研究発表 4

生字幕制作のための音声認識

本間真一

Speech Recognition for Live Closed-Captioning

Shinichi HOMMA

ABSTRACT

This paper describes three real-time closed-captioning systems using speech recognition that are in practical use and a new system under study. The first system employs the "direct method," whose input is the original program sound, and the second one in use employs the "re-speak method," whose input is the speech rephrased by a "re-speaker". The new system employs a hybrid method combining the "direct" and "re-speak" methods. This paper also describes approaches to improving the accuracy of spontaneous speech recognition in order to enlarge the number of closed-captioned TV programs.

1．まえがき

テレビ番組の音声情報を文字で伝達する「字幕放送」は，聴覚障害者や高齢者のための重要な情報伝達手段となっている。また，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って，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ている。

生放送番組へのリアルタイム字幕付与（生字幕制作）は長い間技術的に困難であったが，音声認識技術や高速入力用キーボードの進展によって，ニュースやスポーツなどの番組においても拡充されてきている。音声認識は更に効率的な生字幕制作を実現するために欠かせない技術であり，今後，いっそうの発展が期待されている。

本稿では，これまでに実用化した，音声認識を用いた生字幕制作システムを紹介するとともに[1)2)]，現在開発中の新しいシステムの概要について述べる[3)]。また，本システムで適用可能な番組を拡充するために行っている，対談などの自由発話の認識性能の改善に向けた研究の取り組みについても紹介する[4)~6)]。

2．これまでに実用化した生字幕制作システム

NHKの字幕放送は1983年の朝の連続テレビ小説「おしん」による実験放送を経て，1985年から本放送を開

始した。その後，主に聴覚障害者への情報伝達手段として，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（完プロ）を対象に年々拡充してきた。一方，ニュースをはじめとする生放送番組に対しては，リアルタイムで漢字かな交じり文の日本語字幕を付与することは技術的に困難であったので，字幕放送を実施していなかった。

2.1 旧ニュース字幕制作システム[1)]

当所が音声認識の研究に着手したのは1969年であり，生字幕制作システムの研究に本格的に取り組み始めたのは1996年である。その後，アナウンサーの音声と原稿を大量に収集してニュース用の音声データベースを構築し，音声認識手法の改良を図り，1999年に認識率を97％に向上させた。2000年３月には，スタジオアナウンサーの原稿読み上げ部分に限定した音声認識を利用して，「ニュース７」において日本初の生字幕放送を実現した。

そのときに用いた旧ニュース字幕制作システムの構成を１図に示す[1)]。実際のシステムは障害時のバックアップ等を考慮しているので複雑な構成であるが，１図は音声認識装置とその周辺装置に限定した簡略化した図である。

旧システムではスタジオアナウンサーのマイク出力を直接音声認識装置に接続するダイレクト方式を採用していた。また，男声用と女声用の音声認識装置をそれぞれ１台設置し，音声認識の誤りを修正するための修正装置もそれぞれ設けた。運用にあたっては，タッチパネルで誤りを発見するオペレーターとキーボードで修正をするオペレーターのペアが２組（計４名）必要であった。

なお，旧システムでは字幕が付与できる区間はアナウンサーの原稿読み上げ部分に限定されていたので，2006年に運用を中断し，高速入力用のキーボード（スピードワープロ）[7)]による方式に移行し，現在に至っている＊1。

2.2 リスピーク（復唱）を用いた生字幕システム[2)]

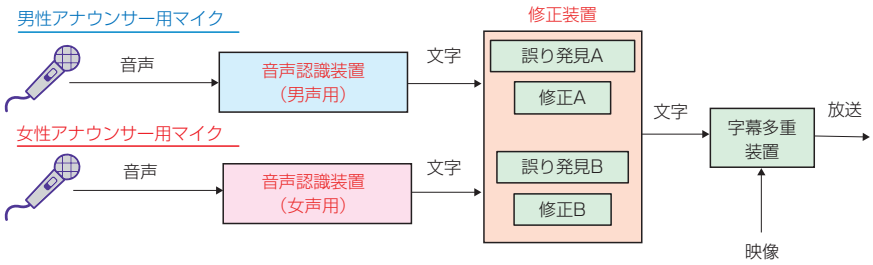
NHKではスポーツ中継，歌番組などに対しても，生字幕放送の拡充を進めてきた。このような番組では，歓声等の背景雑音，話しことば特有の不明りょうな発話などが原因で，番組音声を直接認識させることは困難である。そこで，別の専任の話者（リスピーカー）が番組音声を聞き取り，復唱した音声を認識させるリスピーク方式を採用することで，生字幕放送を実現した。2001年末の「紅白歌合戦」を皮切りに，「オリンピック」「大相撲」「ワールドカップサッカー」「プロ野球」などで生字幕放送を拡充した。

リスピーク方式を用いた生字幕制作システムの構成を２図に示す[2)]。また，リスピーク方式の特徴を以下に列挙する。

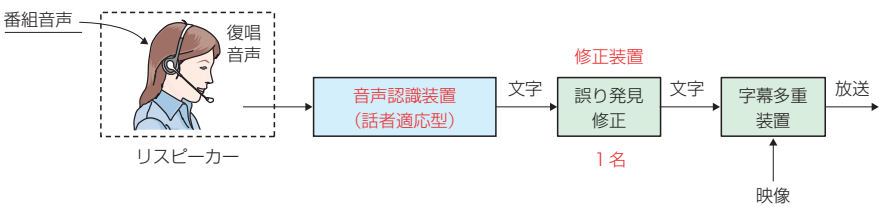
［長所］

・背景雑音が大きい番組や出演者が複数の番組に対応できる。

＊1　（株）スピードワープロ研究所が実施。



1図　旧ニュース字幕制作システム（音声認識周辺）



2図　リスピーク方式による生字幕制作システム

・話者適応型の音声認識装置が使えるので，高い認識率が期待できる。
・実況アナウンスにない拍手や歓声等の補足や，見てわかる内容の要約・省略が可能である。
・認識誤りは，更なる言い換え・言い直しで回避できることがある。

［短所］

・リスピーカーの介在による遅延が発生する。
・リスピーカーが番組音声を完全に復唱できる保証がない。
・リスピーカーの人的コストがかかる。

リスピーク方式は，当初，歌番組やスポーツ中継での運用を想定していたが，音声認識装置に登録可能な語彙(ごい)数を４万語から10万語に拡張したことなどにより，適用可能な番組数は年々増えている。2010年４月からは「スタジオパークでこんにちは」での運用を始めるなど，情報・バラエティー番組への適用も可能となっている。

３．試作したハイブリッドシステム[3)]

3.1　特徴

旧ニュース字幕制作システムで字幕付与が可能であった区間がアナウンサーの原稿読み上げ部分に限定されていた問題を解決するために，ダイレクト方式とリスピーク方式を切り替えるハイブリッド方式のニュース字幕システムを試作した（３図）。3.2節で述べる研究成果を取り入れ，アナウンサーの原稿読み上げ区間だけでなく，記者の現場リポート区間においてもダイレクト方式が適用できるようになった。なお，街頭で一般の人が発話するインタビュー区間など，ダイレクト方式では良好な認識精度が得られない区間においては，リスピーク方式に切り替える運用を想定している。

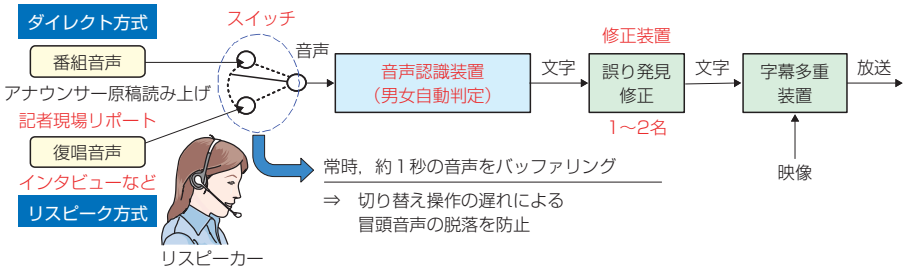
システムの主な特徴を以下に列挙する。

・字幕付与可能な範囲が広がり，音声認識装置をニュース番組の全編に適用できる。
・入力音声の男女自動判別機能を導入し，旧システムでは男女別に必要であった音声認識装置を１台にした。
・修正装置の機能を改善し，従来は４名必要としていたオペレーターを１～２名に削減した。
・リスピーク方式からダイレクト方式に切り替える直後に，常時バッファリングしている約１秒の音声を付加し，切り替え操作の遅れに伴う冒頭の音声の脱落を防止した。

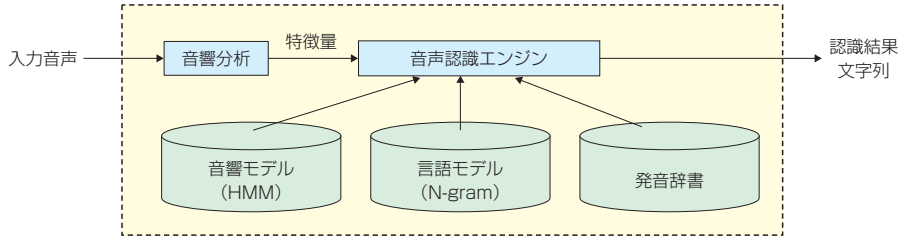
3.2　音声認識装置

４図に示すように，ハイブリッド方式の音声認識装置は音響分析部，音響モデル，言語モデル，発音辞書および音声認識エンジンで構成される。音響分析部では入力音声の波形から発話区間を切り出し，周波数などの特徴量を抽出する。音響モデルは特徴量を事前に学習して得られる統計モデル（HMM）*２であり，言語モデルは電子化されたニュース原稿から各単語の接続

＊２　Hidden Markov Model（隠れマルコフモデル）。音素（母音・子音）ごとの周波数の変化の様子を統計的に表したもの。



3図　新ニュース字幕制作システム（ハイブリッド型）



4図　音声認識装置の構成

確率を事前に学習して得られる統計モデル（単語N-gramモデル）*3である。発音辞書は単語に発音記号を付与したリストである。音声認識エンジンは音響モデル，言語モデルおよび発音辞書を利用して，認識結果となる文字列を探索する。

以下，試作した音声認識装置（新装置）の改善点を旧ニュース字幕システム（旧装置）と比較して列挙する。

・音響分析部では，音素認識*4を用いた発話検出を導入し，発話区間の欠落を減少させた[8]。また，雑音の影響を低減させるフィルター[9]を導入した。
・音響モデルの学習データを増やし，より詳細なモデルを構築した。
・言語モデルの学習データを増やし，新装置では約17年分のニュース原稿で学習した。なお，旧装置では約９年分のニュース原稿で学習した。
・男女並列連続音声認識[8]を導入し，入力音声の男女を自動判定することで，男女別の音響モデルの自動適用を可能にした。なお，旧装置では男女別の音声認識装置を１台ずつ設置していた。
・計算機の性能が向上したので，旧装置と比較して処理速度が高速化され，探索単語数が増加した。

3.3 修正装置

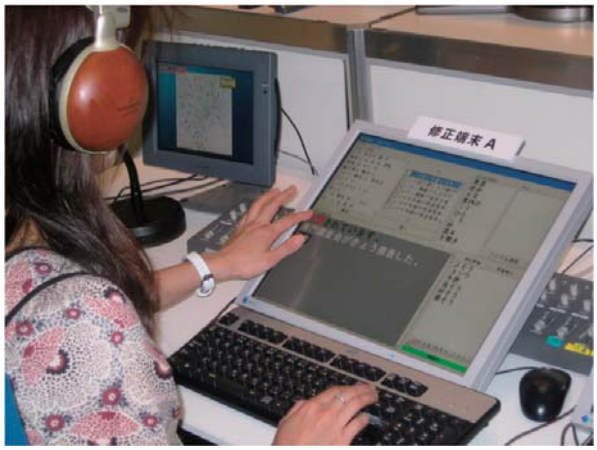
旧システムの修正装置には，誤り発見用と誤り修正用の端末があり，これらをセットにした運用が基本であった。新システムにおける修正装置の台数は，想定される音声認識精度に応じて可変にでき，音声認識精度が高い場合には１台にできるという特徴がある。新装置には，タッチパネルモニターを接続しており，画面上に音声認識装置が出力する単語列をリアルタイムで逐次表示する。オペレーターはヘッドホン経由で認識出力に対応する音声を聞き，誤認識の単語を発見すると，それを指でタッチして選択し，キーボードで速やかに修正する。また，正しい文字列であることを確認したら，速やかにテイクボタン（キーボード上のEnterキー）を押して字幕となるテキストを送出する。修正作業の様子を５図に，修正装置の画面を６図に示す。

3.4 実験

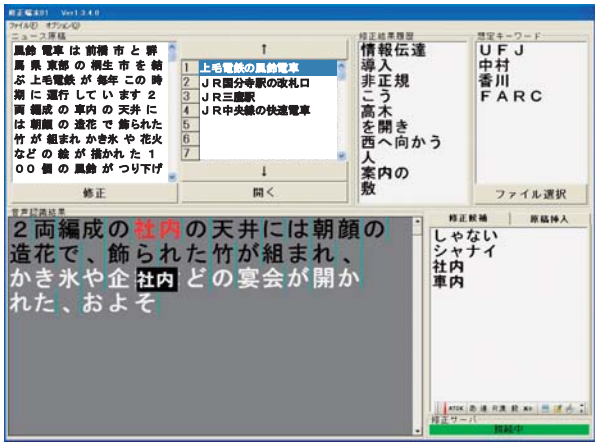
2008年４月30日に放送された「ニュース７」（30分番組）を素材として，生字幕制作実験を行った。ニュース７は１名のアナウンサーのバストショット画面でリード（各ニュース項目の導入部）が読み上げられ，VTRや中継映像に切り替わって本記（詳細な内容）が読み上げられる典型的なパターンのニュース番組（ストレートニュース）である。実験は当所の実験室で行い，音声は放送されたものを用いた。実際の運用では，SE（Sound Effect：効果音）をミックスする前のアナウンサーの音声を直接認識させることができるので，実運用で想定される環境とはやや異なる。

新装置と旧装置の認識性能を認識率で比較した結果と，新装置を用いて修正した後の正解率を１表に示す。

番組全体での認識率は旧装置で91.8％であったが，新装置では96.2％に改善された。ただし，リスピーカーによる発話区間（リスピーク区間）だけで見ると84.4％であり，まだ，改善の余地がある。リスピーカーの発話はインタビュー等でみられる話しことばを復唱するケースが多く，ニュース原稿で学習した言語モデルとはミ



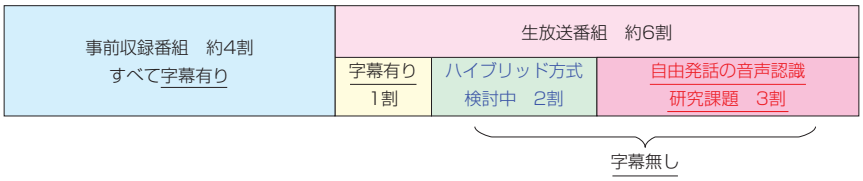
5図　修正装置における修正作業の様子



6図　修正装置の画面

＊３　例えば，「地球」という単語の次には「温暖化」という単語が現れやすいなど，単語どうしのつながりやすさを表すモデル。

＊４　入力音声に含まれる母音や子音を認識し，その記号を順次出力する。例えば，「こんにちは」と発声された音声からは「ｋｏＮｎｉｃｈｉｗａ」が出力される。



7図　NHKの総合テレビにおける字幕放送時間の割合[10)]

1表　新・旧装置による認識率の比較と修正後の正解率

話者（発話の割合）	旧装置	新装置	修正後
アナウンサー（62.6%）	92.4%（雑音無し：95.0%）（雑音有り：88.2%）	96.9%（雑音無し：98.0%）（雑音有り：95.1%）	100.0%
リポーター（29.1%）	94.2%	97.9%	100.0%
リスピーカー（8.3%）	79.8%	84.4%	99.9%
全体	91.8%	96.2%	99.9%

スマッチの大きいことが原因と考えられる。しかし，その発話区間は短く，リアルタイムでの修正が容易なので修正後の正解率は99.9%となった。実験では，リスピーク区間で「ついに」という単語が１つ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ではなかった。

実験に用いた素材では，アナウンサーによる原稿読み上げ区間の割合は総単語数の62.6%であった。旧システムで字幕付与が可能な区間はこの区間に限定されていたが，新システムではリポーター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。従って，１番組中で字幕が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。

4．自由発話音声認識の研究

4.1　字幕放送の現状と研究課題

7図にNHKの総合テレビにおける字幕放送時間の割合（字幕化率）を示す[10)]。現在，約５割の番組に字幕が付与されているが，生放送番組の字幕化率は約１割であり，番組の性質に応じて４種類の方式（ダイレクト方式，リスピーク方式，スピードワープロ方式，一般のキーボード方式）を使い分けて生字幕制作を行っている。

生字幕を付与していない番組のうち，ストレートニュースに関してはハイブリッド方式で字幕化ができる可能性はあるが，その割合は約２割である。従って，残りの３割は現時点では字幕化が困難な状況である。このような番組は原稿や台本が無く，出演者が自由に

… あのーそれぞれの国が分担してっていうのもあるけれども新しい発想っていうのが僕は問われてきてると思うんですねっていうのは …

8図　自由発話区間の書き起こし例

発話をする区間（自由発話区間）を多く含む番組である。リスピーク方式はこのような番組の生字幕制作に有効な手法であるが，復唱精度やコストに関する問題があり，将来的には，番組音声を直接認識する精度を向上させて，ダイレクト方式に移行するのが望ましい。

一般に，自由発話区間を多く含む番組はアナウンサーや記者だけではなく，ゲストも出演し，声質や発話の特徴のばらつきが大きい。また，ニュースとは異なり，読み原稿が無いので，**8図**に示すような話しことば特有の口語表現が多く現れるほか，相づち等による発話の重なりもある。更に，話題に関しては，深い内容となる傾向があり，専門用語も頻出する。これらのことが原因で，試作した装置では75%程度の認識率しか得られず，現在，自由発話の認識精度の向上を目指した研究を進めている。

4.2　音響モデルの改善

従来の音響モデルの学習は音声とその正しい書き起こし文を用いて，音響スコアを最大化するように行われていただけで，正解と不正解（認識誤り）の識別能力については考慮されていなかった。自由発話音声は発声があいまいとなる傾向が強く，この識別能力をいっそう高める必要がある。そこで，認識誤りの傾向を学習することで識別能力を高くする手法の１つである「音素誤り最小化学習[11)]」を導入し，音響モデルの高精度化を行った。また，自由発話音声の不明りょうさ（早口等でみられる発声変形）に着目した更なる改善も行った[5)]。

4.3　言語モデルの改善

ニュース用の言語モデルの学習では大量に存在する電子化されたニュース原稿が利用でき，低コストで学習できるが，自由発話である話しことばの学習では電子化された大量のテキストが存在せず，一般に，人手によって書き起こし文を作成する必要があり，コスト

がかかる。話しことばを多く含むテキストとして，インターネット上で公開されている議会録や放送済みの字幕なども利用可能ではあるが，このようなテキストでは話しことば特有の口語表現が書きことば調に整形されることが多い。そのため，これらのテキストをそのまま学習に利用すると，入力音声と言語モデルがミスマッチとなり，音声認識の性能の向上には結びつかないことがある。

そこで，大量の電子化原稿に少量の自由発話の書き起こし文を最適に混合したテキストを作り，自由発話の特徴を言語モデルに反映させる方法を検討している[5]。また，話しことば特有のフレーズを抽出し，同じ意味を持つフレーズ（同等表現）の出現確率を補正する検討も行っている（9図）[6]。なお，同等表現とは，例えば，書きことばで「〜という」が話しことばでは「〜っていう」に変形するような関係のことである。

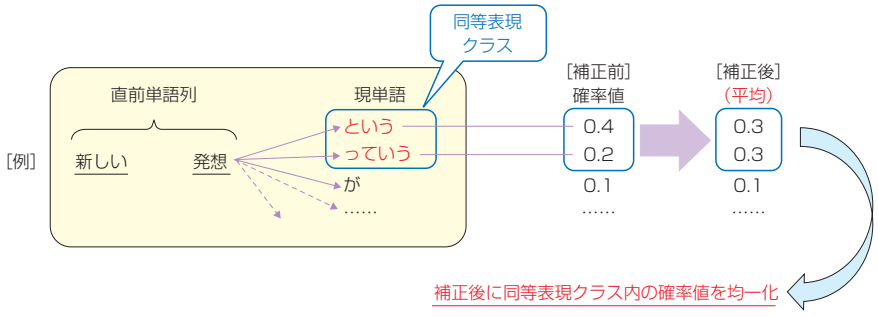
4.4 研究の進展状況

2008年度から開始した自由発話音声認識の研究の進展状況を10図に示す。評価には2008年の5月19日~22日放送の報道情報番組「クローズアップ現代」のゲスト対談部分を使用した。2008年4月に研究を着手した時点での認識率は75.5%であったが，音響モデルと言語モデルの改善によって，2010年3月末時点では86.4%となった。今後，更に，発声変形や話しことばの対策による改善を進めていく。

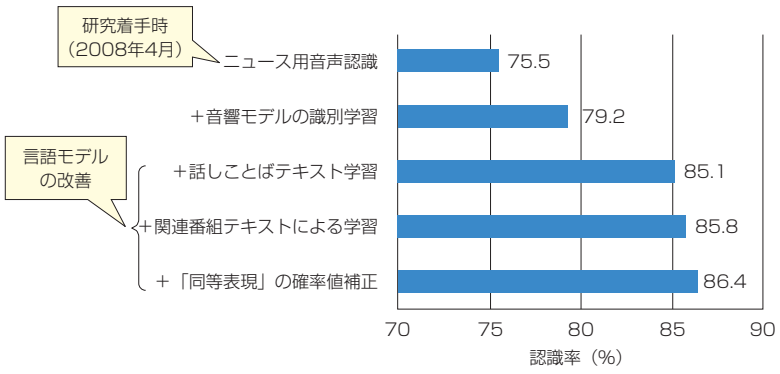
5. むすび

生字幕放送のための音声認識装置について，これまでに実用化した方式と現在検討中のシステムを紹介した。また，更なる生字幕放送の拡充のために，現在取り組んでいる自由発話音声認識の研究についても紹介した。

音声認識装置による初の生字幕放送開始以来，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。今後，放送番組特有の字幕制作条件や音響条件を最大限考慮し，更に難易度の高い番組へ研究対象を移行するとともに，省力化を図った運用性の高い生字幕制作システムを開発していきたいと考えている。また，アーカイブ化された番組検索のためのメタデータ制作[12]など，字幕制作以外の放送分野でのさまざまな音声認識技術の展開を検討していきたい。



9図　同等表現の確率値を補正するイメージ



10図　自由発話音声認識の研究の進展状況

参考文献

1） 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，D-2, Vol.J84-D-2, No.6, pp.877-887（2001）

2） 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，D-2, Vol.J87-D-2, No.2, pp.427-435（2004）

3） 本間，小林，奥，佐藤，今井，都木：“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）

4） 佐藤，奥，本間，小林，今井，都木：“単語に依存した発声変形音素の効率的な識別学習，”音響学会春季講演論文集，2-Q-8, pp.253-256（2010）

5） 本間，佐藤，奥，小林，今井，都木：“報道系対談番組向け自由発話音声認識の改善，”音響学会春季講演論文集，3-Q-17, pp.243-244（2009）

6） 本間，奥，小林，佐藤，今井：“同意の単語連鎖を考慮した自由発話音声認識，”音響学会春季講演論文集，1-Q-9, pp.169-170（2010）

7） 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，D-2, Vol.J78-D-2, No.11, pp.1589-1597（1995）

8） 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., Vol.E90-D, No.8, pp.1286-1291（2007）

9） H. Hermansky and N. Morgan：“RASTA Processing of Speech，”IEEE Trans. on Speech and Audio Processing, Vol.2, No.4, pp. 578-589（1994）

10）総務省：“報道資料：平成20年度の字幕放送等の実績，”http://www.soumu.go.jp/menu_news/s-news/ 02ryutsu05_000007.html

11）D. Povey and P. C. Woodland：“Minimum phone error and I-smoothing for improved discriminative training，”Proc. IEEE ICASSP, pp.I-105-108（2002）

12）小林，奥，本間，佐藤，今井，都木：“コンテンツ活用のための報道番組自動書き起こしシステム，”情処学音声言語情報処理研資，Vol.2009-SLP-77, No.20（2009）

ほんましんいち
本間真一

1992年入局。新潟放送局，技術局を経て，1998年より音声認識の研究開発に従事．現在，放送技術研究所人間・情報科学研究部専任研究員。