Vol.2011-SLP-88 No.4
2011/10/28

音声認識によるリアルタイム字幕放送の進展

今井亨[†]　奥貴裕[†]　小林彰夫[†]

テレビ番組の音声を文字で伝える字幕放送は，聴覚障害者や高齢者への重要な情報保障手段の一つである．1985 年の字幕放送開始以来，リアルタイムの日本語文字入力方法が確立されていなかったため，字幕が付与される番組は長い間事前収録番組に限られていた．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を進め，世界に先駆けて 2000 年に音声認識によるニュース番組のリアルタイム字幕放送を開始した．番組音声を直接認識する本ダイレクト方式の実用化後，スポーツ番組の実況アナウンス等の復唱音声を認識するリスピーク方式の字幕制作システムの実用化などにより，リアルタイム字幕放送は年々拡充されるようになった．また，両方式を併用して認識性能と運用性を高めた，ハイブリッド方式のニュース番組用字幕制作システムの実用化も，現在検討を進めている．本稿では，字幕放送の現状と音声認識を利用した各種字幕制作システムを紹介するとともに，その技術的特徴と実用化の経緯について述べる．

Advances in Real-Time Closed-Captioning for Live Broadcast by Speech Recognition

Toru Imai[†]　Takahiro Oku[†]　and Akio Kobayashi[†]

Closed-captioning for broadcast, which displays spoken words as texts on the TV screen, is one of important media for the hearing impaired and the elderly. Since starting in 1985, closed-captioning has been provided only to prerecorded TV programs due to lack of a real-time input method of Japanese texts. NHK has done extensive research on speech recognition for news with other research institutes and led the world in real-time closed-captioning for broadcast news by speech recognition in 2000. Besides the direct method recognizing the original program sound, NHK realized a re-speaking method where rephrased utterances by another speaker are recognized for captioning of sports programs, resulting in expansion of live closed-captioning every year. Also a new hybrid method combined with both methods will be put into practical use for more accurate and efficient captioning of news programs soon. This paper introduces current situation of closed-captioning for live broadcast and the real-time closed-captioning systems with their technological features and the ways how they were implemented.

1. はじめに

NHK（日本放送協会）は，視聴者からの受信料を主な事業収入として，公共の福祉のために豊かで良い番組を放送するだけでなく，放送技術の研究を業務のひとつとしている[*1]．NHK放送技術研究所（技研）は，将来の新しいメディアやサービスの研究のほか，障害者や高齢者など，すべての人が好みの手段で快適に利用できる「人にやさしい放送」の研究開発[1)]を進めている．その研究成果は，NHKの番組制作に利用されるとともに，メーカーへの特許等の実施許諾，技術協力や共同開発による番組制作機器・放送受信機器を通じて，社会還元されている．

人にやさしい放送のうち，テレビ番組のナレーションやせりふなどの音声を文字で伝える字幕放送は，テレビの音が聞き取りにくい高齢の方や聴覚に障害のある方への重要な情報保障手段となっている．1983年NHK連続テレビ小説「おしん」での実験放送および1985年の本放送以来，字幕放送は長い間ドラマやドキュメンタリーなど放送前に完成している事前収録番組に限られ，生放送へのリアルタイム字幕付与は技術的に困難であった．1990年代には，アメリカにおけるテレビの字幕デコーダー内蔵義務化や高速入力用キーボードによる生放送への字幕付与の普及もあり，日本でも聴覚障害者からニュースなど生放送への字幕を要望する声が高まった[*2*3]．

当時，欧米ではDARPA（米国防総省高等研究計画局）を中心とする英語ニュースの音声認識の研究が盛んに行われており，1996年に技研は日本語ニュース番組のリアルタイム字幕付与を目指して，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と連携した「ニュース音声認識プロジェクト」を発足させた．その後の研究の進展により，スタジオ・アナウンサー部分の目標認識率95%以上が達成され，2000年3月末から世界に先駆けて音声認識によるニュース番組のリアルタイム字幕放送が開始された．アナウンサーの原稿読み上げ部分を直接認識するこのダイレクト方式の実用化後，高速入力用キーボード[2)]による原稿読み上げ部分以外の字幕化や，スポーツ番組の実況アナウンス等の復唱音声を認識するリスピーク方式の字幕制作システムの実用化により，リアルタイム字幕放送は年々拡充されるようになった[3)]．また，両方式を併用して認識性能と運用性を高めたハイブリッド方式のニュース番組用字幕制作システムの実用化も，現在検討を進めている．

[†] NHK 放送技術研究所
NHK Science and Technology Research Laboratories

*1 放送法第二十条「協会は，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．」
*2 放送法第四条の２「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又は図形を見ることが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．」
*3 1999 年 9 月に発生した東海村臨界事故では，聴覚障害者は字幕のないテレビの報道内容が十分わからず，不安が大きかったとして，緊急災害時の字幕の重要性が指摘された．

本稿では，NHKにおける音声認識によるリアルタイム字幕放送について，これまでに実用化された各種方式を紹介し，その技術的特徴と実用化の経緯を述べる．

2. 字幕放送

2.1 クローズド・キャプション

字幕放送（クローズド・キャプション）は，文字が映像上にすでにスーパー・インポーズされているオープン・キャプションと異なり，文字がテレビジョン信号の映像以外の領域に多重されていて，視聴者の好みで表示する／しないの選択が可能である．アナログ放送では，字幕を見るために専用の文字放送受信機が必要だったが，デジタル放送では字幕表示機能が受信機に標準装備されるようになり，リモコンの「字幕」ボタンを押せば字幕が画面に表示される．ワンセグ携帯端末も字幕表示が可能で，電車の中など音を出せない場所では，字幕は健聴者にとっても有効であり，まさにユニバーサルな情報伝達手段となっている．

2.2 字幕化率

字幕放送の拡充は年々進んでいるが，まだすべての番組に字幕が付与されるには至っていない．総務省のまとめによれば，2010年度の総放送時間に占める字幕放送時間の割合は，NHK総合テレビ（デジタル）が56.2%，民放在京キー５局平均（同）が43.8%となっている[4]．字幕放送普及のための行政指針として，総務省が2007年に策定した字幕放送普及行政の指針[5]では，7～24時の生放送を含む字幕付与可能な全ての放送番組（討論番組など複数人が同時に会話を行う生番組は除く）に，2017年度までに字幕付与することを目標としている．NHK総合テレビの2010年度の目標対象番組に対する字幕番組の割合は62.2%であり[4]，目標達成へ向けて効率的な字幕制作の取り組みを進めている．

2.3 字幕制作手段

放送番組には，ドラマやドキュメンタリーなど，放送前に完成している事前収録番組と，ニュースやスポーツなどの生放送番組があり，それぞれ字幕制作手段は異なる．NHK 総合テレビ（7～24 時）の場合，事前収録番組は平均で全体の約４割を占め，すべてに字幕がついている．事前収録番組に対する「オフライン字幕」は，パソコンを利用して人手で文字入力されるほか，字幕表示位置とタイミング，文字色などの調整が入念に行われる．一方，生放送番組に対する「生(なま)字幕」の制作には，リアルタイムでの文字入力が必要になる．アメリカの放送局では，裁判所の速記入力用に開発されたキーボードで，専門のオペレーターが１人で字幕を入力することが一般的だが，日本語は同音異義語が多く仮名漢字変換を要するため，人の話す速さで一般のパソコンから文字入力することは困難となる．

そこで，NHK では現在，キーボードと音声認識を利用した次の４つのリアルタイム字幕制作方式を，番組の性質に応じて使い分けている．

(1) 一般的なキーボードを利用した連携入力方式[2]で，複数の入力者が短い単位で文字をリレー入力する方式【歌謡番組などで利用】
(2) 複数キーを同時押下する特殊な高速入力用キーボード（ステノワード[6]）による方式，すなわち入力者と校正者のペア数組が短い単位で分担して入力するリレー方式【ニュースなど報道番組で利用】
(3) 番組音声を直接認識する，ダイレクト方式の音声認識【ニュース[7]（2000～2006年）と大リーグで利用】
(4) 字幕専用アナウンサーが言い直した音声を認識する，リスピーク方式[8]の音声認識【大相撲やプロ野球などスポーツと情報番組で利用】

キーボード入力方式は，番組の話題や発話スタイル，背景雑音等によらずオペレーターが柔軟に文字入力することが可能だが，熟練した専門オペレーターを複数必要とする．一方音声認識による方式は，番組ごとの辞書の事前学習や発話スタイルおよび背景雑音等の影響を受けることから，適用可能な番組は限られるものの，リスピークや誤り修正の人材を比較的確保しやすいという利点がある．

3. 音声認識によるリアルタイム字幕制作システム

3.1 ダイレクト方式

前述のように，技研は他の研究機関とも連携して，1996年からニュース音声認識の研究を開始した．ニュース番組におけるアナウンサーの音声と原稿を大量に収集してニュース音声データベースを構築し，ニュース向け音響モデルと言語モデルの学習法や，音声認識手法の高精度化の研究を進めた[9]．その結果，ニュースのメイン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が目標の95%[*4]を達成し，世界に先駆けて2000年に音声認識によるニュース番組のリアルタイム字幕放送[7]を開始することとなった[*5]．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．放送での音声認識の実用化は，ニュース音声認識プロジェクトを始め，日頃熱心にご議論いただく音声認識コミュニティーの成果であったといえる[*6]．

このニュースのメインアナウンサー原稿読み上げ向け字幕制作システムは，図１に

*4 平均 40 単語のニュース文は 12 秒で発声されるので，認識率 95%は平均 6 秒に 1 単語の誤りに相当し，リアルタイムで修正できる範囲だが，認識率 90%の平均 3 秒に 1 単語の誤りはリアルタイム修正が困難になる．
*5 1999 年 11 月の衆議院逓信委員会で，NHK 会長が翌年度からニュ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．準備期間が短いことから，誤り修正部分はメーカーと共同で開発したものの，音声認識部分は技研のシステムをそのまま利用して運用を開始した．
*6 「NHK の音声認識を用いた字幕放送の実用化は我々音声認識研究者の努力の勝利と言えよう」中川聖一；“特集－音響学における 20 世紀の成果と 21 世紀に残された課題 第 1 部 4.音声認識”，日本音響学会誌 57 巻 1 号，pp.17 (2001)

Vol.2011-SLP-88 No.4
2011/10/28

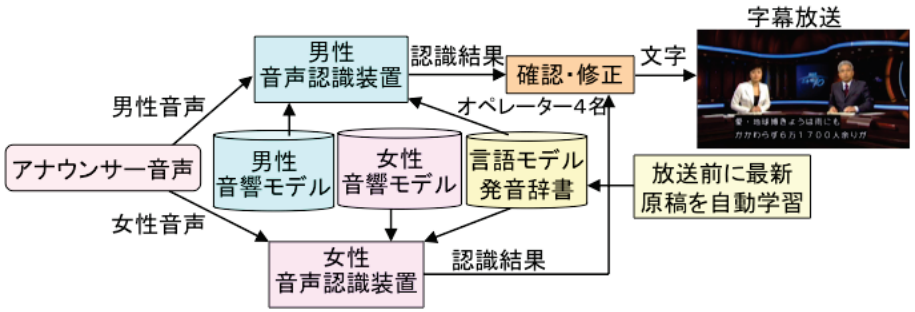

図１　ダイレクト方式音声認識によるニュース番組の字幕制作（2000～2006 年運用）

図２　ニュース字幕制作における誤り修正（誤り発見者と修正者のペア２組）

示すように男女別々の音声認識装置で構成され，認識結果の確認と修正を計４名のオペレーター（誤り発見者と修正者のペア２組）が担当した（図２）．音声認識結果はリアルタイムに確認され，誤りがあればタッチパネルやキーボードで即座に修正の後，字幕が送出された．音声認識のデコーダーは，音響モデルに性別および音素環境依存のHMM，言語モデルにNグラムモデル（語彙サイズ約２万単語）を用い，第１パスでバイグラムによる単語依存N-best探索（複数の直前単語でひとつの静的な木構造化単語辞書を共有して探索）を行いながら，発話途中で繰り返し第２パスのトライグラム・リスコアリングを実行する逐次２パスデコーダー[10]とした．言語モデルは，放送直前に最新のニュース原稿（読み原稿用に手書きで加筆修正される前の電子原稿）を自動的に取得し，重み付け学習した[11]．

ニュース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た．字幕表示は画面下に最大 15 文字×２行で，3.5 秒おきのページ更新を基本としたが，平均約 10 秒の字幕表示までの遅れ時間（音声認識に１秒以内＋確認・修正に５～６秒＋ページ更新のためのバッファリング時間）や，すでにスーパー・インポーズされているオープン・キャプションとの文字の重なりなどの課題が残った．

メインアナウンサー原稿読み上げ部分の字幕放送開始の翌年には，特殊な高速入力用キーボード（ステノワード[6]）による方式を，原稿読み上げ以外の中継や対談項目で採用し，ニュース項目によって制作手段を使い分けつつ字幕対象番組を拡充した．しかし，音声認識は原稿読み上げ部分の限定的な運用だったこともあり，経費削減のため2006年からNHKのニュース番組は高速入力用キーボードによる字幕制作方式に一本化されている．なお，従来よりも少人数での運用が可能で認識性能を高めた，後述するハイブリッド方式の字幕制作システムの実用化を今後予定している．

3.2　リスピーク方式

ニュースの字幕放送開始後，スポーツや音楽・情報番組の生放送にも字幕を付与したいとの要望があり，多様な話者の発話スタイルや背景雑音等の影響を考慮して，字幕専用のアナウンサー（字幕キャスター）がヘッドホンで番組音声を聞きながら，基本的には番組中の実況アナウンサーや解説者の言葉を復唱（場合によっては内容を要約）し，その音声を認識する「リスピーク方式[*7]」を採用することとした（図３）．NHKは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[12]，2001年12月の紅白歌合戦以来，大相撲，プロ野球，オリンピック，W杯サッカー，情報番組などを字幕化している[13]．リスピーク方式によれば，背景雑音が大きく，出演者が複数いるような番組でも，静かなスタジオで字幕キャスターが１人で内容をアレンジして発話することにより，字幕付与が可能である．スポーツ番組では，字幕の表示遅れ（５～10秒）をなるべく小さくし，読んでわかりやすい字幕とす

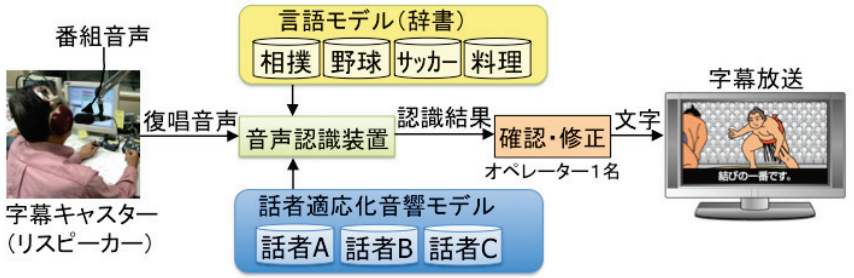

図３　リスピーク方式音声認識によるスポーツ番組等の字幕制作（運用中）

*7 リスピーク方式は，英国 BBC が 2001 年 9 月にはスヌーカー，ゴルフ，テニス，陸上競技の４種目で字幕放送を実施しており，現在は英国など欧州や南米の放送局での生字幕制作方式の主流になっている．

るために，字幕キャスターは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[8]，実況アナウンサーが説明しない拍手や歓声など場内の様子を補足するなどしている．音声認識の誤り部分は，字幕キャスターが再発声して再度認識させることもあるため，誤り修正者は１人で対応している（平均認識率90～95%）．

音声認識で用いる言語モデルは，番組ジャンルごとに学習テキストを用意し，各番組専用に学習している（語彙サイズ約 10 万単語）．音響モデルは，字幕キャスターごとの適応学習によって，認識率の向上を図っている．リスピーカー１人での復唱作業は疲労を伴うことから，数 10 分ごとに別のリスピーカーへ交替することが通常であるが，生放送中の交代では音声認識のオンライン処理を中断することなく，音響モデルを即座に切り替えられるようになっている．なお，大リーグの実況アナウンスを日本のスタジオでつける場合には，リスピーク方式ではなくダイレクト方式で字幕を制作している．

言語モデルの学習については，ニュース番組では記者が入稿した電子原稿を学習に利用できるものの，他の生放送番組（特にワイドショーなどの情報番組）では番組構成表１枚～数枚程度の事前情報しか得られないことも多く，学習用関連テキストの収集に現状では手間を要している．そのため，現状では音声認識で対応可能な番組は，言語モデルの学習に十分なテキストが得られるか，定期的に放送することで学習データの収集コストを結果的に低くできる番組に限られる．字幕放送をさまざまな番組に拡充していくためには，語彙の選定，新出単語の登録や学習テキストの収集，低出現頻度単語の対策など，音声認識の専門家でなくとも効率よく事前準備できるようにすることが，今後の課題となっている．

3.3 ハイブリッド方式

ニュース番組の場合，現状の音声認識技術で，スタジオ・アナウンサーの原稿読み上げや記者による現場リポート，さらにアナウンサーと記者の落ち着いた対談の認識性能は，実用化レベル（平均単語認識率 95%以上）にある．一方，それ以外のインタビュー部分や編集済みビデオ素材，自由発話の多い対談などは，大きな背景音，音楽，不明瞭な発声，速い話速，くだけた話し言葉などの影響により，認識性能が低下する．

そこで，十分高い認識率が得られる条件の音声を直接認識するダイレクト方式と，それ以外のインタビュー部分などをリスピーカーが復唱して認識するリスピーク方式を併用し，新たなハイブリッド方式のリアルタイム字幕制作システム[14] の開発を進めてきた（図４）．音声認識の入力音声は，ニュース項目に応じてリスピーカー自身が番組音声とリスピーク音声の間を手動で切り替えるが，切り替え時に発話冒頭が欠落しないよう，音声をバッファリングしている．認識結果の確認・修正は，インターフェースの改善によって複数の修正オペレーターの同時作業を可能とし[15]，操作性の向上と，番組の難しさに応じた修正者数の削減（１～２名）を実現した（図５）．ハイブリッド方式によれば，従来困難だった音声認識のみによるニュース番組全体の字幕付与

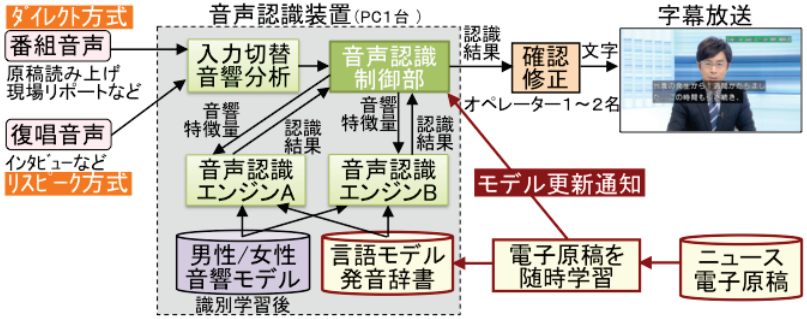

図４　ハイブリッド方式音声認識によるニュース番組の字幕制作（運用予定）

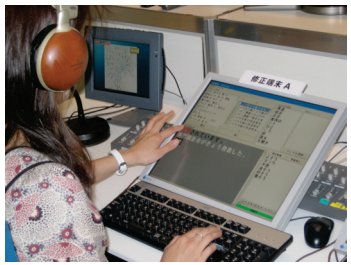

図５　認識誤り修正作業の様子（修正オペレーター１～２名）

図６　東日本大震災ニュースでの字幕制作の様子

が可能になり，さらにエンドレス音素認識による音声区間検出や男女並列連続音声認識[16]，音響モデルの選択的識別学習[17-18]や放送中の言語モデル自動更新対応[19]など，最近の音声認識の要素技術の改善によって，認識性能と運用性が向上した．本方式は，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として期待されており，総務省の字幕放送普及目標を確実に達成するため，今後ニュース番組での実用化が予定されている．また，2011年3月の東日本大震災直後のニュース番組では，字幕による情報提供を充実させるため，技研で試作した本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による字幕放送を実施した（図６）．

4. おわりに

本稿では，NHK の生放送番組の字幕をリアルタイムで制作するため，技研で研究開発した音声認識技術をベースに実用化された，各種字幕制作システムの概要を紹介し，その技術的特徴と実用化の経緯について述べた．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が，その実用化のポイントは，
(1) 代替手段が限られる中での，実用化に対する視聴者，国，放送局の強い要望
(2) インハウス（企業内）導入におけるチャレンジングな意思決定
(3) 最初から完璧を目指すのではなく，できるところからの実現
(4) 全自動ではなく，人手による確認・修正やリスピーカーを含めたシステム設計
にあったと考えている．

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，現状の音声認識技術は残念ながらどのような音響条件の番組でも字幕化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，話題が多様に変化する情報番組（ワイドショー）の音声認識性能の改善および事前準備の効率化，現在リスピーク方式で対応している番組をダイレクト方式へ移行していくための改善研究，ローカル放送局でも運用可能ないっそう効率的で運用コストの低い字幕制作システムの開発などを，引き続き検討していきたいと考えている．

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します．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，豊橋技術科学大学・中川聖一教授の諸先生方にご指導をいただきました．ここに深く感謝の意を表します．

参考文献

1) 伊藤崇之：高齢者・障害者のメディアアクセスに関する話題－人に優しい放送をめざした研究開発－，信学会２種研究会，サイバーワールド第９回研究会，pp.1-6 (2008).
2) 総務省，三菱総研：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，http://www.soumu.go.jp/main_sosiki/joho_tsusin/b_free/pdf/060810_1.pdf (2006).
3) 今井亨：リアルタイム字幕放送のための音声認識，信学技報，SP2009-52, WIT2009-58 (2009).
4) 総務省：報道資料：平成 22 年度の字幕放送等の実績，http://www.soumu.go.jp/menu_news/s-news/01ryutsu05_01000012.html (2011).
5) 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/menu_news/s-news/2007/071030_2.html (2007).
6) 西川俊，高橋秀知，小林正幸，石原保志，柴田邦博：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，Vol.J78-D-II，No.11, pp.1589-1597 (1995).
7) 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：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，Vol. J84-D-II，No.6, pp. 877-887 (2001).
8) 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：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, Vol.J87-D-II, No.2, pp.427-435 (2004).
9) 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:ニュース音声認識システムの検討, 音講論集（秋）, 3-1-9, pp.103-104 (1997).
10) 今井亨, 田中英輝, 安藤彰男, 磯野春雄: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, Vol.J84-D-II, No.9, pp.1942-1949 (2001).
11) 小林彰夫，今井亨，安藤彰男，中林克己：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，Vol.40，No.4，pp.1421-1429 (1999).
12) 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，http://www3.nhk.or.jp/pr/marukaji/m-giju093.html (2003).
13) 服部多栄子，椎名努，堂免大規：生字幕放送サービスーシステムとサービスの概要ー，映情学技報，BCT2004-24，Vol.28，No.5，pp.17-20 (2004).
14) 本間真一，小林彰夫，奥貴裕，佐藤庄衛，今井亨，都木徹：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, 映情学誌, Vol.63, No.3, pp.331-338, (2009).
15) 大出訓史，三島剛，江本正喜，今井篤，都木徹：効率的なリアルタイムニュース字幕修正システム，映情学年大，6-3 (2003).

Vol.2011-SLP-88 No.4
2011/10/28

16) T. Imai, S. Sato, S. Homma, K. Onoe, and A. Kobayashi: Online speech detection and dual-gender speech recognition for captioning broadcast news, IEICE Trans. Inf. & Syst., Vol.E90-D, No.8, pp.1286-1291 (2007).
17) D. Povey and P.C. Woodland: Minimum phone error and I-smoothing for improved discriminative training, Proc. IEEE ICASSP, pp.I-105-108 (2002).
18) 今井亨，小林彰夫，佐藤庄衛，本間真一，奥貴裕，都木徹：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, pp.113-120 (2008).
19) 今井亨，本間真一，小林彰夫，奥貴裕，佐藤庄衛，都木徹：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，音講論集（秋），1-1-16，pp.37-40 (2008).