードマップ

音声認識を用いた字幕放送の実用化例の紹介に先立ち、字幕放送を取り巻く現状と、字幕放送の拡充目標について述べる。

2.1 字幕放送

字幕放送で表示される文字は、クローズド・キャプションとも呼ばれ、番組のタイトル文字などのように、番組映像の上に文字がスーパーインポーズされているオープン・キャプションとは異なる。字幕の文字データは番組映像とは別に放送信号に多重されていて、受信機の機能で字幕文字が画面に表示される。したがって、必ず表示されるオープン・キャプションに対して、クローズド・キャプションの場合は、視聴者が必要に応じて、字幕の表示・非表示を選択することができる。

一般的なアナログ放送の受信機にはこの字幕を表示する機能がなかったため、字幕を見るためには専用の文字放送の受信機をテレビとは別に用意する必要があったが、デジタル放送の受信機には字幕表示機能が標準で装備されるようになった。完全デジタル化を終えた現在、テレビ受信機のリモコンの「字幕」ボタンを押せば、どの受信機でも字幕を楽しむことができるようになっていく。また、ワンセグを受信できる携帯端末でも字幕を楽しむことができるので、電車の中などの音を出せない公共の場で番組を楽しむ際には、健常者も字幕を利用する

ことがあり、字幕放送は、障害の有無を問わずユニバーサルな情報伝達手段となっている。

2.2 字幕の制作方法の分類

放送番組はその制作過程の違いにより、次の2種に分類することができる。一つは、放送前に番組の制作を終え、VTRなどのメディアに収録された番組を放送する事前収録番組である。ドラマやドキュメンタリーなどの番組がこれに相当する。もう一つは、番組を収録しながら放送する生放送番組であり、ニュース番組やスポーツ中継番組がこれに相当する。平成24年度前期のNHK総合テレビで7時から24時に放送された番組を調べると、約3割の番組が事前収録番組に相当する。これらの事前収録番組では、放送前に字幕の表示位置や表示タイミング、文字色などを入念に調整して人手で文字を入力して字幕を制作することができる。現在、事前収録番組に分類される番組のすべてにこうした字幕がついている。

一方、生放送番組に字幕を付与するためには、番組音声をもその場でリアルタイムに文字に変換する必要がある。アメリカの放送局では、裁判所の速記のために開発された特殊なキーボードを用いて、専門のオペレーターが一人で生放送番組の字幕を入力するのが一般的である。しかし、日本語の場合は、同音異義語が多く、かな漢字変換を要するので、同様の手段で字幕を制作することができない。そのため、音声認識を利用するなどの効率の良い字幕制作手段が必要とされている。

NHKでは、音声認識や各種キーボードを利用した効率の良い字幕制作方式を、番組の性質に応じて使い分けて、生放送番組の字幕を制作している。現在、生放送番組の字幕を制作するために用いられている方法は次の5つである。

- (1) パソコンなどに用いられている一般的なキーボードを利用する方式。複数の入力者が短い発話単位の文字をリレー方式で入力する。
- (2) ステノワード（西川、1995）と呼ばれる複数のキーを同時に押下する特殊な高速入力用キーボードを利用する方式。入力者と校正者のペアが数組で、短い発話単位の文字をリレー方式で入力する。
- (3) 音声認識で直接番組の音声を認識する方式（ダイレクト方式）（安藤、2001）。音声認識の認識誤りは人手で修正する。
- (4) 字幕制作のために、番組の音声を専用のアナウンサーが言い直した音声を認識する方式（リスピーク方式）（松井、2004）。
- (5) 認識率の高い部分にダイレクト方式を利用し、それ以外の部分でリスピーク方式を利用するハイブリッド方式（本間、2009）。

キーボードを用いて入力する(1)、(2)の方式は、番組の話題や発話スタイル、背景雑音などの制限を受けること

なく、オペレーターが柔軟に文字を入力することができる。しかし、入力するオペレーターに熟練した専門技術が必要となり、リレー方式であるため、多くの熟練オペレーターを確保する必要がある。これらの方式では、熟練オペレーターの確保が課題となる。一般のキーボードを用いる方式(1)は、人の話す速さで入力できる文字の量が限られているため、歌謡番組など字幕の分量が少ない番組で利用されている。高速入力キーボードを用いる方式(2)では、オペレーターに高度な専門性が要求され、オペレーターの確保と育成が必要である。入力できる分量の制約は小さく、様々な番組に字幕を付与することができる。しかし、確保できるオペレーターの制限により、主にニュースなどの報道番組で利用されている。

一方、音声認識を用いた(3)、(4)、(5)の方式は、番組ごとに認識に用いる辞書を事前に学習する必要がある、ここで学習された辞書の良し悪しで認識精度が決まる。また、原稿の読み上げ音声、対談音声など、入力音声の発話スタイルや背景雑音の影響を受けて認識精度が左右される。音声認識を利用した方式は、これらの理由により適応できる番組が限られているが、(1)、(2)の方法に比べ、字幕制作に関わる人的負担が音声認識により小さくなるため、リスピーク話者や認識誤りを修正するオペレーターなどの必要な人材に高度な専門性は要求されず、確保しやすいという利点がある。各方式の詳細は次章で述べるが、ダイレクト方式(3)はニュース番組(2000～2006年)と大リーグ野球の中継番組で利用され、リスピーク方式(4)は大相撲やプロ野球などのスポーツ中継番組と情報番組で利用されている。また、2012年3月より、定時のニュース番組においてハイブリッド方式(4)が利用されるようになった。

2.3 字幕の拡充目標

字幕放送は年々拡充されてきているが、すべての番組で字幕放送が行われている状況にはまだ至っていない。2010年度現在で、総務省がまとめた各放送局の総放送時間に占める字幕放送の割合(総務省, 2011)は、NHK総合で56.2%、民放在京キー5局で43.8%である。

2007年、字幕放送の普及のために総務省は、「視聴覚障害者向け放送普及行政の指針(総務省, 2007)」を策定している。この指針では、話者の多い討論番組など技術的に字幕を付与できない番組、外国語の番組、音楽番組、権利処理上の理由で字幕を付与できない番組を除き、7時～24時に放送されるすべての字幕付与可能な番組に、2017年までに字幕を付与することを目標としている。2010年度現在で、この普及目標の対象となる放送番組における字幕番組の割合は、NHK総合で62.2%であり、2017年の目標達成に向けて、効率的な字幕制作方法の研究開発と字幕放送の拡充を進めていく。

また、同指針は策定から5年後を見通しに見直しを行う

ことになっており、策定から5年を経た現在、見直しに向けた検討が進められている(総務省, 2012)。

3. 音声認識を用いたリアルタイム字幕制作システム

3.1 音声認識

音声認識は、人の声をコンピュータで自動的に文字に変換する技術である。現在の音声認識では、日本語の母音や子音などの音素で観測される音響信号を学習した統計モデルと、日本語文章中での単語と単語の繋がりやすさを評価し、入力音声信号に対応付けられた単語列の文章らしさを評価する統計モデルを用いた確率統計的手法が主流である。ここで前者を音響モデル、後者を言語モデルと呼び、これらのモデルを精度よく学習することにより、数万単語の大語彙を用いて連続発声された文章を精度よく認識できるようになる。

生放送番組の字幕制作に用いられる音声認識システムには、「精度よく認識できること」、「音声の発話から字幕の表示までの時間差を短縮するため、逐次、リアルタイムに認識結果が出力されること」が要求される。また、認識誤りを効率良く修正できるシステムであることが要求される。以下に、生放送番組の字幕制作に用いられてきた4つの音声認識システムを紹介する。

3.2 ダイレクト方式

前述のとおり、放送技術研究所では他の研究機関と連携して、ニュース音声認識の研究を開始した。最初の目標は、ニュース番組のアナウンサーの音声の認識であり、放送されたニュースのアナウンサーの音声とニュース原稿を大量に収集して、音声データベースを構築した。このデータベースを元に、ニュース向け音響モデルと言語モデルの学習法を検討し、音声認識の高精度化の研究を進めた(小林, 1997)。

ここでの研究開発の成果により、ニュース番組のアナウンサーがスタジオで原稿を読み上げた音声の単語認識率を95%まで改善することができた。この95%の単語認識率は、生放送の字幕制作に音声認識を利用するために目標とする認識率である。NHKの放送するニュース番組は、一文が平均40単語からなり、一文あたりの読み上げ時間は12秒であるため、95%の単語認識率の認識結果には、平均すると6秒に1単語の誤りが出力され、この認識誤り単語は人手で修正することができる。しかし、単語認識率が90%に低下すると、平均して3秒に1単語の認識誤り単語を修正する必要がある、リアルタイムの修正が困難になる。

そして、2000年3月、ニュース番組のアナウンサーによる原稿読み上げ部分に限定しつつも、音声認識を利用したニュース番組での字幕放送が、世界に先駆けてNHK「ニュース7」で開始された。これは、1999年11月の衆議院通信委員会でNHK会長が「翌年度から、ニ

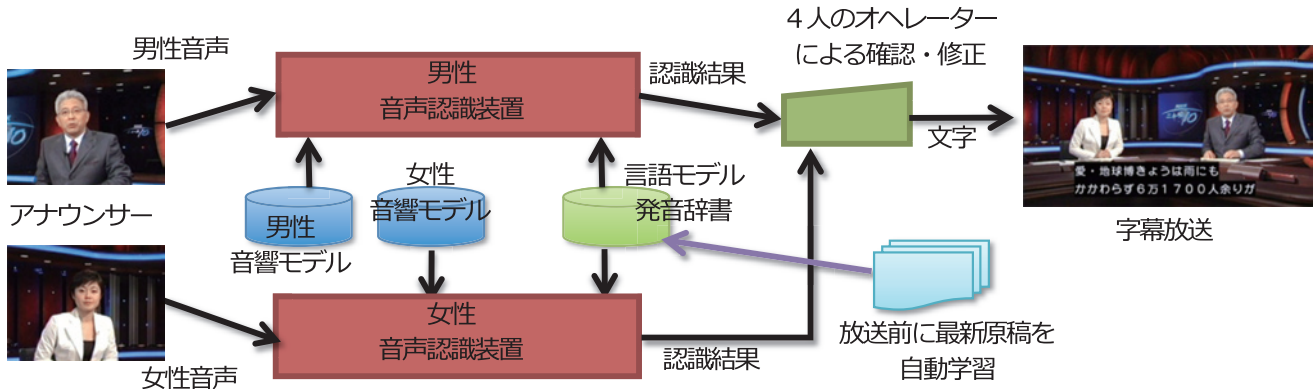


図1 ダイレクト方式の音声認識によるニュース番組の字幕制作（2000～2006年）

ユースのメインアナウンサー部分で音声認識を利用した字幕放送を開始し、段階的に拡大したい。」と答弁したことが発端となった。その後、短い準備期間で認識誤りを人手で修正するシステムをメーカーと共同で開発し、音声認識部分は放送技術研究所の実験システムをそのまま利用して運用を開始することになった。現場にいる記者からのレポートやインタビュー部分には字幕がなく、アナウンサーの原稿読み上げだけの部分的な運用ではあったが、日本での生放送番組への字幕付与は初めてのことであり、聴覚障害者や高齢者からの反響は大きかった。

この字幕制作システムの構成を図1に示す。このシステムでは、男性用と女性用の2式の音声認識装置が稼働しており、ニューススタジオのアナウンサーの性別に応じて、各アナウンサーに取り付けられたマイクの出力がそれぞれの音声認識装置に入力される。この2つの音声認識は同じ認識アルゴリズムを用い、共通の言語モデルを用いているが、男性または女性の音声だけから学習した音響モデルを用いる点が異なる。音素の特徴が大きく異なる男女の音声を別々にしたこと、男女共通の音響モデルを用いる場合よりも高い認識率が得られ、マイクごとに話者を特定できるニューススタジオの特性を生かしたシステムである。

この音声認識に用いられている音響モデルは、前後の音素によって異なる音素の特徴を考慮した音素環境依存モデルを学習したHidden Markov Modelであり、言語モデルは、放送直前に最新のニュース原稿を自動的に取得して重みづけ学習したモデルである（小林，1999）。ここで学習できる原稿は、アナウンサーが読み原稿用に手書きで加筆修正（図2）する前の原稿であるが、この学習により、最新のニュースを認識するために必要な固有名詞や人名などを認識できるようになる。認識に用いられる語彙は約2万単語であり、各単語が連続して出現する確率がNグラム言語モデルとして学習される。

音声認識には、上述の音響モデルと言語モデルを用いて、第1パスで2単語連鎖確率を用いて単語依存N-best探索（複数の直前単語で一つの静的な木構造化単語辞書

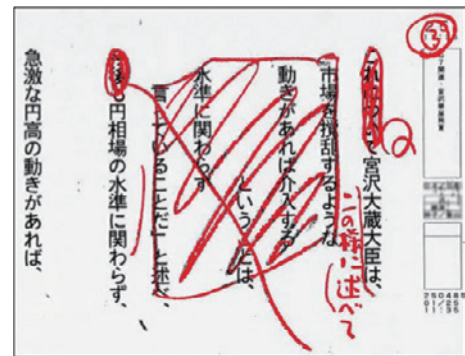


図2 アナウンサーの読み原稿。コメントを読み上げる時間などを調整するために、直前まで手書きで加筆修正される。



図3 ダイレクト方式によるニュース字幕における誤り修正の様子。誤り発見者と修正者のペア2組で修正。

を共有して探索）を行いながら、発話途中で繰り返し第2パスを実行し、3単語連鎖確率を用いてリスクアリングする逐次2パスデコーダー（今井，2001）が用いられた。

認識結果の誤りは、誤りの発見者と修正者のペア2組、計4名のオペレーターが修正する（図3）。リアルタイムに出力される音声認識結果をオペレーターが確認し、誤りがあれば、誤り部分をタッチパネルで指定し、それ

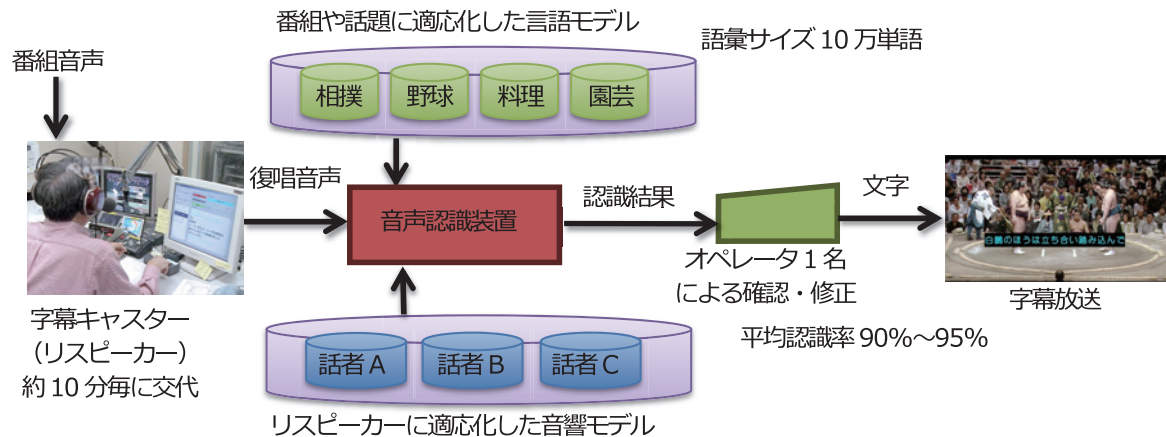


図4 リスピーク方式の音声認識によるスポーツ中継や情報番組の字幕制作（運用中）

をキーボードで即座に修正し、字幕を放送する。ニュース番組の字幕には、特に正確さが求められ、誤りのないことはもちろん、要約や書き換えなしに、一字一句アナウンサーの発話通りの字幕が求められる。

字幕は、画面の下部に最大15文字の2行で表示し、読みやすさのために、3.5秒で更新するのが基本となる。このため、音声が発声されてから字幕が表示されるまでは、平均10秒の遅れが生じる。この遅れの原因は、音声認識に1秒以内、認識結果の確認と修正に5～6秒の遅れが生じるのに加え、字幕表示の更新のためのバッファリング時間が必要になるためである。また、遅れ時間の課題に加え、すでにスーパーインポーズされているオープン・キャプションの文字との重なりも課題として残された。

2000年に開始されたスタジオのアナウンサーの原稿読み上げ部分の字幕放送は、翌年、ステノワードと呼ばれる特殊な高速入力キーボード（西川，1995）を用いる方式を原稿の読み上げ部分以外の中継や対談のニュース項目で併用することにより、字幕を付与することのできる番組が増えた。しかし、音声認識は原稿読み上げ部分での限定的な運用であったのに対し、高速入力キーボードを用いた方式はニュース項目を問わず運用可能であったため、経費削減を理由として2006年から2012年3月までの間、NHKのニュース番組の字幕制作は高速入力キーボードを用いた方式に一本化された。しかし、2012年3月現在、認識性能が高く、従来よりも少人数で運用可能な後述のハイブリッド方式が一部のニュース番組の字幕制作に利用されている。

また、アメリカ大リーグ野球中継の日本語の実況は、アナウンサーが現地のスタジアムではなく、日本の静かなスタジオで実況することが多い。このような実況では、ダイレクト方式の音声認識を用いることができ、字幕制作に利用されている。

3.3 リスピーク方式

ニュース番組での字幕放送開始後、生放送のスポーツ中継や音楽番組、情報番組にも字幕を付与したいとの要望が高まった。しかし、これらの番組では、多様な話者が様々な発話スタイルで発声するため、十分な認識率が得られない。例えば、スポーツ中継では、実況アナウンサーの感嘆詞や、得点シーンなどを興奮した口調で実況する部分が認識精度の低下の要因となる。さらに、スタジオでの収録と異なり、背景雑音の混入を避けられない場合も多く、背景雑音に起因する認識精度の低下も課題となる。

このような番組では、静かなスタジオにいる字幕キャスターが、ヘッドフォンで番組音声を聞きながら音声認識のために番組中のアナウンサーや解説者の言葉を復唱した音声を認識する「リスピーク方式」が採用されている（図4）。このリスピーク方式は、英国BBCが2001年9月に、スヌーカー、ゴルフ、テニス、陸上競技の4種目の中継で字幕放送を実施するために採用した方式であり、現在は英国など欧州や南米の放送局の生放送番組の字幕制作方式の主流となっている。

NHKでもメーカーと共同でリスピーク方式の音声認識によるリアルタイム字幕制作システムを開発した（NHK，2003）。この方式は、2001年12月の紅白歌合戦をはじめとし、オリンピック、ワールドカップサッカー、大相撲中継、プロ野球へ字幕を付与するために利用されている（多部，2004）。ここで挙げられたスポーツ中継では、画面に映し出される各選手のプレーなどと字幕の内容に不一致があると、競技の進行を理解することが難しくなるため、5～10秒の字幕の表示遅れをなるべく小さくし、読んでわかりやすい字幕を制作しなければならない。字幕キャスターは画面を見てわかることを省いたり、内容を簡潔にまとめて言い換えたり（松井，2004）、拍手や歓声などの競技場の音の様子を実況アナウンサーのコメントに補足するなどの工夫をしている。また、情報番組など複数の出演者の発話が交錯するような番組の

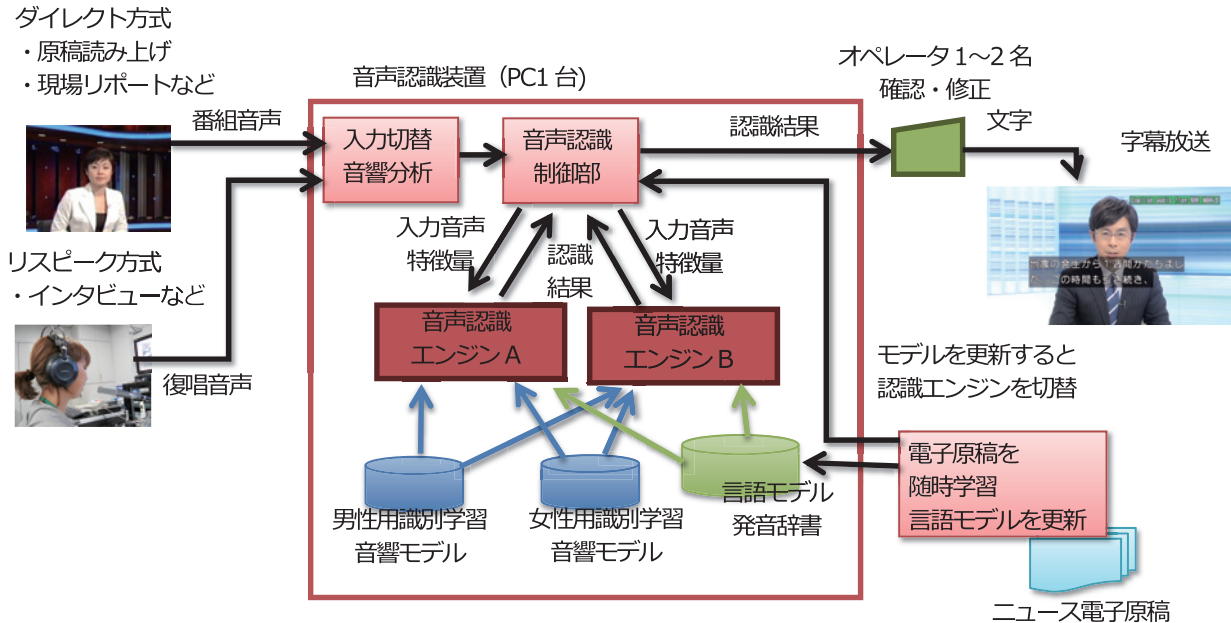


図5 ハイブリッド方式の音声認識によるニュース番組の字幕制作（運用中）

音声を直接認識するのは困難であるが、本手法では一人の字幕キャスターが内容を整理して復唱することができるため、視聴者にわかりやすい字幕を提供することができる。

このリスピーク方式で用いる音声認識では、予め字幕キャスターごとに適応学習した音響モデルを用いて認識率の向上を図っており、それらの音響モデルは番組途中の字幕キャスターの交代に対応できるように、瞬時に切り替えられるようになっている。上述のとおり、様々な工夫をしながらの復唱を行う字幕キャスターの疲労に配慮し、数十分ごとに別の字幕キャスターに交代しながら字幕放送を行うのが通常である。

一方、言語モデルには、スポーツ競技などの番組のジャンルごとの学習テキストから学習した語彙サイズ約10万単語の大規模な番組専用の言語モデル用いている。特にワイドショーのような情報番組は、広範な話題を有しており、それぞれの話題を認識できる言語モデルの学習が課題となっている。記者が入稿した最新のニュース原稿をデータベースから学習できるニュース番組と異なり、情報番組で得られる事前情報は、A4紙面数枚の構成表しか得られないことも多く、番組の大まかな進捗がわかるだけである。そのため、この情報をもとに番組の話題に関連するテキストを人手で手間をかけて収集して言語モデルを学習している。

これらの理由により現状では、言語モデルの学習に十分なテキストが得られる番組か、定期的に同種の話題の字幕を制作することで、結果的に学習データの収集コストを下げられる番組が本方式の適用対象となっている。今後、字幕放送を様々な番組に拡充していくためには、音声認識の専門家がこなしている語彙の選定や新出単語

の登録、学習テキストの収集、必要な単語を認識できるようにするための各種対策などの事前準備を、専門家でもなくても効率良くできるような仕組みを作っていく必要がある。

3.4 ハイブリッド方式

2000年に開始されたダイレクト方式の音声認識は、その後の技術の進展により、アナウンサーの原稿読み上げ部分だけでなく、記者による現場からのリポート、アナウンサーと記者の比較的落ち着いた対談でも実用的な認識率が得られるようになった。一方、それ以外のインタビューや編集済みのビデオ素材、自由発話の多い対談などの部分では、背景にある雑音や音楽のレベルが高かったり、発音が速くて不明瞭であったり、くだけた会話口調であったりするため、字幕制作に十分な認識率95%が得られていない。

そこで、高速入力キーボード方式による字幕が付与されていないニュース番組に字幕を付与することを目的として、ハイブリッド方式のリアルタイム字幕制作システム（本間，2009）の開発を進めてきた。図5に示すとおり、この方式では、十分に高い認識率が得られる条件の番組音声をダイレクト方式で直接認識し、それ以外のインタビューなどの部分ではリスピーク方式を採用して字幕アナウンサーの復唱音声を認識する。従来のダイレクト方式では困難だった音声認識のみによるニュース番組全体への字幕付与が、このハイブリッド方式により可能になった。

ダイレクト方式を使うか、リスピーク方式を使うかは、字幕キャスターまたは修正オペレーターがニュース項目ごとに判断して、音声入力を切り替える。ここで、余裕



図6 認識誤り修正作業の様子。確認・修正のためのインターフェースが改善され、修正オペレータの人数は1～2名に削減された。

を持って音声入力を切り替えることができ、切り替え時に発話の冒頭部分が欠落しないように、入力音声をバッファリングし、切り替え時刻から1秒程度さかのぼった入力音声から認識されるようになっている。

また、認識結果の確認・修正のインターフェースも改善され、操作性の向上により複数の修正オペレーターが同時に効率良く作業を行えるようになった（大出，2003）。さらに、番組ごとの音声認識の難しさに応じて、修正オペレーターの人数を増減できるようになった。通常の定時に放送される5分から10分のニュース番組であれば、1～2名のオペレーターで字幕放送を実現できる（図6、図7）。

ハイブリッド方式の音声認識装置も、番組音声と復唱音声を高精度に効率良く認識するために改善された。従来のダイレクト方式の音声認識では、男性話者の音声と女性話者の音声の入力を分けておかなければならなかったが、本方式では、男女並列連続音声認識（Imai, 2007）を採用することにより、入力音声の話者の性別をシステムが自動的に認識して、男性用と女性用の音響モデルから適切なモデルを使用して認識できるようになった。また、同技術で用いられているエンドレス音素認識により、現場からの中継リポートなどの背景雑音の多い入力音声でも適切に発話の区間を検出できるようになった。

さらに、使用する音響モデルの学習にも選択的識別学習（Povey, 2002）（今井，2008a）が採用された。この学習方法では、学習音声を実際に認識システム認識し、その認識システムで起こる認識誤りを学習することができるため、従来の音響モデルの学習法に比べ、認識誤りを少なくすることができる。

一方、言語モデルは、字幕を制作している番組の途中に入稿された最新の電子原稿を反映できるようになった。この言語モデルの自動更新（今井，2008b）機能は、音声認識装置内に2系統の音声認識を用意することによ

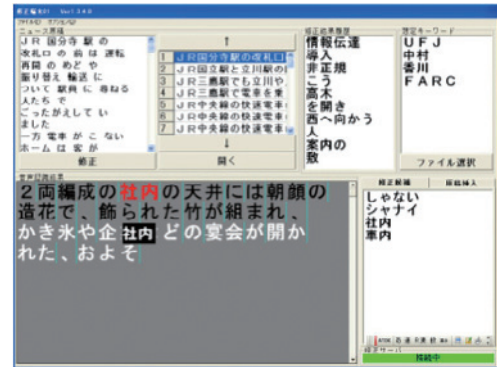


図7 改善された修正・確認インターフェース。認識誤り単語をタッチしてキーボードで修正するのが基本だが、各種修正候補単語をタッチしても修正可能である。他の修正者の修正作業も確認でき、同時共同作業が可能になった。

り実現されている。常時2系統の音声認識のうちの1系統が稼働しており、新しい言語モデルが学習されると、稼働していない音声認識エンジンにそのモデルが読み込まれ、発話の合間にこの最新のモデルを読み込んだ音声認識エンジンに切り替えられる。

このハイブリッド方式は、従来よりも効率的で運用コストの低い字幕制作システムであるという長所を生かして、2012年3月より、16時からの5分間の定時ニュース番組の字幕付与に採用された。さらに、高速入力キーボードなど、コストの高い字幕制作の利用が難しいローカル放送に字幕を制作する手段としても期待されており、2012年9月より、関東ローカル番組の「首都圏ネットワーク」と「首都圏ニュース845」に本方式を用いて字幕が放送されるようになった。これらのローカル放送は、主にアナウンサーの原稿読み上げ音声で構成される定時ニュースとは異なり、収録・編集済みの映像を再生する特集コーナーがあったり、それらの特集を見た感想を対談風に述べたりするため、音声認識の対象としての難易度は高い。しかし、収録・編集済み映像の部分では、確認・修正インターフェースに事前に読み込んでおいた原稿を、番組の進行にしたがって送付したり、対談部分ではリスピーク方式を活用したりするなど、字幕放送制作現場のノウハウとともに実用化に至っている。以降、大阪や名古屋などのからローカル放送での字幕放送が順次実現していく予定である。今後、総務省の字幕放送普及の目標を確実に達成していくため、音声認識の適用範囲の拡大と種々の字幕制作手法の進展が望まれる。

4. 字幕制作に向けた音声認識の課題

4.1 災害報道対応

2011年3月、多くの人命と財産を失う事となった東日本大震災の時には、NHKをはじめとする報道機関が重要な役割を担うこととなった。人々の安全と安心を守り、

人命と財産に関わる情報を適切に伝えることは、公共放送の使命であり、本震災時にも災害に関わる情報を克明に報道し、それは長期に渡った。

災害の状況を、障害者をはじめとする多くの方々に伝えるため、いち早く立ち上がったのは高速入力キーボードを用いた字幕制作方式であった。災害報道の字幕は、障害者だけでなく、避難所でワンセグ放送を利用して災害の状況を知ろうとした方々の役にも立ったと思われる。しかし、災害報道が長期化するにつれて、高速入力キーボードで字幕を付与できる量の限界に近づいた。そこで、字幕による情報提供を充実させるため、まだ実験用の試作機であったハイブリッド方式のシステムを臨時で放送センターに設置し、ニュース番組の一部で音声認識を用いた字幕放送を実施した。

図8は東日本大震災のニュースで臨時に使用した音声認識による字幕制作の様子である。この経験を通して、災害報道の字幕を制作するための音声認識の課題が明らかになった。前述のとおり、音声の認識率は100%ではないため、必ず誤り修正者が必要であり、災害時の修正者の確保が第1の課題となる。また、災害報道を認識するためには災害に関する情報を音声認識の言語モデルに学習させる必要があり、災害に関連した単語や言い回しを災害に応じて迅速に収集して学習する方法の開発が第2の課題である。そして第3の課題は、音声認識を適用できるニュース項目の推定である。災害報道では、被災現場や避難所などの様々な場所の様々な音環境下からの中継があったり、スタジオで災害の専門家との対談があったりするため、認識精度の推定が困難であった。実際に、どの項目で音声認識を字幕制作に利用できるかは、現場担当者の経験にもとづいて決められた。東日本大震災の災害報道にもとづいて、音声認識を適用できる項目を検証しておく必要がある。現在放送技術研究所では、上記の第2の課題と第3の課題について検討を進めている。

4.2 広範な話題に対応した音声認識

NHKの放送するワイドショーのような情報番組には、リスピーク方式による字幕が付与されている。これらの情報番組は、非常に広範な話題を扱うことが特徴である。このような番組の字幕制作にあたっては、これらの話題を認識するために、言語モデルの学習に必要なテキストデータを人手で収集しており、この作業にかかるコストが大きい。

このような広範な話題に対応した言語モデルの自動学習を目的として、簡易な番組構成表の記載から、認識に必要な話題を推定し、その話題に言語モデルを適応化するための研究開発も進めている。



図8 東日本大震災ニュースで臨時に使用した音声認識による字幕制作の様子。

4.3 ダイレクト方式の拡充

地方局におけるローカル放送への字幕付与の現場など、多くの番組制作現場では、リスピーク方式にかかる字幕キャスターのコストは大きく、ダイレクト方式による字幕制作が望ましく、ダイレクト方式を適用できる番組の拡充が望まれている。

今後、音響モデルや言語モデル、認識アルゴリズムの改善などにより、より多くの番組で、字幕キャスターなしで字幕を制作できる音声認識システムの研究開発を進める必要がある。

5. おわりに

本稿では、NHKの生放送番組にリアルタイムで字幕を制作するために研究開発してきた音声認識技術を紹介し、本技術にもとづいて実用化されてきた各種の字幕制作システムの概要を述べた。音声認識技術は、今やNHKのリアルタイム字幕制作に欠かせないものとなっている。

リアルタイム字幕放送の開始以来、聴覚障害者や高齢者からは、「字幕のおかげで家族といっしょに番組を楽しめるようになった。」「スポーツ選手の心理や対戦の意味合いが字幕でわかり、楽しみが広がった。」といった声が寄せられている。リアルタイム字幕放送は年々拡充されてきたが、残念ながら現状の音声認識技術は全ての番組の音声を認識して字幕を制作できるまでには至っていない。今後は、総務省の字幕放送普及目標を確実に達成するためにも、さらなる研究開発を続けていきたい。

謝辞

本稿で述べた音声認識による字幕制作システムは、

NHK編成局, 報道局, 放送技術局, 技術局, 放送技術研究所, (株)NHKグローバルメディアサービス, (株)パナソニックの連携によるものであり, 研究開発および実用化にあたられた関係各位に深く感謝する。また, 音声認識の研究開発にあたり, 早稲田大学・白井克彦名誉教授, 小林哲則教授, 東京工業大学・古井貞熙特命教授, 電気通信大学・尾関和彦名誉教授, 豊橋科学技術大学・中川聖一教授の諸先生方にご指導頂いた。ここに深く感謝の意を表する。

引用文献

- D. Povey and P. C. Woodland (2002). Minimum phone error and I-smoothing for improved discriminative training. *Proceedings of IEEE ICASSP*, I-105-108
- NHK (2003). NHK技術情報 自動音声認識による生字幕制作システムを開発 2003年9月5日 <<http://www3.nhk.or.jp/pr/marukaji/m-giju093.html>> (2012年9月10日)
- T. Imai, S. Sato, S. Homma, K. Onoe, and A. Kobayashi (2007). Online speech detection and dual-gender speech recognition for captioning broadcast news. *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 8, 1286-1291
- 安藤彰男, 今井亨, 小林彰夫, 本間真一, 後藤淳, 清山信正, 三島剛, 小早川健, 佐藤庄衛, 尾上和穂, 世木寛之, 今井篤, 松井淳, 中村章, 田中英輝, 都木徹, 宮坂栄一, 磯野春雄 (2001). 音声認識を利用した放送用ニュース字幕制作システム 電子情報通信学会論文誌, Vol. J84-D-II, No. 6, 877-887.
- 伊藤崇之 (2008). 高齢者・障害者のメディアアクセスに関する話題 一人にやさしい放送を目指した研究開発— 電子情報通信学会2種研究会サイバーワールド第9回研究会, 1-6.
- 今井亨, 田中英輝, 安藤彰男, 磯野春雄 (2001). 最ゆう単語列逐次比較による音声認識結果の早期確定 電子情報通信学会論文誌, Vol. J84-D-II, No. 9, 1942-1949
- 今井亨, 小林彰夫, 佐藤庄衛, 本間真一, 奥貴裕, 都木徹 (2008a). 放送用リアルタイム字幕制作のための音声認識技術の改善 第2回音声ドキュメント処理ワークショップ, 113-120
- 今井亨, 本間真一, 小林彰夫, 奥貴裕, 佐藤庄衛, 都木徹 (2008b). リアルタイム字幕制作のためのモデル自動更新に対応した音声認識 音響学会秋季講演論文集, 1-1-16, 37-40
- 大出訓史, 三島剛, 江本正喜, 今井篤, 都木徹 (2003). 効率的なリアルタイムニュース字幕修正システム 映像情報メディア学会年次大会, 6-3
- 小林彰夫, 今井亨, 安藤彰男, 宮坂栄一, 赤松裕隆, 中川聖一, 小黒玲, 尾関和彦, 古井貞熙, 鈴木順子, 白井克彦 (1997). ニュース音声認識システムの検討 音響学会秋季講演論文集, 3-1-9, 103-104
- 小林彰夫, 今井亨, 安藤彰男, 中林克己 (1999). ニュース音声認識のための時期依存言語モデル 情報処理学会論文誌, Vol. 40, No. 4, 1421-1429
- 総務省, 三菱総研 (2006). 国内外における視聴覚障害者向け放送に関する調査研究 2006年3月24日 <http://www.soumu.go.jp/main_sosiki/joho_tsusin/b_free/pdf/060810_1.pdf> (2012年9月10日)
- 総務省 (2007). 視聴覚障害者向け放送普及行政の指針の公表 2007年10月30日 <http://warp.ndl.go.jp/info:ndljp/pid/283520/www.soumu.go.jp/s-news/2007/071030_2.html> (2012年9月10日)
- 総務省 (2011). 平成22年度の字幕放送等の実績 2011年8月30日 <http://www.soumu.go.jp/menu_news/s-news/01ryutsu05_01000012.html> (2012年9月10日)
- 総務省 (2012). デジタル放送時代の視聴覚障害者向け放送の充実に関する研究会報告書の公表 2012年5月23日 <http://www.soumu.go.jp/menu_news/s-news/01ryutsu05_02000022.html> (2012年9月10日)
- 多部多栄子, 椎名努, 堂免大規 (2004). 生字幕放送サービス—システムとサービス概要— 映像情報メディア学会技術報告, BCT2004-24, Vol. 28, No. 5, 17-20
- 西川俊, 高橋秀知, 小林正幸, 石原保志, 柴田邦博 (1995). 聴覚障害者のためのリアルタイム字幕表示システム 電子情報通信学会論文誌, Vol. J78-D-II, No. 11, 1589-1597.
- 松井淳, 本間真一, 小早川健, 尾上和穂, 佐藤庄衛, 今井亨, 安藤彰男 (2004). 言い換えを利用したリスピーク方式によるスポーツ中継のリアルタイム字幕制作 電子情報通信学会論文誌, Vol. J87-D-II, No. 2, 427-435.
- 本間真一, 小林彰夫, 奥貴裕, 佐藤庄衛, 今井亨, 都木徹 (2009). ダイレクト方式とリスピーク方式の音声認識を併用したリアルタイム字幕制作システム 映像情報メディア学会誌, Vol. 63, No. 3, 331-338.



さとう しやうゑい
佐藤 庄衛

1991年東北大・工・電気卒。1993年同大大学院修士課程修了。同年NHK入局。仙台放送局を経て、1995年より同放送技術研究所勤務。博士(工学)。音声認識を用いた生放送番組への字幕付与の研究に従事。電子情報通信学会, 日本音響学会, 情報処理学会, 映像情報メディア学会会員。

Closed-Captioning of Live Broadcasts by Using Speech Recognition

Shoei Sato¹⁾

Closed-captioning, which displays the spoken content of a broadcasting program as text, is an important media for supporting people who have difficulty following speech. NHK has provided closed-captioning since 1985 for the hearing impaired and the elderly. Until 2000, closed-captioning could only be provided for prerecorded TV programs, because of the lack of a real-time input method of Japanese text. NHK's extensive research on speech recognition conducted in cooperation with other research institutes led to the start of real-time closed-captioning for broadcast news in 2000. The technologies of this system continued to be developed, and this effort eventually extended NHK's live captioning capability to sports programs and talk and variety shows. This paper describes the technologies developed for real-time captioning of broadcast news and their subsequent developments. It also introduces a new system that is currently under study and discusses its technical issues through the illustrative example of live captioning of broadcasts of the Great East Japan Earthquake.

Keywords

Closed-Captioning, Speech Recognition, Live Captioning

¹⁾ NHK Science and Technology Research Laboratories