ヒューマンインタフェース 80−11
音声言語情報処理 23−11
（1998．10．17）

ニュース番組自動字幕化のための音声認識システム

今井亨　小林彰夫　尾上和穂　安藤彰男

NHK放送技術研究所

〒157-8510 東京都世田谷区砧1-10-11

E-mail: {imai, akio, onoe, ando}@strl.nhk.or.jp

あらまし 高齢者や聴覚障害者への放送サービスを充実させるため、音声認識を利用した放送番組の自動字幕化を検討している。本報告では、現在開発中のニュース音声認識システムの概要を述べる。アナウンサーの音声を認識するデコーダーは、bigramを用いた単語依存*N*-bestに基づく第1パスと、trigramによるリスコアリングの第2パスで構成される。語彙サイズを5Kから65Kまで変化させ、音素ネットワークがリニアと木構造の場合について、認識率と処理時間を調べた。さらに、ニュースの特徴を生かした時期依存言語モデルと、電子原稿を利用した認識結果の修正について述べる。

キーワード 連続音声認識、ニュース、字幕

A Broadcast News Transcription System for Captioning

Toru Imai, Akio Kobayashi, Kazuo Onoe, and Akio Ando

NHK (Japan Broadcasting Corp.) Science & Technical Research Laboratories

1-10-11 Kinuta Setagaya, Tokyo 157-8510 Japan

E-mail: {imai, akio, onoe, ando}@strl.nhk.or.jp

Abstract Automatic captioning for TV shows is required by old ages and hearing impaired. This paper describes a broadcast news transcription system for captioning, which is under development. A decoder converting announcers' speech into texts consists of two passes: the first pass based on word-dependent *N*-best search with bigram and the second pass for rescoring with trigram. Recognition accuracy and processing time were examined with a linear structured or tree structured phoneme network for some vocabulary sizes from 5K to 65K. This paper also describes a time dependent language model updated with latest news and post-correction of the transcriptions by electronic draft scripts.

Key words continuous speech recognition, broadcast news, captioning

1．はじめに

テレビの文字放送による字幕サービスは、日本ではいくつかのテレビ番組に対して10年以上前から行われている。しかし、字幕文字はキーボード等を用いて手入力で作成されているため、高齢者や聴覚障害者が望む字幕付き番組を、大量に提供できていないのが現状である。一方、米国では、裁判所の速記用に開発された特殊キーボードを用いることにより、生放送であっても時間遅れなく字幕文字を入力することができるので、字幕付き番組が充実している。日本語の場合、かな漢字変換、同音異義語の選択という過程を経るため、訓練された人でも、人間の話す早さで日本語を入力するのは困難である。そこで我々は、連続音声認識の技術を利用して、放送の字幕文字を自動生成する研究を行っている。

字幕サービスはいろいろなジャンルの番組で求められているが、中でもニュース番組への要求は高い。ニュースには、記者らが事前に用意した電子原稿が存在するが、それをそのまま字幕に用いるわけにはいかない。アナウンサーが読み上げている原稿は、記者が用意した電子原稿のプリントアウトに、ディレクターらが手書きで修正を加えたものである。その修正は、放送直前や放送中であったり、突然別のニュースに差し替えられることもある。このように、電子原稿は必ずしも放送で読まれる内容と一致しないため、実際のニュース音声をもとに、字幕が作成されることが望ましい。そこで、ニュース番組の字幕を、連続音声認識によって自動的に生成（最終的には人が確認・修正）するシステムの検討を進めている（図1）。

本稿では、現在開発中のニュース音声認識システムの概要を報告する。まず、デコーダーにおける語彙サイズと音素ネットワークの検討結果について述べる。次に、ニュースの特徴を生かした時期依存言語モデルと、電子原稿を利用した音声認識結果の修正について述べる。

2．デコーダー [1]

アナウンサーの音声を認識するデコーダーは、図2に示すように、2パスで構成される。第1パスでは、言語モデルにbigram、音響モデルにtriphone-HMMを用いて、ViterbiビームサーチによるN-best探索を行う（単語依存N-best[2]）。すなわち、HMMの各状態で、直前単語が異なる最大n個（$n \ll N$）のパスを残しながら探索を進める。直前単語の同じパスが同一状態に達した時は、最大スコアのパスを残す。単語終端では、現単語のラベル、最終状態の最大n個のスコア、それぞれの直前単語終端へのポインタを残すことで、単語ラティスを生成していく。次単語へは、現単語の最良スコアのパスのみを進める。文末で単語ラティスを再帰的にトレースバックして、N-best文を得る。

ビームサーチの枝刈りには、対数尤度差に基づく大域的ビーム幅と、より狭い単語終端ビーム幅を併用する。また、バックオフ時に接続可能な単語を、unigramの上位K単語に制限して、処理量を削減する。

第1パスの探索では、全単語の発音に従って展開されたtriphoneの音素ネットワークにおいて、各ノード（triphone）に対応するHMMを1フレーム分進める処理（HMM-step）と、ノード間（単語内および単語間）の遷移の処理（grammar-step）を行う。音素ネットワークは、1音素目から単語独立なリニアなものと、語頭部分の音素を複数単語で共有する木構造[3]の両者を検討した。

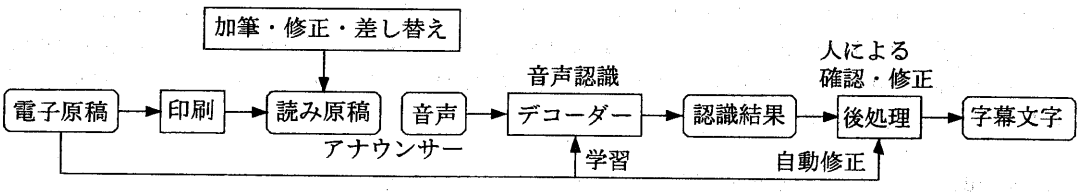

図1　ニュース音声認識システム

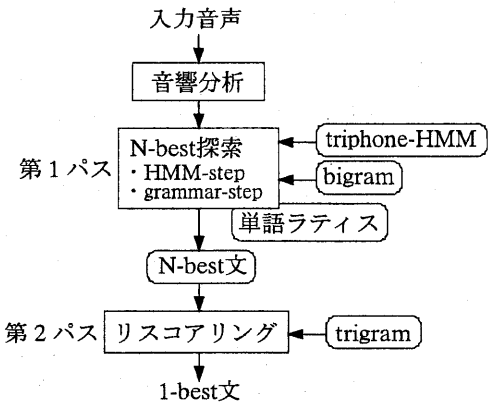

図2　デコーダー

第2パスでは、*N*-best文をtrigramによってリスコアリングし、最良スコアの文を認識結果として出力する。

2.1 音素ネットワーク構造

2.1.1 リニア

リニアな音素ネットワークは、1音素目で現単語を特定できるので、単語間ですぐにbigramを適用できる。しかし、その構造から、単語始端でアクティブなノード数が一気に増加してしまう。これを避けるために、アクティブなノードの個数に応じて閾値を動的に制御することが、処理時間の短縮に有効である[4]。

2.1.2 木構造

木構造の音素ネットワークは、語頭部分のノードを複数単語で共有するため、語頭でのアクティブなノード数を削減できる。しかし、現単語を特定できるノードに処理が進むまで、スコアにbigramを適用できない。そこで枝刈りには、ノードを共有する単語のうちで最大のbigramを使うことが広く行われている[5]。そのためには、次のノードに遷移した時、共有単語リストが変化したなら、直前単語に依存した最大bigramを求めなければならず、実装時には処理量を削減する工夫が必要になる。語彙サイズが小さい場合には、あらかじめ全ノードで直前単語ごとに最大bigramを計算しておき、テーブル化しておくことが可能だが、語彙サイズが大きくなるにつれて、必要とするメモリー量は増加する。そこで、語彙サイズに応じて、アクティブになる割合が高い語頭のLレベル目までの全ノードと、L+1レベル目のunigram上位K単語に関するノードのみ、事前に最大bigramを計算しておくことにする。

木構造のネットワークは、動的に生成していく手法も提案されているが[3,5]、ここでは静的[6,7]なものを検討した。ただし静的な場合、共有単語リストが異なる前ノードで決定された直前単語を引き継ぐので、1-bestでもnは大きい方が望ましい。

2.2 認識実験

NHKニュースの認識実験を、異なる語彙サイズ（5K, 10K, 20K, 40K, 65K）と音素ネットワーク構造で行った。実験条件を表1に示す。言語

表1　実験条件

音響分析	16kHzサンプリング ハミング窓幅25ms、周期10ms 39次元パラメータ (12次MFCC+対数パワー、各々の△、△△)
音響モデル	男女別不特定話者、triphone-HMM 8混合ガウス分布、音素数42 モデル数（ニュース語彙20Kの場合） 論理HMM=5,844 物理HMM=1,366 (男), 1,396 (女) 共有化状態数=2,356 (男), 2,420 (女) 学習データ：日本音響学会・ATR連続音声 男54名12H、女58名15H
言語モデル	bigram（第1パス）, trigram（第2パス） バックオフスムージング、Good-Turing cut-offはそれぞれ1, 2 学習データ：NHKニュース・データベース ('91.4.1～'96.7.10)
デコーダー	状態内保存パス数n=4 第1パス出力文数N=200 大域的ビーム幅160、単語終端ビーム幅50 言語スコア重み14、挿入ペナルティ0 バックオフ・接続可能単語数K=2,000 リニア・最大目標アクティブノード数20,000
評価データ	NHKニュース[8] ('96.7.11～'96.7.14) スタジオアナウンサー・クリーンスピーチ 男女各50文（性別既知） 計100文（3,986単語） 文平均長13.2sec.
マシン	DEC Alpha 600MHz メモリーサイズ750MB

モデルは、評価日の前日までの5年3カ月分の電子原稿から学習した。ネットワークのノード数は、語彙サイズ20Kの場合、リニアの時に160K、木構造の時に85Kである。木構造の場合の最大bigramテーブル化は、総サイズが約300MBになるよう、上記の語彙サイズに対してそれぞれ、語頭のレベル数L=12, 3, 2, 1.2, 1.1とした（非整数は次のレベルの一部まで）。

認識結果を表2に示す。認識率は、音素ネットワーク構造によらず、語彙サイズ20Kまで上昇し、20Kを超えるとほとんど変化しなかった。木構造の音素ネットワークは、語彙サイズによらず、リニアなものよりも認識率が高かった。これは、木構造の場合、語頭に近いほど実際よりも大きなbigramで枝刈りをすることになり、リニアの時に語頭付近で枝刈りされてしまう単語が、生き残る可能性が高いためと思われる。

処理時間は、語彙サイズ20K以下では木構造が優位だが、最大bigramを十分にテーブル化しておくことができなかった40K以上では、リニアな場合よりも遅くなった。

2.2.1 各処理の効果

認識率と処理時間を改善するための各処理の効果について調べた。第2パスのtrigramによるリスコアリングにより、語彙サイズ20Kで木構造の場合、認識精度が3.1%改善された。バックオフ時に接続可能な単語の制限は、認識率の低下を0.5%に抑えて、処理時間を92%に削減した（リニアの場合は0.2%の低下で処理時間を85%に削減）。最大bigramのテーブル化は、認識率に影響を与えずに、処理時間を4%に削減した。より狭い単語終端ビームの併用は、語彙サイズ20Kで木構造の場合、認識率が不変で処理時間を95%に、リニアの場合、認識率の低下を0.4%に抑えて、処理時間を92%に削減した。

2.2.2 負荷分布

デコード時の負荷分布を、図3に示す。アクティブなHMMを1フレーム進める処理（HMM-step）に全体の56%、単語内および単語間での処理（grammar-step）に全体の39%の処理時間を要した。

├第1パス	98%
│ ├音響分析	2%
│ ├HMM-step	56%
│ │ └ガウス分布出力確率密度計算	38%
│ └grammar-step	39%
│ └最大bigram計算（テーブル以外）	7%
└第2パス	2%

図3　負荷分布（木構造、語彙サイズ20K）

2.2.3 アクティブ・ノード

デコード時のアクティブなノード数の変化を、ある評価文について調べた（図4）。音素ネットワークが木構造の場合には、リニアな場合よりもアクティブなノード数が全般に少なく、平均で約1/3に削減されていた。

次に、木構造（語彙サイズ20K）の場合、どのレベルのノードがどれくらいの割合でアクティブになるのかを、女性の50評価文で調べた。表3に示すように、平均的には、レベル1のノードは96.4%がアクティブになっており、レベル2のノードは19.6%がアクティブになっていた。アクティブなノードについて、レベルごとの分布を調べると、レベル1が22.0%、レベル2が45.5%であった。すなわち、全レベルのアクティブなノードのうち、67.5%がレベル2以下（レベル3以下では85.5%）であるので、低レベルでの最大bigramを事前にテーブル化しておく効果は

表2　認識結果

語彙サイズ	ﾃｽﾄｾｯﾄﾊﾟｰﾌﾟﾚｷｼﾃｨ		OOV	リニア・ネットワーク		木構造ネットワーク	
	bigram	trigram		認識精度	処理時間(xRT)	認識精度	処理時間(xRT)
5K	67.4	35.2	5.2%	73.9%	2.1	77.4%	1.9
10K	72.3	36.3	2.0%	78.3%	2.6	82.2%	2.3
20K	76.1	37.5	0.5%	79.7%	3.1	83.5%	2.8
40K	77.5	38.3	0.3%	79.9%	3.7	83.3%	5.9
65K	77.8	38.4	0.2%	79.9%	4.4	83.5%	9.5

RT=実時間

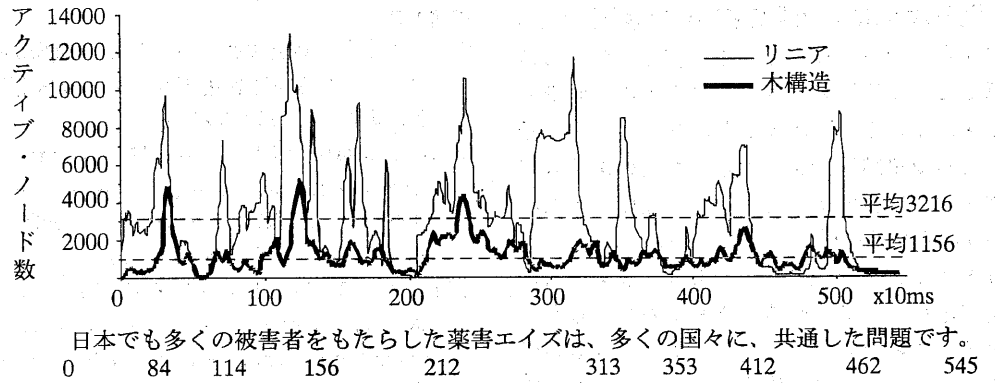

図4　アクティブ・ノード数の変化（語彙サイズ20K）

表3　各レベルでのアクティブ・ノード分布（木構造・語彙サイズ20K）

レベル	1	2	3	4	5	≧6	≧1
ノード数	179	1,823	6,207	12,592	15,842	48,324	84,967
累積	0.2%	2.4%	9.7%	24.5%	43.1%	100%	
アクティブ率	96.4%	19.6%	2.3%	0.6%	0.2%	0.03%	0.9%
アクティブ分布	22.0%	45.5%	18.0%	9.7%	3.2%	1.6%	100%
累積	22.0%	67.5%	85.5%	95.2%	98.4%	100%	

高いと考えられる。

3．時期依存言語モデル[9,10]

ニュース番組の連続音声認識では、ニュースの特徴を生かした言語モデルを構築することが重要である。ニュース番組では、一つの話題が数日間にわたって続くことが多いので、最近のニュース原稿に重みをつけて語彙の設定とn-gramの学習を行うことで、パープレキシティの削減と認識率の向上が期待できる。そこで、認識対象のニュース番組の前日あるいは数日間のニュース原稿を使った、言語モデルの適応化（時期依存言語モデル）を検討している。

時期依存言語モデルは、次のようにして構築する。まず、長期間のニュース原稿から作られたn-gram言語モデルと、最近のニュース原稿から作られたn-gram言語モデルの線形補間によって、両者の最適な確率重みを求める。次に、両言語モデルの学習データの単語数によって正規化しつつ、テキスト重みを求める。このテキスト重みによって長期間のニュース原稿と最近のニュース原稿を足し合わせ、頻度に応じて語彙を設定しなおす。以上をテキスト重みが一定になるまで繰り返して、最終的な言語モデルを時期依存言語モデルとして採用する。

約5年間の長期間のニュース原稿と、評価日の前日のニュース原稿を使って時期依存言語モデル（語彙サイズ20K、bigram）を構築して、評価実験を行った。その結果、テストセット・パープレキシティーは、複数の評価日の平均で、7.4%削減された。また、ニュース音声の認識実験（評価はテストセット・パープレキシティー削減率4.4%の1日）では、認識精度が0.7%向上するという結果が得られた。

今後は、評価日当日の電子原稿を使った時期依存言語モデルを検討する予定である。

4．電子原稿による認識結果の修正 [10,11]

アナウンサーが読み上げる原稿は、記者らが用意した電子原稿のプリントアウトに、加筆・修正を加えたものであることが多い。したがって、音声認識結果の文字列と、もとの電子原稿の文字列を照合することで、ある程度、認識誤りの検出と訂正を行うことができる。

提案する認識結果の修正法では、まず、当日の電子原稿の中で、音声認識結果の文に最も類似する文を検出する。文のマッチングは単語単位で行い、ひらがなと漢字ごとの文字スコアを利用する。文のマッチング・スコアが閾値未満の場合には類似文がなかったものとして、修正を行わない。類似文が検出された場合には、認識結果と電子原稿の単語不一致区間を見つけ、単語数の差が一定値以下の場合にはその区間を認識誤りとみなして、もとの電子原稿で置き換える。評価実験では、認識精度が83.5%（表2の木構造20K）から90.2%まで向上するという結果が得られた。

5．まとめ

本稿では、現在開発中のニュース音声認識システムの概要を報告した。デコーダーは、bigramを用いた単語依存*N*-bestに基づく第1パスと、trigramによるリスコアリングの第2パスで構成される。語彙サイズを5Kから65Kまで変化させ、音素ネットワークがリニアと木構造の場合について検討した。その結果、認識率は木構造の音素ネットワークが、語彙サイズによらずリニアなものよりも高かった。処理時間に関しては、語彙サイズが20Kを超える場合には、現時点ではリニアな方が優位であった。また、ニュースの特徴を生かした時期依存言語モデルと、電子原稿を利用した音声認識結果の修正について述べた。今後は、さらにニュースの特徴を生かしたシステムを検討する予定である。

参考文献

[1] 今井亨、尾上和穂、小林彰夫、安藤彰男、“ニュース音声認識用デコーダーの開発”、音講論集、3-1-12 (1998.9).

[2] R. Schwartz and S. Austin, "A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses," ICASSP-91, pp. 701-704 (1991.5).

[3] H. Ney, R. Haeb-Umbach, B.-H. Tran, and M. Oerder, "Improvements in Beam Search for 10000-Word Continuous Speech Recognition", ICASSP-92, pp. 9-12 (1992.3).

[4] 服部浩明、渡辺隆夫、畑崎香一郎、吉田和永、江森正、古賀真二、“ビームサーチを用いた大語彙音声認識方式の検討”、音講論集、2-6-5 (1997.3).

[5] J.J. Odell, V. Valtchev, P.C. Woodland, and S.J. Young, "A One Pass Decoder Design for Large Vocabulary Recognition," Human Language Technology Workshop, pp. 405-410 (1994.3).

[6] G. Antoniol, F. Brugnara, M. Cettolo, and M. Federico, "Language Model Representations for Beam-Search Decoding," ICASSP-95, pp. 588-591 (1995. 5).

[7] 野田喜昭、松永昭一、嵯峨山茂樹、“単語グラフを用いた大語彙連続音声認識における近似演算手法の検討”、信学技報、SP96-102 (1997.1).

[8] 安藤彰男、宮坂栄一、“ニュース音声データベースの構築”、音講論集、2-Q-9 (1997.3).

[9] 小林彰夫、今井亨、安藤彰男、“ニュース音声認識のための時期依存言語モデル”、音講論集、2-1-17 (1998.9).

[10] A. Kobayashi, K. Onoe, T. Imai , and A. Ando, "Time Dependent Language Model for Broadcast News Transcription and its Post-Correction," to appear in ICSLP-98 (1998. 12).

[11] 尾上和穂、今井亨、安藤彰男、“記者原稿を用いたニュース音声認識結果の修正法”、音講論集、1-6-6 (1998. 3).