音素履歴木を用いたフレーム同期型SSS-LR文音声認識

6G－2

清水 徹　門前 聖康†　松永 昭一
ハラルド シンガー
ＡＴＲ音声翻訳通信研究所　†山形大学

1　はじめに

連続発声された文音声の認識では、文節発声に比較して文法的曖昧性が増加する。従来用いられていた音素同期型HMM-LR[1]では、1つのLRスタックを1つのセルで表現するために、文法的曖昧性の増加に伴い同一音素列に対する仮説数も増加する。このため、音素照合が重複し処理時間が多くなる問題点があった。また、音素同期型探索では尤度の正規化が必要であるが、正規化の精度が悪いと認識率が低下する問題点があった[2]。本論文では、音素履歴木を用いたフレーム同期型SSS-LR連続音声認識手法を提案する。本手法は、統語解析部に(1)音素履歴木(2)状態ネットワークを導入し、文法的曖昧性の増加に起因する仮説数の増加を抑えた効率的なフレーム同期型処理を実現した。なお、HMMモデルは音素照合の高精度化を図るためSSSアルゴリズムで作成したコンテキスト依存HMMモデルを使用した[3][4]。また、本手法を用いた特定話者・不特定話者文認識実験結果について述べる。

2　フレーム同期型SSS-LR認識系の構成

本論文で提案するフレーム同期型SSS-LR音声認識系の構成を図1に示す。

探索部 入力フレーム番号, 音素履歴, 予測音素,HMM状態の4つで決まる仮説をGrid仮説と呼ぶ。Grid仮説をOne Pass Vitervi サーチに基づき各時刻毎に展開する。仮説の評価値に基づき枝刈りを行なう。

統語解析部 探索部の検出音素に基づきLRパーザを駆動し、音素履歴木および状態ネットワークを展開する。展開結果で得られる予測音素（コンテキスト依存HMMモデルの場合はさらに後続音素）を探索部に送る。

状態ネットワーク LRスタックをグラフ構造化したもので、ノードはスタックの状態、アークはアクション実行時の遷移を表す(文献[5]のFSAネットワークと等価)。ネットワークは音素検出に際してのみ動的に展開する。一度展開したノード系列に対するアクションは、パーザを駆動せずトレースのみで高速に行なう。

音素履歴木 音素系列を木構造で表したもので、Grid仮説と状態ネットワークを結び付ける。ノードには、状態ネットワークで得られる予測音素の和集合を保持する。

図2にHMMモデル,Grid仮説, 音素履歴木, 状態ネットワークの相互関係を示す。

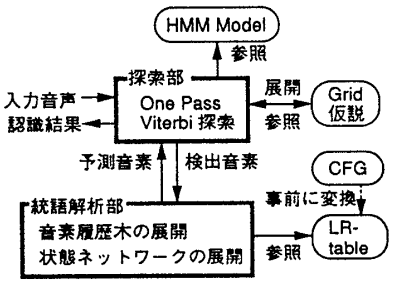

図1: フレーム同期型SSS-LR音声認識系の構成

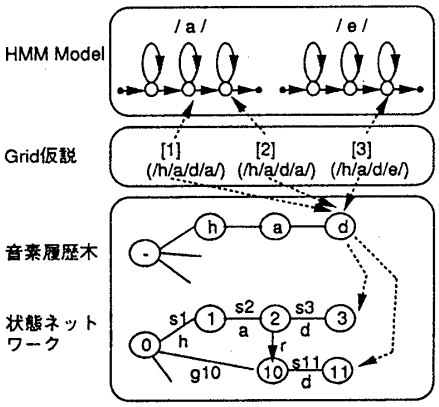

図2: HMMモデル,Grid仮説, 音素履歴木, 状態ネットワークの相互関係

3　音素履歴木の機能

探索部と統語解析部の分離 Grid仮説は、入力フレーム番号、音素履歴、予測音素、HMM状態でのみ区別される。従って、音素履歴木の1つのノード(例えば、"h-a-d"の"d")が指している状態ネットワークのノード(""h-a-d"の"d"ではノード3,11の2状態)数はGrid仮説の数に影響を与えない。すなわ

Frame Synchronous SSS-LR Sentence Recognition using a Phoneme Hypotheses Trie
Tohru Shimizu, Seikou Monzen†, Shouichi Matsunaga and Harald Singer
ATR Interpreting Telecommunications Research Laboratories
†Yamagata University

表 1: 特定話者文節認識実験における文節認識率 (%) と 1 文節あたりの CPU 時間 (sec)

beam width	top1	top5	top10	cputime
50	84.7	87.7	87.7	0.8
100	90.7	94.6	94.7	1.9
200	93.6	98.7	99.0	5.1
400	93.9	99.3	99.7	13.9
800	94.0	99.4	99.9	39.1
250*	91.0	98.6	−	6.3

* は音素同期 SSS-LR

ち、音素履歴木は同一音素系列に対する文法仮説のパッキング機能を有する。音素同期 HMM-LR や北らのフレーム同期型処理手法 [5] ではこの種のパッキングは行なわれていない。逆に、音素履歴木や状態ネットワークの展開を仮説の音素検出時のみ行なうことにより音素照合部の曖昧性は統語解析部の負荷にならない。同一 Grid 仮説のマージや N-best 候補を求めるための仮説の枝刈りは探索部のみで実行可能である。(音素履歴, 予測音素,HMM 状態が同じ仮説はスコアの最も大きい仮説を残せば良い。音素系列が異なっても、状態ネットワークのノード,HMM 状態が同一の仮説は以降の時刻の音響尤度が等しいので、その時点で尤度の大きい順に N 個の候補を残せば N-best が求められる。) 音素同期探索では、かなり特殊なタスクでスタックの併合が実現されている [6] 以外、音素系列の長さが異なる仮説のスタックの併合は行なわれていない。

仮説比較の高速化 フレーム同期型処理で頻繁に発生する同一 Grid 仮説 (入力フレーム番号, 音素履歴, 予測音素,HMM 状態が同じ仮説) のマージが音素履歴木のノードの比較のみで高速に行なうことができる。

4 認識実験

音声資料は国際会議に関する問い合わせタスク (文節認識 : 12 対話,701 文節、文認識 : 7 対話,136 文章)、文脈自由文法は、タスク依存型文節内文法 (perplexity 2.66) および文文法 (perplexity 2.79)、音響モデルは HMnet[4](特定話者モデル : 5240 単語の偶数番目の単語から学習, 状態数 600、不特定話者モデル : 話者重畳型 (285 名), 構造決定 = 5240 単語の偶数番目の単語, 状態数 200, 学習 = 音韻バランス文 50 文 (文節発声)) を用いた。音響分析は、標本化 12kHz, フレーム周期 5ms, ハミング窓 30ms、特徴量は、1 〜 16 次 LPC ケプストラム,1 〜 16 次 Δ LPC ケプストラム, log パワー, Δ log パワーを用いた。

特定話者モデルを用いた文節および文認識結果を表 1, 表 2にそれぞれ示す。文節認識実験では音素同期型の結果も併せて示す。表 1より、ビーム幅 200 のフレーム同期型の認識率・ cpu 時間はともにビーム幅 250 の音素同期型の性能を上回り、フレーム同期型処理の有効性が確認された。また、表 2より、文認識では文節認識に比較してかなり大きなビーム幅が必要であるものの、ある程度のビーム幅を確保すれば (3,200、メモリサイズ 11Mbyte) 文音声認識が可能であることが示された。また、表 3に示す不特定話者モデルを用いた文認識実験では、話者によっては特定話者モデルと同程度の認識率が得られたが、平均的には低い認識率にとどまった。これは、特定話者モデルの状態数が 200 であることも原因の一つと考えられる。

表 2: 特定話者文認識実験における文認識率 (%) と 1 文章あたりの CPU 時間 (sec)

beam width	top1	top5	top10	cputime
100	50.0	52.2	52.2	3.7
200	56.6	59.6	59.6	5.6
400	63.2	70.6	70.6	10.1
800	65.4	76.5	76.5	21.1
1600	69.1	80.9	80.9	44.3
3200	70.6	82.4	83.1	87.4

表 3: 不特定話者文認識実験における文認識率 (%)

speaker	top1	top5	top10
MAU	68.4	84.6	84.6
MIK	43.8	51.8	54.5
MTK	47.3	65.2	67.9
FAK	70.5	81.3	86.6
FAS	38.4	58.9	62.5
FNY	37.5	62.5	70.5
平均	51.0	67.4	72.4

5 むすび

本論文では、連続発声された文音声の認識を目的とした認識系として、音素履歴木を用いたフレーム同期型 SSS-LR 連続音声認識手法を提案した。また、特定話者・不特定話者音声認識実験を行ない、本手法により文音声認識が可能であることを示した。今後、探索の高速化について検討を行なう予定である。

参考文献

[1] 北, 川端, 斎藤: "HMM 音韻認識と拡張 LR 構文解析法を用いた連続音声認識", 情処論, 31, 3, pp.472-480 (1990).

[2] 野田, 嵯峨山: "前向き尤度を用いた A^* ビーム探索による HMM-LR 音声認識"信学技報 SP94-23 (1994).

[3] 鷹見, 嵯峨山: "逐次状態分割法による隠れマルコフ網の自動生成", 信学論 (D-II), J76-D-II, 10, pp.2155-2164 (1993).

[4] 永井, 鷹見, 嵯峨山, シンガー: "隠れマルコフ網と一般化 LR 構文解析を統合した連続音声認識", 信学論 (D-II), J77-D-II, 1, pp.9-19 (1994).

[5] 北, 矢野, 森元: "LR パーザ制御による One-Pass 型連続音声アルゴリズム"信学技報 SP94-13 (1994).

[6] 南, 鹿野, 高橋, 山田: "音韻環境依存 HMM と候補のマージを用いた不特定話者大語彙連続音声認識"音講論 2-7-5, pp.79-80 (1993).