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

Real-Time Closed-Captioning System Using Speech Recognition of Direct Program Sound and Re-Spoken Utterances

正会員 本間 真一[†], 正会員 小林 彰夫[†], 奥 貴裕[†],
佐藤 庄衛[†], 正会員 今井 亨[†], 正会員 都木 徹[†]

Shinichi Homma[†], Akio Kobayashi[†], Takahiro Oku[†], Shoei Sato[†], Toru Imai[†] and Tohru Takagi[†]

Abstract A new real-time closed-captioning system for Japanese TV news broadcasts is described. The system is based on a hybrid automatic speech recognition system that switches its speech input between the original program sound and speech rephrased by a “re-speaker”. By improving our speech recognizer and its error correction system, we decreased the number of correction operators needed from four in our previous news system to one or two, depending on the difficulty of the speech recognition. Experimental results show that the system could be used at broadcasting stations for captioning typical news programs more easily and by fewer staff.

キーワード：字幕放送, 音声認識, ダイレクト方式, リスピーク方式, 認識誤り修正

1. ま え が き

放送番組の発話内容を文字で伝える字幕放送（クローズド・キャプション）は、主に聴覚障害者への情報保障手段として年々拡充が図られてきた¹⁾。最近では、字幕の表示機能を標準で装備したデジタル放送受信機やワンセグ携帯端末の普及により、字幕放送は健聴者にとっても有効な、ユニバーサルな情報伝達手段となってきた。ドラマやドキュメンタリーなどの事前制作番組の字幕は、放送前にある程度の時間をかけて制作することが可能であるが、ニュースやスポーツなどの生放送番組においては、番組音声からリアルタイムで字幕（以下、生字幕と呼ぶ）を制作しなければならない。ところが、一般的なパソコンのキーボード等による入力手段では時間遅れが大きく、生字幕の制作は困難であるため、さまざまな工夫や技術的な支援が必要となる。

NHK では、これまでに音声認識技術を用いたリアルタイム字幕制作システムを開発し²⁾、2000 年に一部のニュース番組において字幕サービスの提供を開始した。しかしながら、当時の音声認識の性能で生字幕の制作が可能であっ

たのは、十分に高い認識率が得られるアナウンサの原稿読み上げ部分に限定されていた。また、音声認識の誤りを発見し、即座に修正するために、4 名のオペレータを必要とした。2001 年からは、中継や対談をはじめとした、アナウンサの原稿読み上げ部分以外にも字幕を付与するために、複数のオペレータがリレー方式で入力する特殊な高速入力用キーボード（ステノワード）³⁾ による方式を導入した。そして 2006 年からは、ニュース番組の字幕制作では部分的な運用であった音声認識による方式を中断し、ステノワードによる方式に一本化され、現在に至っている。

一方、スポーツ番組の生字幕制作では、直接の音声認識が困難であるため、別の専任の話者（リスピーカ）が番組音声を取り、復唱した音声（リスピーク）を認識させるリスピーク方式^{4)~6)}を採用するケースが多い。この方式では、歓声等の背景雑音や、相槌等に見られる複数話者の発話の重なりを回避できるほか、話し言葉口調の発話を書き言葉口調に言い換えることができるため、高い音声認識性能が得られ、かつ、読みやすい生字幕を制作できるというメリットがある。しかし、リスピーカの要員確保が必要であり、疲労の観点から、リスピーカの交代についても配慮をしなければならないというデメリットも存在する。

2006 年度の字幕放送の時間数の割合⁷⁾は、技術的に字幕を付することが困難な生放送番組等を除くと、NHK 総合が 100.0%、民放キー 5 局平均が 77.8%であったが、生放送も含む総放送時間に対しては、NHK 総合が 43.1%、民

2008 年 6 月 26 日受付, 2008 年 12 月 18 日最終受付, 2009 年 1 月 20 日採録

[†]NHK 放送技術研究所 (人間・情報)

(〒157-8510 世田谷区砧 1-10-11, TEL 03-5494-3369)

[†]NHK Science and Technical Research Laboratories

(1-10-11, Kinuta, Setagaya-ku, Tokyo 157-8510, Japan)

放キー 5 局平均が 32.9%であり、依然として字幕が未付与である番組の比率は高い。2007 年度に総務省が策定した新たな字幕放送普及行政の指針⁸⁾では、生放送も含む 7~24 時の字幕付与可能なすべての放送番組に、2017 年度までに字幕付与することが目標として定められ、効率的な字幕制作手段の確立が放送局の重要な課題のひとつになっている。特に、すべてのニュース番組に字幕を付与するためには、突発的な緊急ニュースに対応でき、かつ、要員の確保が困難な地方放送局においても運用可能な、簡易で低コストのリアルタイム字幕制作システムの開発が求められる。

前述のステノワード方式は、一般のパソコン等のキーボードのカナ入力やローマ字入力とは異なり、速記文字用にあらかじめ設定された文字列を、10 個のキーの組合せで一度に入力する特殊なものであり、この操作に熟達したオペレータを複数名必要とする。しかし、緊急ニュースへの対応には、24 時間体制の要員確保が必要であり、ステノワード方式のキーボード入力可能な人材を全国の放送局で常時確保することは、現実的には困難である。

一方、オペレータの人数や負荷の削減をねらいとした自動音声認識技術は、字幕放送を効率的に拡充する手段として、その進展が期待されている。しかし、これまでの音声認識技術では、背景雑音等による音響環境の変動や話し言葉口調等の不明瞭な発話スタイルの影響を受けると、高い認識精度を確保できないことがあった。そこで筆者らは、高い認識率が得られる条件の音声区間は直接認識し、それ以外はリスピーク方式で認識する、ハイブリッドな音声認識によるリアルタイム字幕制作システムを試作した。本システムの主な特徴は以下の通りである。

- ハイブリッド方式の採用により、字幕付与可能な範囲が広がり、音声認識だけでニュース番組の全編をカバーできるようになった。

- 音響モデルと言語モデルの高精度化、発話区間の検出性能の向上、入力音声の男女別自動判別機能の導入などにより、原稿読み上げだけでなく、記者による現場リポートなどの音声認識の性能を改善した。

- 一般的なキーボードとタッチパネルを利用した修正装置の機能改善により、従来は 4 名を必要としていたオペレータ数を、1~2 名に削減可能にした。この修正装置は操作の習熟が容易であり、人材確保の面でステノワード方式よりも有利である。

以下、2 章で本システムの概要を述べ、3 章では本システムによって行った生字幕制作実験の結果を、従来システムと比較して報告する。そして 4 章で課題を整理して考察し、最後に 5 章でまとめを述べる。

2. システムの概要

試作したハイブリッド型のリアルタイム字幕制作システムの概略を、図 1 に示す。本システムは、入力音声を切り替えるスイッチ、音声認識装置、および修正サーバと修正

端末からなる修正装置で構成される。本システムは放送局内に設置されることを想定しており、出力される字幕データは、放送局内で映像音声信号に多重された後、放送電波を介して受信機に向けて送信されることになる。

2.1 ハイブリッド方式

ニュース番組の場合、スタジオアナウンサの原稿読み上げや記者による現場リポート音声であれば、高い認識精度が得られる。このような区間は、番組音声を直接認識させる（ダイレクト方式）。一方、大きな背景雑音や、主として話し言葉に見られる不明瞭な発話を含む音声を認識させると、誤りが増加するため、このような音声区間では、専任の話者（リスピーカ）が番組の内容を復唱し、これを認識させる（リスピーク方式）。なお、このリスピーカは、アナウンサではなく、特別な発話の訓練を受けていない一般の人を想定する。そして、番組音声の性質に応じて、図 1 のスイッチをリスピーカが操作して、両方式の切り替えを行う。本稿では、これを“ハイブリッド方式”と呼ぶこととする。本方式により、字幕付与可能な範囲が広がり、音声認識だけでニュース番組の全編をカバーできるようになる。

2.2 音声切替

リスピーク音声は、番組音声に対して、時間的に常に遅れる性質があるため、リスピークが終了して番組音声側にスイッチが切り替えられた直後に、音声認識装置の入力となる次の番組音声区間の冒頭が脱落するおそれがある。よって、これを回避するため、常時約 1 秒間の音声をバッファリングしておき、番組音声側にスイッチが切り替えられた直後に、これを次の番組音声区間の冒頭に付加して認識させる機能を有している。なお、本システムのスイッチは、2 入力音声入力システムを有する音声認識装置の音響分析部において、ソフトウェア的に実現している。

2.3 音声認識装置

生字幕制作の実施にあたり、音声認識装置は、高い認識精度に加えて、入力音声に対する認識結果出力の遅れが蓄積しないこと（リアルタイム性^{*}）が求められる。筆者らが開発した音声認識装置は、音響モデルに性別および音素環境依存の HMM（隠れマルコフモデル）、言語モデルに電子化されたニュース原稿から学習した N グラムモデルを用い、第 1 パスでバイグラムによる単語依存 N-best 探索、第 2 パスでトライグラムによるリスクアリングを行うタイプのものである。リアルタイム性を実現するため、発話終了後に認識結果を確定するのではなく、発話の途中で、認識精度を有意に劣化することなく、認識結果を早期確定する機能をもつ⁹⁾。

筆者らが 2000 年に開発したシステム²⁾（旧システム）の音声認識装置（旧装置）に対して、今回開発した音声認識装置（新装置）において改良を行った内容を以下に列挙する。

^{*} 本論文では、音声からはほぼ一定の遅れで文字が出力され続けており、その遅れ時間がしだいに蓄積されない場合、「リアルタイム性」があるものとする。

図 1 ハイブリッド型リアルタイム字幕制作システムの概略
Overview of the hybrid real-time closed-captioning system.

- 音響特徴量の抽出の際に、低い変調周波数を抑圧することにより雑音等の影響を軽減させる RASTA(Relative SpecTrA) フィルタ¹⁰⁾を適用した。

- 音素認識による発話区間検出¹¹⁾を導入した。

- 男女並列連続音声認識機能¹¹⁾を導入したことにより、入力音声の男女別を自動判定し、性別のあった音響モデルの自動適用が可能になった。なお、旧装置では、男女別の音声認識装置をそれぞれ1台ずつ設置する必要があった。そのため、番組毎にレギュラ出演するアナウンサの性別は特定できるが、記者の性別は特定できず、男女別の自動判定機能をもたない旧装置では運用が煩雑となるという理由から、記者の発話の音声認識は事実上不可能であった。

- 音響モデルの学習データ量を増やし（旧装置：男性48.3時間・女性38.6時間、新装置：男性340時間・女性250時間）、音響モデルの高精細化を図った（旧装置：8混合ガウス分布HMM、新装置：16混合ガウス分布HMM）。

- 言語モデルの学習データ量を増やした（旧装置：約9年分のニュース原稿、新装置：約17年分のニュース原稿）。また、語彙サイズを大きくし（旧装置：20K、新装置60K）、未知語率の削減を図った。

- 旧装置は、Alpha21264 500MHz のCPUを搭載したUNIXマシンで動作していたが、新装置では、Xeon 3.60GHzとなり、処理速度が高速になった。

2.4 修正装置

修正装置は、大出らが開発した装置¹²⁾をベースに改良を加えたものであり、修正サーバと、1台ないし2台の修正端末から構成される。なお、旧システム²⁾の修正装置では、誤り発見用と誤り修正用の端末があり、これらを1セット

にした2台1組単位による運用が基本であったのに対し、本装置における修正端末の台数は、想定される音声認識精度に応じて可変にできるところに特徴がある。修正端末には、タッチパネルモニタが接続されており、この画面上には、音声認識装置が出力する単語列がリアルタイムで逐次表示される。オペレータ（修正者）は、ヘッドホン経由で認識出力に対応する音声を聞き、誤認識の単語を発見したら、これを指でタッチして選択し、キーボードによって速やかに修正する。また、正しい文字列であることを確認したら、速やかにテイクボタン（キーボード上のEnterキー）を押下して字幕となるテキスト（字幕データ）を送出していく。

修正端末における修正作業の様子を図2に、修正端末の画面構成を図3に示す。また、本修正装置の特徴を以下に列挙する。

- 旧システムの修正装置が、誤り発見用と修正用の2台1組単位を基本としていたのに対して、本装置では各修正端末単独で誤り発見と修正が可能になったため、充分認識率が高い場合には修正端末を1台とし、誤りが増えるにつれて修正端末を2台以上に増やしていくことができる。したがって、最低必要なオペレータは、旧システムの2名から本装置の1名へと削減可能となった。

- 画面に提示中の認識出力と、修正者がヘッドホンでモニタリングする発話内容を一致させるため、発話冒頭の単語を画面提示した後（約1秒後）に音声再生を開始させる。

- 修正者が音声を聞き直すことができる機能をもつ（聞き直し機能）。また、聞き直しによる時間遅れを取り戻すため、聞き直し終了直後に、話速変換機能¹³⁾を使用して音

声を高速再生し、実時間に追いつく機能をもつ（高速再生機能）。

- 各オペレータの修正担当範囲を明確化するため、自分側の担当範囲の文字色を黒色、相手側の担当範囲の文字色を白色にするよう画面表示を制御する。これにより、2名のオペレータが重複することなく、複数の誤り箇所を同時に修正することが可能となるため、旧システムのような誤り発見用端末と修正用端末を2台1組にしたシステムに比べて効率よく修正できる。

- 認識結果出力に遅れが生じた際に、相手側の担当範囲の音声を、話速変換機能を利用して高速再生することにより、音声再生の遅延の蓄積を防止する。

- 修正者の視線の移動を極力少なくするため、タッチした誤り単語の直下の画面エリアで文字列の修正を行う（インライン変換機能）。

- 修正者のワンタッチにより、単語単位の範囲で修正すべき文字列を一括選択できる。また、連続した複数の単語を一括選択する際には、開始単語と終了単語の2点をタッチするか、その範囲をドラッグすればよい。なお、単語区切りの表示を明確化するため、単語間に縦線を挿入する。

- 選択された単語または単語列の同音異義語リストを、修正候補として提示する。

- あらかじめ放送音声に現れることが予想できる重要な固有名詞や専門用語（想定キーワード）を画面上に表示し、ワンタッチで修正に利用できる。また、同一単語の認識誤りの繰り返しに対処するため、直前に修正した数単語（修正履歴）を画面上に表示し、ワンタッチで修正に利用できる。

- 放送に用いられる予定のニュース原稿を画面上に表示し、これを修正者が参照できる。また、その原稿に含まれる単語をワンタッチで修正に利用できる。

- 認識誤りが多く、修正作業が難航した場合であっても、字幕表示の遅れを蓄積しないようにするため、修正端末画面に提示された後に、送出されないまま一定時間（20秒に設定）が経過した単語は、同単語から発話終了までの範囲を「…」などの記号に置換し、強制的に字幕送出する（タイムアウト機能）。

なお、これまで4名の修正者を養成した結果、本端末の操作を習熟するためには、約1ヶ月の訓練期間が必要であることがわかっている。一方、ステノワード方式は、高速な日本語入力が可能とはいえ、速記文字用にあらかじめ設定された文字列に対応する複数キーの組合せ入力を習熟しなければならず、入力者の養成には2年から3年以上かかる¹⁴⁾とされている。本論文で提案する修正装置は、一般のパソコン等のキーボードと同じカナ入力やローマ字入力による操作であり、しかも最低1名で認識誤り箇所のみを修正すればよい。したがって、特殊なキーボード入力に熟達したオペレータを複数名必要とするステノワード方式に比べて、本修正装置は人員確保の面で有利である。

図2 修正端末における修正作業の様子。
Operation of error correction at a correction client.

図3 修正端末の画面。
Screen shot of a correction client.

3. リアルタイム字幕制作実験

3.1 実験準備

本システムの性能を評価するため、ニュース番組を用いてリアルタイム字幕制作実験を行った。本実験のオペレーションは、修正端末の操作法を十分に習熟した修正者2名と、特別な発話の訓練を受けていない一般の話者によるリスピーカ1名によって行うこととした。各人は、実験前の準備段階において以下の作業を行う。

(1) 言語モデルと発音辞書の作成

報道情報システム（記者が投稿したニュース原稿*を一括管理するサーバ）から、実験対象番組の放送開始時刻より遡って、6時間以内に出稿されたニュース原稿をダウンロードし、これにより言語モデルの適応化を行う¹⁵⁾。また、新出単語に対しては、発音を自動付与し、これを人手でチェックした後、発音辞書に追加する。

(2) オペレータ自身の事前学習と修正端末の準備

修正者とリスピーカは、実験の直前にニュース原稿を読んで、放送予定の内容を確認する。また修正者は、想定され

* ニュース番組でアナウンサーが読み上げる原稿は、記者が投稿し、ディレクターやアナウンサーが手書きで加筆修正を施したものである。したがって、その原稿をそのまま字幕にすることはできない。

表 1 話者および音響環境別の単語数
Number of words classified by speakers and acoustic conditions.

話者	単語数 (番組全体に占める割合)	
アナウンサ	1840 (62.6%)	clean: 1139 (38.7%) noisy: 701 (23.8%)
リポータ	856 (29.1%)	
リスピーカ	244 (8.3%)	
全体	2940 (100.0%)	

る単語をあらかじめキーボード入力することによって、修正端末のかな漢字変換の精度を上げる。さらに、修正端末に対して、想定キーワードの登録を行う。リスピーカは、番組構成表を見て、リスピークの実施が想定される項目を事前に確認する。

3.2 リスピークの条件

リスピークは、背景雑音が大きい音声区間、話し言葉口調の発話（インタビュー、記者会見など）、および、複数話者による対談区間において実施することを基本とした。また、リスピーカは、番組音声と認識出力を放送中に常時モニタし、認識精度が低下する条件下では、随時速やかに番組音声に割り込んでリスピークを行うこととした。一方、オープン・キャプション方式*による字幕とクロズド・キャプション方式による字幕の重複を避けるため、オープン・キャプション方式による字幕が存在する番組音声区間では、リスピークを行わないこととした。

3.3 ストレートニュースによる実験

最初の実験用の番組素材には、1名のアナウンサのバストショット画面でリード（各ニュース項目の導入部）が読み上げられ、VTRや中継映像に切り替わって本記（詳細）が読み上げられるパターンで構成される、典型的なニュース番組（ストレートニュース）を取り上げることとした。実際に用いたのは、2008年4月30日に放送された「NHKニュース7」（30分番組）である。なお、本実験は、研究所内の実験室内において行い、音声はオンエアされたものを用いた。放送局内では、アナウンサの音声は、SE（Sound Effect）を混入する前のものを直接認識させることが可能であるため、将来の実運用で想定される環境とはやや異なる。以下では、アナウンサの音声区間のうち、SEの混入がない区間を“clean”，SEの混入がある区間を“noisy”と区別して表記する。

本実験に用いた番組素材の単語数を、話者と音響環境別に分類した結果を表1に示す。旧システムで字幕付与が可能であったのは、アナウンサの原稿読み上げ区間の62.6%であったが、本システムでは、リポータ（記者と気象予報士）の発話区間を含めて番組全体100.0%の字幕付与が可能になった。したがって、一番組中で字幕を付与できる範囲の割合（カバー率）は、約1.6倍に向上したことになる。

本システムの音声認識装置（新装置）と、旧システムの

表 2 新旧装置による音声認識性能比較と修正結果（単語誤り率）
Comparison between new and conventional speech recognizers and post-correction results (WER).

話者	旧装置		新装置		修正後
アナウンサ	7.6%	clean: 5.0% noisy: 11.8%	3.1%	clean: 2.0% noisy: 4.9%	0.0%
リポータ	5.8%		2.1%		0.0%
リスピーカ	20.2%		15.6%		0.1%
全体	8.2%		3.8%		0.0%

音声認識装置（旧装置）の性能を単語誤り率（Word Error Rate: WER）によって比較した結果と、新装置によるリアルタイム字幕制作実験で得られた修正後のWERを表2に示す。なお、旧装置のデータは、実験終了後に保存した番組音声を用いて得たものである。本表から、番組全体で見た場合のWERは、旧装置では8.2%であったのに対し、新装置では3.8%であったことから、WERの削減率は53.7%となり、大きな改善が得られた。ただし、リスピーカによる発話区間（リスピーク区間）だけで見ると15.6%で、改善の余地が残る結果であった。また、新装置による認識結果をリアルタイムで修正した後のWERは0.1%であった。具体的には、リスピーク区間で「ついに」という1単語が修正時に脱落しただけであり、文意を損ねる致命的な誤りは見られなかった。

リスピークの発話内容を調査したところ、実験に用いた番組素材のうちリスピーク区間は11文からなり、このうち9文において正確なリスピークが実行された。また、残りの2文において一部の単語の脱落が存在したが、文意を損ねるような違いは見られなかった。

入力音声に対する字幕提示の遅れは、誤りの修正がない場合、平均約5秒（音声認識による遅れが約1秒、修正者の確認と送出で約4秒）で、現行の放送に用いられている高速入力用キーボード方式による遅れと比較して、有意な差は見られなかった。連続した認識誤りの発生に伴って修正作業が行われた際には、やや遅れが目立つところも見られたが、本実験では、2.4節で述べた修正作業が難航した際に強制送出を行う「タイムアウト機能」は動作することにはなかった。修正作業により字幕送出の遅れが一時的に蓄積しても、その後の音声認識誤りのない区間では、修正者は未確定で蓄積している音声認識結果に対して、テイクボタンを押下して送出するだけであるため、字幕送出の遅れを解消することができる。したがって、修正装置の出力においても、音声認識装置の出力と同様に遅れが蓄積せず、リアルタイム性の確保が認められた。

以上の結果より、リスピーク区間に改善の余地が残るものの、修正装置がもつ多様な機能の有効活用によって、正確でかつ遅れの少ない生字幕送出が可能であったことから、ストレートニュースにおける本システムの有効性が確認できた。

* 文字列を映像上にスーパーインポーズする方式。テロップとも呼ばれる。

3.4 大型ニュース番組（ニュースショー）での性能

前節では、本システムがストレートニュースの生字幕制作において有効であることを示したが、本節では、より高度な演出形態をもつ大型ニュース番組（ニュースショー）への適用の可能性を調査した結果を報告する。

ニュースショーは、リードと本記を繰り返すストレートニュース的な演出形態に加えて、男女複数のアナウンサ（キャスタ）、記者、ゲストなどの多人数の話者が出演し、解説、対談、現場中継などを織り交ぜて構成されるニュース番組のことを指す。たとえば、「NHK ニュースウオッチ 9」（60分番組）では、番組中の約 5 割がアナウンサの原稿読み上げ、約 1 割が記者の現場リポートであるが、残りの約 4 割において、大きな背景音や、主に話し言葉に見られる不明瞭な発話が含まれる。

この番組によって字幕制作実験を試みた結果、ストレートニュース的な演出形態の区間は、前節と同等水準の音声認識精度と修正精度が得られるが、解説や現場中継の区間では、音声認識精度が劣化し、2 名の修正者でも修正が困難となるケースが見られた。

話し言葉口調の解説や対談区間は、現状ではリスピーク方式で対応することになるが、「NHK ニュースウオッチ 9」では、60 分番組の約 4 割をリスピークしなければならないことを考慮すると、表 2 に示されるように、リスピーク区間の認識精度が充分でないため、これが長時間続くと、修正に時間がかかって字幕送達の遅れが蓄積し、最終的にはタイムアウト機能が動作することになる。また、番組音声の完全な自由発話になると、番組音声にリスピークが追従できなくなり、生字幕制作が不能になる現象も発生する。そこで、特にリスピークが困難となりやすいニュース項目の末尾で行われるアナウンサ（キャスタ）のコメント（後受け）区間の WER を、「ニュースウオッチ 9」で調査した¹⁶⁾（表 3）。この結果において、ダイレクト方式を用いた場合には、WER が 41.1%とさきわめて高いため、修正作業が不能となり、また、リスピーク方式を用いた場合には、ストレートニュースとほぼ同等の 16.1%の WER が得られるが、番組音声にリスピークが追従できず、単語の脱落が生じて文意を損ね、誤りの修正も困難となった。

この原因を調べるため、あるニュース項目における原稿読み上げ区間とコメント区間の話速とパープレキシティ（平均単語分岐数）の比較を行った（表 4）。これより、コメント区間で音声認識が困難になるのは、話速が速いことと、発話内容と言語モデルのミスマッチに原因があることが明らかになった。また、話速が速いことが、リスピークで番組音声に追従できなくなる原因にもなっていると考えられる。発話内容と言語モデルのミスマッチが生じる原因としては、発話スタイルが原稿読み上げ口調ではなく、口語調となりやすいことに加えて、ストレートニュースに比べて、それぞれのニュース内容を深く掘り下げるため、一般的なニュース番組において多用されない言い回しや単語が多く

表 3 コメント区間の音声認識性能と修正結果（単語誤り率）
Recognition and post-correction results of commentary speech segments (WER).

方式	単語数	音声認識出力	修正後
ダイレクト方式	299	41.1%	修正不能
リスピーク方式	248	16.1%	2.0%

表 4 話速とパープレキシティの比較
Comparison of speech rate and perplexity between read and commentary speech.

発話区間	話速	パープレキシティ
原稿読み上げ	7.2 mora/s	15.8
コメント	9.2 mora/s	115.0

出現するためであると考えられる。

以上から、本システムをニュースショーの字幕制作に適用するためには、速い話速への対策と、言語モデルのさらなる高精度化が重要な課題といえる。

4. 考察と課題

4.1 ハイブリッド方式の効果

3.3 節の実験結果から、ストレートニュースにおいて、正確でかつ遅れの少ない生字幕制作が可能であることを確認した。少人数のオペレータによってニュース番組全編の字幕付与を実現できた理由としては、2.3 節で述べた音声認識装置の音響特徴量、音響モデル、言語モデルの改善の結果として得られた認識精度の向上と、2.4 節で述べた修正装置の機能の改善、および、ハイブリッド方式の採用によるところが大きいと考えられる。

ハイブリッド方式の有効性は、次のように解釈できる。3.3 節の実験において、話者がリスピーカとなった区間、すなわち、リスピークが実行された割合は 8.3%であり、その発話内容には、わずかな単語の脱落は存在したものの、番組音声に対して文意を損ねるような違いは見られなかった。一方、もしニュース番組全体をリスピーク方式^{4)~6)}だけで字幕化しようとする、十分な発話の訓練を受けたアナウンサの原稿読み上げ部分は、一般の話者が長時間正確に復唱し続けることが困難であり、単語の脱落や言い誤りの生じるおそれがある。したがって、音声認識精度がそもそも高い区間では、ダイレクト方式が有効である。そして、音声認識が困難な部分だけをリスピーク方式とするハイブリッド方式により、ニュース番組全編に対する字幕付与が可能となる。

以上の効果により、ステノワード方式や筆者らの旧システムと比較した場合、約 1ヶ月の訓練期間で養成できる修正者 2 名と、特別な発話の訓練を受けていないリスピーカ 1 名で運用できることから、オペレータの人材確保や運用コストの面に関して有利となった。さらに、男女並列連続音声認識機能の導入により、旧システムにおいて男女別に 1 台ずつ必要としていた音声認識装置が 1 台に削減されたこと、および旧システムにおいて誤り発見端末と修正端末

に分かれていた修正装置を一体化したことにより、システム全体の設備規模もコンパクトになった。

4.2 リスピークの課題

一方で、リスピーク区間の音声認識精度に改善の余地があることや、ニュースショーにおいて生字幕制作が困難となることも明らかになった。

3.3節の実験のリスピーク区間においてWERが15.6%と高くなる原因としては、認識対象がインタビューや記者会見などの話し言葉であるのに対し、言語モデルは、主に書き言葉からなるニュース原稿で学習しているため、認識対象と言語モデルにミスマッチがあることが挙げられる。また、リスピーカに特別な発話の訓練を受けていない話者を想定しているため、アナウンサに比べて発声がやや不明瞭である影響なども考えられる。これらの改善に際しては、リスピーカの発話の際に、話し言葉で学習した言語モデルや、話者適応を施した音響モデルに切り替えることが有効であると推察される。

なお、ストレートニュースでは、表1からもわかるとおり、リスピーク方式への切り替えを要する番組音声区間（リスピーク区間）の比率は小さい。このため、音声認識精度が低い状況であっても、修正者は修正端末の聞き直し機能を利用して、正確な修正作業を遂行できる。また、高速再生機能を有しているため、修正作業に伴う字幕送出の遅延を取り戻すこともできる。ところが、ニュースショーではリスピーク区間の比率が大きくなるため、聞き直し機能を繰り返し利用して修正すると字幕送出の遅れが蓄積し、タイムアウト機能の動作に至るケースも見られることから、現状の性能では完全な生字幕制作は困難である。

また、リスピークの発話内容の精度、すなわち、リスピーカが正確に番組音声を復唱できるかという点についても考慮する必要がある。ストレートニュースでは、リスピークの際に見られたわずかな単語の脱落が、文意を損ねることはなかったが、ニュースショーでは、話速の速さなどの影響から、番組音声にリスピークが追従できなくなる問題が生じた。リスピーク精度の改善策には、リスピーカのトレーニングが挙げられるが、人間の音声認識能力を測定したデータ¹⁷⁾などを鑑みると、100%の精度の保証はできない。今後、背景雑音や口語調の発話スタイルにも頑健な音声認識の研究を推進し、将来的にはリスピーク方式からダイレクト方式に順次移行させていきたいと考えている。

5. む す び

新たに試作したリアルタイム字幕制作システムは、番組音声の「ダイレクト認識方式」と、リスピーカが番組音声を言い直した音声を認識する「リスピーク方式」の切り換えによる、ハイブリッド型の音声認識を採用した。また、従来のシステムでは、認識誤りを修正するオペレータの数は4名であったが、音声認識装置と修正装置の改良によって、これを1名または2名に削減した。本システムによって、

簡易な演出のニュース番組（ストレートニュース）の生字幕制作実験を行ったところ、1名のリスピーカと2名の修正者によって、正確でかつ遅れの少ない字幕送出が可能であることを確認した。また、本システムを高度な演出形態の大型ニュース番組（ニュースショー）に適用するための課題を整理した。

将来の実運用へ向けて、すでに本システムの検証作業を始めている。放送局内に本システムを設置して、入力音声にSEの混入がない環境を構築し、5～12分のストレートニュース8番組を素材として生字幕制作テストを行ったところ、修正者が1名であっても、正確でかつ遅れの少ない生字幕制作を実現できることが確認された。今後もさらなる検証作業を継続することによって、システムの信頼性を確認するとともに、突発的に発生する緊急ニュース番組や、よりコンパクトな運用環境が求められる地方放送局発のニュース番組（ローカルニュース）への適用の可能性についても検証を行っていく予定である。

本システムが、生字幕放送番組のいっそうの拡充に寄与できるよう、今後もいっそうの改善を図っていきたい。

〔文 献〕

- 1) 今井亨：“音声認識を利用した情報保障～放送・講演の字幕制作の取り組み～”，音響学会講演論文集，2-8-12 (Mar. 2007)
- 2) 安藤，今井，小林，本間，後藤，清山，三島，小早川，佐藤，尾上，世木，今井，松井，中村，田中，都木，宮坂，磯野：“音声認識を利用した放送用ニュース字幕制作システム”，信学論，J84-D-II，6，pp. 877-887 (June 2001)
- 3) 西川，高橋，小林，石原，柴田：“聴覚障害者のためのリアルタイム字幕表示システム”，信学論，J78-D-II，11，pp. 1589-1597 (Nov. 1995)
- 4) 松井，本間，小早川，尾上，佐藤，今井，安藤：“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”，信学論，J87-D-II，2，pp. 427-435 (Feb. 2004)
- 5) 鈴木，菊地，坂口：“リスピーク音声切り分け・作業分散（わんこそば）方式によるリアルタイム字幕制作システム”，第2回デジタルコンテンツシンポジウム，8-7 (June. 2006)
- 6) M. Marks: “A Distributed Live Subtitling System”，BBC R&D White Paper (Sep. 1989)
- 7) 総務省：“報道資料 平成18年度の字幕放送等の実績”，<http://www.soumu.go.jp/s-news/2007/070629.9.html> (June 2007)
- 8) 総務省：“報道資料 視聴覚障害者向け放送普及行政の指針の公表”，<http://www.soumu.go.jp/s-news/2007/071030.2.html> (Oct. 2007)
- 9) 今井，田中，安藤，磯野：“最ゆう単語列逐次比較による音声認識結果の早期確定”，信学論，J84-D-II，9，pp. 1942-1949 (Sep. 2001)
- 10) H. Hermansky and N. Morgan: “RASTA Processing of Speech”，IEEE Trans. on Speech and Audio Processing，2，4，pp. 578-589 (Oct. 1994)
- 11) T. Imai, S. Sato, A. Kobayashi, K. Onoe, and S. Homma: “Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News”，IEICE Trans. Inf. & Syst., E90-D，8，pp. 1286-1291 (Aug. 2007)
- 12) 大出，三島，江本，今井，都木：“効率的なリアルタイムニュース字幕修正システム”，2003年映情学年次大，6-3 (Aug. 2003)
- 13) 清山，今井，三島，都木，宮坂：“高品質リアルタイム話速変換システムの開発”，信学論，J84-D-II，6，pp. 918-926 (June 2001)
- 14) 総務省：“報道資料 「デジタル放送時代の視聴覚障害者向け放送に関する研究会」報告書の公表”，<http://www.soumu.go.jp/s-news/2007/070330.19.html> (Mar. 2007)
- 15) 小林，今井，安藤，中村：“ニュース音声認識のための時期依存言語モデル”，情処学論，40，4，pp. 1421-1429 (Apr. 1999)
- 16) 本間，尾上，小林，佐藤，今井，都木：“番組音声とリスピーク音声の自動認識を利用したニュース字幕制作実験”，2007年映情学年次大，11-2 (Aug. 2007)
- 17) 中川：“音声認識研究の動向”，信学論，J83-D-II，2，pp. 433-457 (Feb. 2000)

ほんま しんいち
本間 真一 1992 年，早稲田大学理工学部電子通信学科卒業。同年 NHK 入局。新潟放送局，技術局を経て，1998 年より放送技術研究所に勤務。音声認識の研究に従事。正会員。

こばやし あきひろ
小林 彰夫 1991 年，早稲田大学理工学部電気工学科卒業。同年 NHK 入局。岡山，広島放送局を経て，1996 年より放送技術研究所に勤務。音声認識の研究に従事。正会員。

おく たかひろ
奥 貴裕 2001 年，慶應義塾大学理工学部システムデザイン工学科卒業。2003 年，同大学院修士課程修了。同年 NHK 入局。放送技術局を経て，2007 年より放送技術研究所に勤務。音声認識の研究に従事。

さとう しろう
佐藤 庄衛 1991 年，東北大学工学部電気工学科卒業。1993 年，同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に勤務。音声認識の研究に従事。博士（工学）。

いまい とおる
今井 亨 1987 年，早稲田大学理工学部電気工学科卒業。同年 NHK 入局。大阪放送局を経て，1990 年より放送技術研究所に勤務。音声認識の研究に従事。博士（情報科学）。正会員。

たかぎ とおる
都木 徹 1979 年，電気通信大学電気通信学部通信工学科卒業。1981 年，同大学院修士課程修了。同年，NHK 入局。長野放送局を経て，1984 年より，放送技術研究所に勤務。声質変換方式，音声知覚，話速変換方式などの研究に従事。博士（工学）。正会員。