音声言語情報処理　34－28
（2000．12．22）

〔招待論文〕 ニュース音声自動字幕化システム

安藤彰男

NHK放送技術研究所
東京都世田谷区砧1－10－11
ando@strl.nhk.or.jp

あらまし

平成12年3月27日から、ＮＨＫニュース番組「ニュース７」で字幕放送が開始された。この字幕放送は、音声認識技術を利用して、リアルタイム字幕を試みた世界でも初めての例である。

テレビニュース番組に対する字幕放送を実現するためには、リアルタイムで字幕原稿を制作する必要がある。日本語の場合には、仮名漢字変換などに時間がかかるため、アナウンサーの声に追従して字幕原稿をキーボード入力することは困難であり、いままで、我が国ではニュースの字幕放送は実現されていなかった。そこで、音声認識技術を利用することとした。

本稿では、「ニュース７」字幕放送を実現するために開発したニュース音声認識システム、及び音声認識結果を人手で即座に修正するシステムについて解説する。

キーワード：　ニュース放送、リアルタイム字幕、音声認識、認識誤り修正

A Simultaneous Subtitling System for Broadcast News Programs

Akio ANDO

NHK Science and Technical Research Laboratories
1-10-11 Kinuta Setagaya Tokyo 157-8510 JAPAN
ando@strl.nhk.or.jp

Abstract

NHK launched a caption broadcasting on its news program "News7" on March 27, 2000. It is the first trial of "live" captioning services by speech recognition.

To realize captioned broadcasting for news program, it is necessary to create captioned manuscripts in real time. Entry of captioned manuscripts cannot keep pace with the speed of speech in Japanese language, because a certain length of time is required for selection among homonyms. For this reason, the caption broadcasting had not been realized in Japan, and therefore, we tried to use speech recognition technology.

This paper described a newly developed broadcast news transcription system and recognition error correction system used in the caption broadcasting of "News 7".

Key words:　Broadcast news,　Live captioning system,　Speech recognition,　Recognition error correction

1. はじめに

お年寄りや耳の不自由な方への放送サービスとして、話している内容を字幕で表示する字幕放送の拡充が求められている。従来の字幕放送は、放送日の数日前までに制作が完了した番組が対象であり、キーボードによる原稿入力など、人手による要約作業によって制作されている。

一方、生放送番組であるニュースに対して字幕放送を実現するためには、リアルタイムで字幕原稿を作成する必要がある。日本語の場合、同音異義語の選択などに時間がかかるため、特殊なキーボードを利用しても、声に追従して字幕原稿を入力することは難しい。日本語のニュース番組に対する字幕放送を実現するためには、音声認識などの自動的手段を用いる必要がある。

ニュース番組を対象とした音声認識に関しては、米国で、DARPA (Defense Advanced Research Projects Agency) の HUB4 を中心として、放送ニュース音声を対象とした大語彙連続認識の研究が進められている(1)。米国では、既にキーボード入力によるニュース番組の字幕放送が実現されている。従って、同プロジェクトは、リアルタイム性を追及するものではなく、また、発声内容の正確な文字化よりも、音声認識結果を用いた情報検索を目指したものであった。

NHK では、ニュース番組に対する字幕原稿をリアルタイムで作成することを目標として、ニュース音声認識の研究を進めている。この一環として、ニュース番組中のアナウンサーがスタジオ内で原稿を読む部分に対象とした実用化システムを開発した。このシステムを利用して、平成１２年３月２７日から、NHK のニュース番組「ニュース７」の字幕放送を試行的に実施している。これは、音声認識技術を利用してニュース字幕放送を実施した、世界でも初めての試みである。

ニュース字幕制作システムは、音声認識システム、認識誤り修正システム、そして言語モデル学習システムから構成される (図１参照)。このうち、オンライン処理を行うのは、音声認識システムと、認識誤り修正システムであり、言語モデル学習システムは、オフラインで処理を行う。

以下、２．～４．で、ニュース字幕制作システムの各サブシステムについて説明する。５．で実用化状況について述べた後、６．で今後取り組むべき課題について述べる。

2. ニュース音声認識システム

ニュース音声認識システム(2)は、音響モデルとしてHMM (Hidden Markov Model)、言語モデルとして単語バイグラム、単語トライグラムを用いた統計的音声認識手法を利用している。図２に、システムのブロック図を示す。

2.1 音響分析

音響分析部では、入力音声を、サンプリング周波数16kHz、量子化精度16ビットでデジタル化し、ハミング窓を用いて、短時間周波数分析を行う。分析フレームの長さは25ms、フレーム周期10msとした。周波数分析結果から、各フレームごとに、12 次元の MFCC (Mel Frequency Cepstrum Coeficient)を計算する。さらに、MFCC の各次元ごとに、５フレーム分の係数列に対する最小二乗直線の傾きを求めることにより、12次元のΔ-MFCC 係数を計算する。Δ-MFCC 係数に対して同様の操作を行うことにより 12 次元のΔΔ-MFCC 係数も計算する。また、各フレームごとの対数パワーと、そのΔ、ΔΔも求め、これら全てをまとめて、各フレームごとに 39 次元の音響パラメータを得る。音響分析条件を、表１に示す。

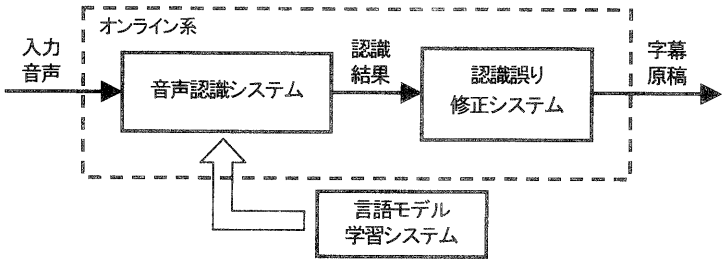

図1 ニュース字幕制作システム

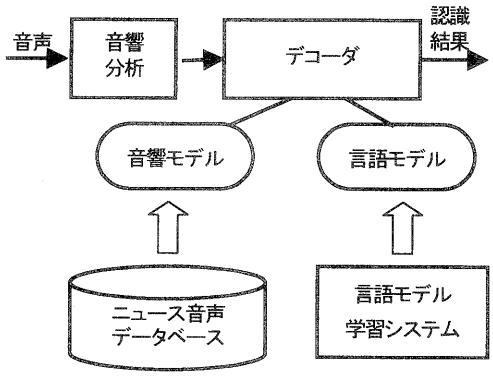

図2 ニュース音声認識システム

表1 音響分析条件

標本化周波数	16kHz
分析窓	ハミング窓
フレーム長	25ms
フレーム周期	10ms

表2 デコーダのパラメータ

状態内保存パス数:	4
第1パス出力文数:	200
ビーム幅:	160
単語終端ビーム幅:	100
言語スコア重み:	14
挿入ペナルティ:	0

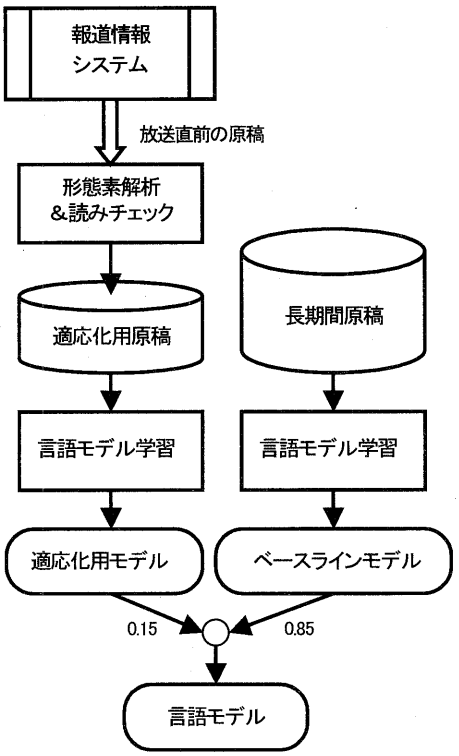

図3 言語モデルの学習

2.2 音響モデル

音響モデルとしては、性別依存の不特定話者トライフォンHMMを用いた。HMMのタイプは、8混合のガウス分布を用いた連続分布HMMである。音素数は42である。HMMは、男性24名、女性21名のNHKのニュース担当アナウンサーが発声した、音素バランス文各100文と、「NHK ニュース音声データベース」(3)を用いて学習した。

2.3 言語モデル

言語モデルとしては、単語バイグラム、単語トライグラムを利用する。本システムでは、形態素で単語を定義する。言語モデルの学習方法については、3．で述べる。

2.4 デコーダ

認識候補を選択するためのデコーダには、2パスデコーダを採用した。第1パスでは、音響モデルにトライフォンHMM、言語モデルに単語バイグラムを用いて、Viterbi ビームサーチによる単語依存N-best 探索(4)を行う。探索の際、各候補のスコアとしては、音響モデルを用いて計算した音響スコア（対数確率）と単語バイグラムを用いて計算した言語スコア（対数確率）の重み付け平均により算出する。第2パスでは、第1パスの出力として得られたN-best文に対して、それぞれの言語スコアを、単語トライグラムを用いて再計算することにより、rescoring を行う。第1パスと第2パスでの、音響スコアと言語スコアとの重みは、同じ値を利用した（音響スコアに対して、言語スコアを14倍して加算）。

デコーダの第1パスでは、単語のネットワークを木構造で表現している。木構造のネットワークは、語頭部分のノードを共有するため、語頭でのアクティブなノードの数を制限でき、処理量が削減されるという長所を持っている。いかし、単語を特定できるノードに処理が進むまでは、バイグラムを用いた言語スコアの計算ができない。そこで、枝刈りの際、前単語とノードを共有する単語とのバイグラムのうち最大の値を利用した。この際、アクティブになる割合の高い語頭のLレイヤーまでの全ノードと、L+1レイヤー目の単語のうち、ユニグラム確率の高い順にK番めの単語に対応するノードについて、事前に最大バイグラムを計算しておくことにした。デコーダ第1パスのパラメータを表2に示す。

なお、実用上は、少ない遅れ時間で認識候補を確定することが望ましいため、ニュース字幕制作システムでは、逐次2パス方式(5)を採用した。

3. 言語モデル学習システム

ニュースでは、新しい単語が次々に出現する。そこで、記者が逐次入稿してくる原稿を利用した言語モデルの適応化手法" TDLM" (6)を採用した。図3に、言語モデル学

習のフローを示す。

「ニュース7」字幕制作では、原則として放送開始の20分前に、NHK報道情報システムをアクセスして、過去6時間の間に作成された最新のニュース原稿を取り込む。この原稿を、形態素単位に分割して、適応化用原稿とする。あわせて、新たに出現した単語について、その読みを付与する。単語の読みは、形態素解析時に自動的に得られた結果を、人手でチェックして、確定する。適応化用原稿から、単語バイグラムと単語トライグラムを計算し、適応化用モデルを作成する。この際、バックオフスムージングにはGood-Turing法を利用した。その際のバイグラム、トライグラムに対するcutoffは、それぞれ1，2とした。

一方、1991年4月から蓄積した長期間のニュース原稿より、事前にベースラインモデルを作成しておく。語彙は、頻度順に20,000語を選択する。このモデルを、適応化用モデルを利用して適応化し、言語モデルを作成する。なお、語彙としては、両モデルの語彙の和集合で定めた。

報道情報システムのアクセス開始から、適応化された言語モデルを音声認識システムが読み込むまでの処理は、約8分で完了する。

4. 認識誤り修正システム(7)

音声認識では、認識率100%の達成は極めて困難である。従って、誤りのない字幕放送を実施するためには、認識結果中の誤りを、人間が修正する必要がある。このような人間による修正を、少ない遅れ時間で実現するため、認識誤り修正システムを開発した。システムのブロック図を図4に示す。システムは、2系統の発見・修正端末と、音声サーバ、修正サーバから構成される。

4.1 音声サーバ

音声認識結果を即座に修正する作業では、修正の基準となるのは、あくまで音声情報である。音声認識システムでは、トライグラムなどの単語の接続情報を利用して認識を行うため、誤りを含んだ認識結果であっても、一見したところでは自然な日本語に見えることが多い。従って、認識誤りを見逃さないためには、視覚にのみ頼って不自然な文字列を発見するのではなく、視覚と聴覚を連動させた作業、すなわち、音声を聴きながら、対応する認識結果を連続的に確認していくという作業が必要である。音声サーバは、なるべく認識結果に同期させた音声呈示を行うため、話速変換の技術(8)を利用して、音声呈示を行う。

また、音声サーバは、検出された無音区間に合わせて、音声を2つの修正端末に分配する。たとえば、入力音声の

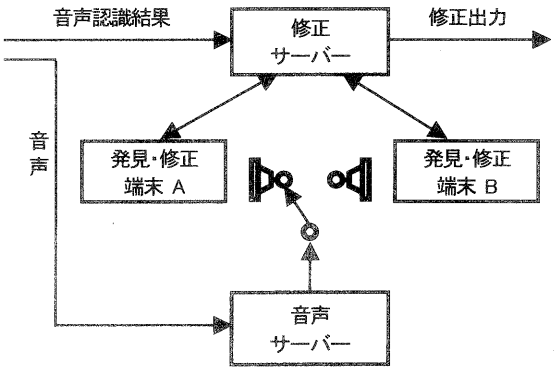

図4 認識誤り修正システム

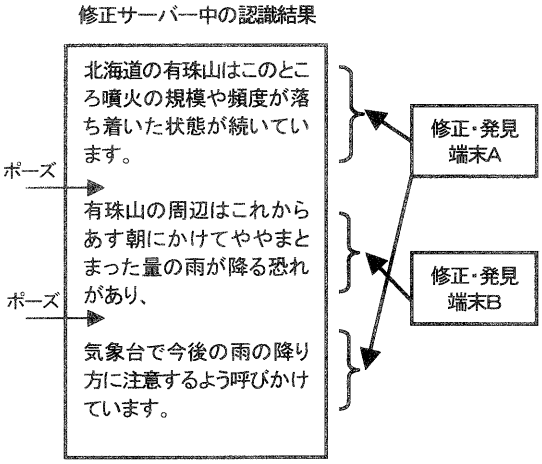

図5 修正サーバーと2系統の端末による誤り修正

最初の部分は、修正端末Aに分配する。入力音声中に、無音区間が検出された場合には、その後の入力音声は修正端末Bに分配する。以後、無音区間が検出されるたびに、入力音声は2系統の端末に順番に分配する。

4.2 修正サーバ

修正サーバは、音声認識結果の確認・修正作業を制御するサーバである。2系統の端末の画面上に、その端末が確認・修正すべき認識結果を表示する。また、修正された認識結果を出力する（図5参照）。

4.3 修正端末

人間は、認識誤りを修正する作業に注意を集中していると、その間に発生した認識誤りを聞き逃したり、あるいは、認識結果が誤っていることは分かるが、正解は何であったかを聞き漏らしたりする。そこで、修正の精度向上と

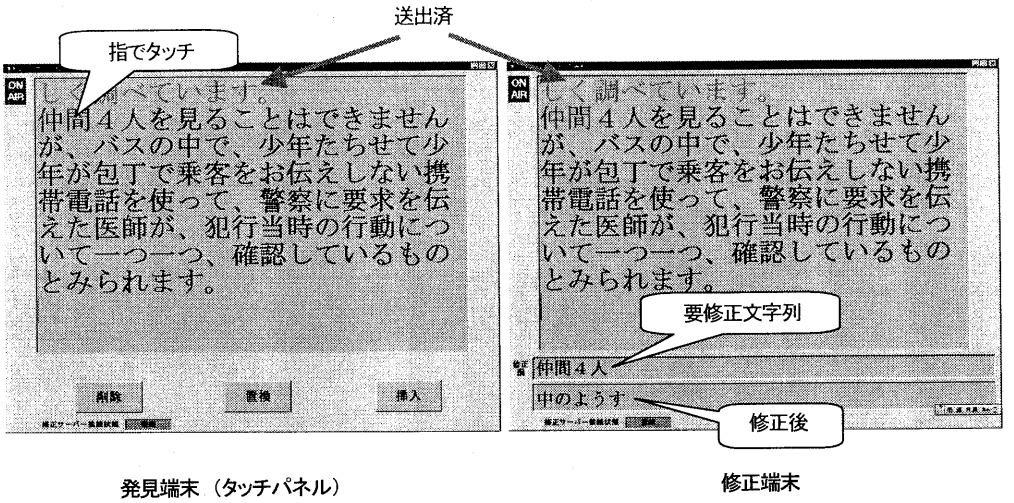

図6 誤り発見・修正端末の画面

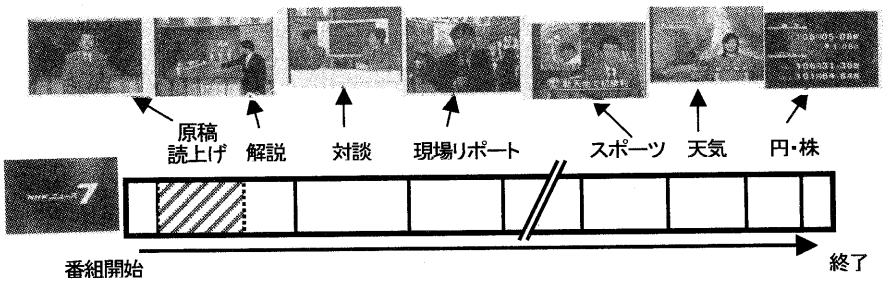

図7 「ニュース7」字幕放送の実用化状況

斜線部：実用化済み

迅速化のため、修正作業を、誤り発見と、発見された誤りの修正という２つの作業に分ける方式を考案し、発見端末と修正端末の２種類の端末を開発した。端末の画面の例を、図６に示す。

5. システムの実用化

実用化システムは、図１に示したニュース字幕制作システムと、事前に用意された原稿をそのまま字幕として送出するシステムとのハイブリッドシステムで構成した。ただし、多くのニュース項目は、原稿が放送直前まで確定しないため、事前に用意した原稿を送出しているのは、字幕放送全体の 15%程度である。これらは、字幕表示の遅れが許されない場合（例えば、次のニュース項目が、地方局発の場合）や、修正オペレータの疲労防止などのために利用されている。残りの約 85%は、本稿で解説する字幕制作システムが利用されている。

平成１２年３月２７日の「ニュース７」字幕放送開始以来、毎日、このシステムを利用して字幕放送を実施しているが、いまのところ、障害などによる放送中断は起こっていない。音声認識を利用した字幕制作システムでの表示遅れ時間（音声発声から、字幕表示までの時間）は、約 10 秒である。この中には、音声認識の遅れ時間（2~3 秒）、確認・修正に要する時間（約５秒）、字幕として 15 文字×2 行で表示するためのバッファリングの時間などが含まれている。

6. 今後の研究課題

「ニュース７」字幕放送の実用化状況を、図７に示す。

現在字幕放送が実用化されているのは、アナウンサーが原稿を読み上げている部分のみであり、しかもスポーツなどの部分は、字幕放送を実施していない。実用化の際の目標としているのは、95%以上の単語正解精度の達成である。

現状での認識性能を示すため、2000年6月1日~7日に放送されたNHKニュース番組「ニュース7」、「おはよう日本」を利用して、テストセットを作成し、認識実験を行った。テストセットの構成を表3に示す。表3のうち、「スポーツ」のsports.cleanと、「雑音ありニュース」は、現場のさまざまな音がミックスされた音声である。NHK放送センターから放送する場合には、アナウンサーの音声のみを取り出すことが可能であるが、地方から編集済みの音が送られてくる場合には、現場音がミックスされた状態で認識する必要があるため、これらをテストセットに加えている。音響モデルは、1996年6月1日~6月30日、1997年6月1日~7月31日、1998年4月1日~9月30日、および1999年4月1日~2000年5月30日に放送された、「ニュース7」、「おはよう日本」などのニュース番組の音声によって学習した(3)。言語モデルは、1991年4月1日から、テストセット中の音声データが放送された時点までに作成された記者原稿をベースラインとして、放送直前6時間分の原稿でTDLMにより適応化した。認識結果を表4に示す。

すでに実用化されているニュース読み上げ部分については、95%の目標性能が達成済みであり、その他の部分に対する認識性能の向上が課題である。自由発話のうち、解説部分については、88%程度の認識精度が得られている(9)。スポーツのうち雑音の少ない部分(sport.clean)も88%程度である。現場リポートは、今回のテストセットでは90%を超える認識性能が得られている。当面は、これらの部分の実用化を目指して、研究を推進する。

表3 NHKテストセット2000の概要

	コンディション	ID	文章数
1	ニュース読み上げ	studio_news	339
2	自由発話(解説、対談など)	spontaneous	160
3	スタジオ・レポート	studio_report	30
4	天気	weather	340
5	スポーツ	sports.clean	139
		sports.noisy	254
6	現場リポート	field_report	109
7	雑音ありニュース	noisy_news	140

話者は全て男性

表4 認識実験結果

	ID	単語正解精度
1	studio_news	98.32%
2	spontaneous	78.05%
3	studio_report	87.86%
4	weather	77.21%
5	sports_clean	88.09%
	sports_noisy	72.15%
6	field_report	91.77%
7	noisy_news	96.56%

7. まとめ

「NHKニュース7」字幕放送で利用している、字幕制作システムについて解説した。今後は、ニュース番組の完全字幕化と、ニュース字幕放送の拡充を目指して、さらなる研究開発を進めていく。

参考文献

(1) Proceedings of DARPA Speech Recognition Workshop, February 1996, Morgan Kaufmann

(2) 今井(亨)、小林、尾上、安藤："ニュース番組自動字幕化のための音声認識システム"，情報処理学会研究報告 SLP-23-11 pp.59-64 (1998-10)

(3) 安藤、宮坂："ニュース音声データベースの構築"，日本音響学会講演論文集，2-Q-9，1997-3

(4) R. Schwarz, et al, "A Comparison of Several Approximate Algorithms for Finding Multiple (N-best) Sentence Hypotheses, Proc. of IEEE ICASSP-91, pp.701-704, 1991-5

(5) 今井(亨)、小林、佐藤、安藤："逐次2パスデコーダを用いたニュース音声認識システム"，信学技報 SP99-129 (1999-12)

(6) 小林、今井(亨)、安藤、中林："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 pp.1421-1429 (1999)

(7) 後藤、今井(亨)、清山、今井(篤)、都木、安藤、磯野："ニュース音声認識結果のリアルタイム修正装置"，2000年信学総合大会，A-15-15 (2000-3)

(8) 清山、今井(篤)、都木："ポータブル話速変換器の開発"，NHK技研R&D Vol.52 pp.61-68 (1998-8)

(9) 本間、小林、佐藤、今井(亨)、安藤："ニュース解説を対象にした音声認識の検討"，信学技報NLC，SP&情処SLP資料(本研究会資料) (2000-12)