卒業論文

連続音声認識システムの構築に関する研究

東北大学工学部 電気・情報系
阿曽研究室 4 年
中野 英行

目 次

1 序論 **4**

1.1 研究の背景 4
1.2 連続音声認識 4
1.3 連続音声認識システムの構成 5
1.4 研究の目的 8
1.5 本論文の構成 8

2 HMnet を用いた音素ラベリング **9**

2.1 はじめに 9
2.2 HMnet(Hidden Markov Network) 9
2.3 HMnet を用いた音素ラベリング 12

3 評価実験 **16**

3.1 はじめに 16
3.2 特定話者での実験 16
3.3 不特定話者での実験 17
3.4 まとめ 21

4 結論 **22**

表目次

3.1 全音素の正解率 17
3.2 挿入率 17
3.3 脱落率 17
3.4 全音素の正解率 19
3.5 挿入率 19
3.6 脱落率 20

図目次

1.1 連続音声認識システムの概念図 7

2.1 left-to-right HMM 10
2.2 ergodic HMM 10
2.3 HMnet 11
2.4 逐次状態分割法によって構成されたHMnetの例 12
2.5 Viterbiアルゴリズム 15

第１章

序論

1.1　研究の背景

音声は情報伝達手段として，手軽でしかも能率のよいものであるが，それ以外にもいくつかの利点がある。すなわち，あらためて特別な訓練を受ける必要がなく，道具が不要で，別の行動をしながら情報のやりとりができる。
音声によって人間とコンピュータの間のコミュニケーションをスムーズに行なうためには，コンピュータには人間が発声した音声から言語情報を自動的に抽出する処理部と，伝達すべき音声を生成する処理部が必要である。そのうち，音声信号から言語情報を自動的に抽出し，音声に対応するテキストを自動的に決定することは通常，音声認識と呼ばれている。これまで，音声認識に関する研究は数多くなされており，現在では簡単な孤立単語の認識装置が製品化され，その結果簡単な機械との対話が実現した。しかし，利用する側の人間にとって語彙数や発声速度などの制約は少なくない．このような背景のもと，話し言葉のような自然発話を対象にした連続音声認識の研究に重点が置かれるようになったがその理論や技術は必ずしも確立されていない。そこで本論文では，従来からある基本的な考え方を参考にして連続音声認識システムを実現するにはどうしたらよいかを考えることを目標とする。まず本章では，自然発話を対象にした連続音声認識の基礎となる，文音声認識を中心とした連続音声認識とシステムについて簡単に説明する。

1.2　連続音声認識

文は，一般に，なんらかの構造と意味をもつので，文音声は文の構造や意味に関する制約を受けたものである。したがって，文音声のような連続音声認識では自然言語処理が音響音韻的処理と並んで重要である。また，少語彙の孤立単語認識では，認識の基本単位を認識対象の単語そのものにすることが最良と考えられる (認識の基本単位が大きいほどコンテキストの違いによる音響的変動を吸収でき，認識精度は上る) が，一般に連続音声認識の場合には，モデルの数が膨大なものとなるので，実際上単語を認識の単位とすることはできない。通常，音素や半音節などを認識の基本単位として用いることになる。少語彙の

孤立単語認識などと比較して連続音声認識において次のような点が特に難しい問題である。

- 連続音声の発声速度は，各単語を区切って発声する場合に比べてかなり速く，連続音声中の音素の音響的変動は非常に大きい
- 音素コンテキストによるの音素の音響的変動が大きい
  例えば，
  a r a y u r u....
  の2つのaは前後の音素が違うので，同じaでもその特徴はかなり異なる
- 連続音声中の単語あるいは文節の間の境界は，音響的には特別のものではなく，他の部分の音韻間の境界と基本的に同じであって，それらを音響的に区別することは難しい

1.3 連続音声認識システムの構成

以下に代表的な連続音声認識システムの構成を示す。

- 1. 音響処理部
  入力された波形データを短区間 (通常 5ms から 10ms 程度) に切りわけ，短区間ごとに音声の特徴をよく表現した特徴量へと変換する。現在最もよく用いられている特徴量は cepstrum である。cepstrum は人間の発声機構を考えて考案された特徴量であり，音声の特徴をよく表現している。
  人間は声帯で一定の励振波形を発声させ，声道の形を変えることで波形を変化させて様々な音声を発声する。つまり音声は声帯で発声された励振波形に声道の伝達関数を畳み込むことで生成される。cepstrum は，入力波形をフーリエ変換した後に対数をとり，更に逆フーリエ変換をして低次の項をとることで，声道の形からくる特徴をうまく取り出している。また，cepstrum の時間的な動きも特徴量にいれるために，cepstrum の時間方向の微係数も特徴量とするのが一般的である。

- 2. 音素認識部
  - 音素セグメンテーション
    音声信号を適当な信号処理によって音素に対応した時間区間に区分化できれば，音素認識が比較的容易になる。このような時間区分化を音素セグメンテーションと呼ぶ。
    音素セグメンテーションの方法には，陽 (explicit) に音素セグメンテーションを行なう方法と音素ラベリングの結果として音素セグメンテーションを行なう方法があるが，自然な発話の連続音声の信号の音素セグメンテーションを自動的に

正確に行なうことは困難なので，現時点では，後者の方法がよく用いられている。

- 音素ラベリング
  得られた特徴量ベクトル系列をもとに，音素の認識を行なう。あらかじめ各音素ごとにその音素の特徴をよく表現するような音素モデルを用意しておき，特徴量ベクトル系列と音素モデルを連結したものとの距離を計算，その距離が最も近いものを出力する。出力は音素ラベル系列となる。

- 3. 言語処理部
  音素ラベル系列から，単語辞書や文法辞書，言語モデルなどを用いて文を生成する。まず，辞書の単語・文節に対応する音素記号列と音素ラベル系列との記号列マッチングなどによって，単語・文節の候補を選ぶ。次に，得られた文節の候補をもとにして，構文情報や意味情報，文脈情報あるいはタスクによる制約などを利用しながら，文を生成する。

図 1.1 に連続音声認識システムの概念図を示す。

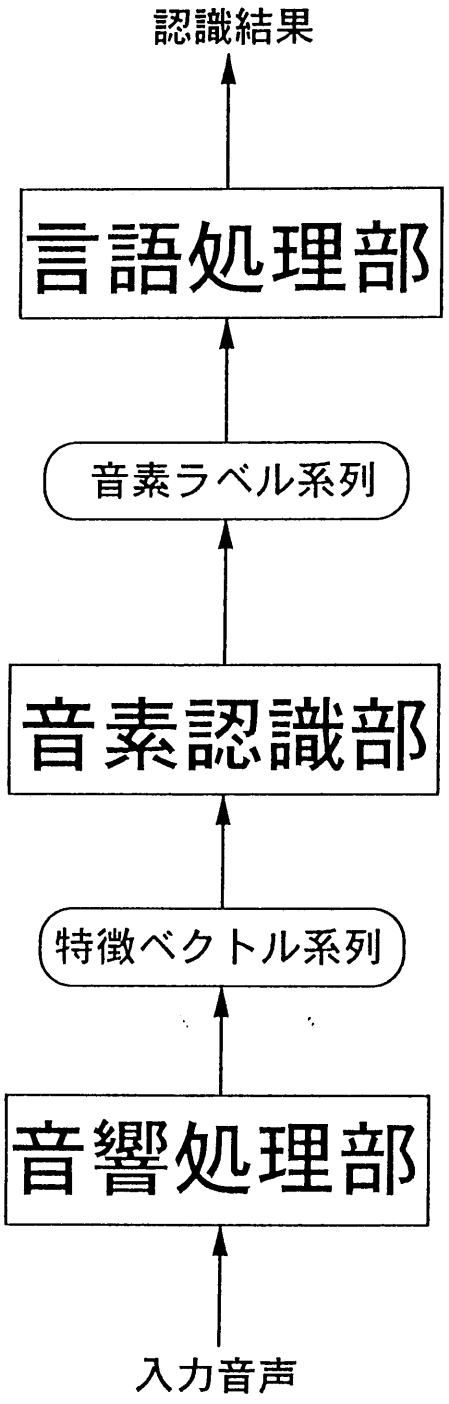

図 1.1: 連続音声認識システムの概念図

1.4 研究の目的

従来から知られている基本的な考え方や基礎技術を参考にして，連続音声認識システムを実現するにはどのようにすればよいかを考える。

1.5 本論文の構成

第 1 章 本研究の背景，及び本研究の目的を述べる。

第 2 章 HMnet を用いた音素ラベリングについて説明する。

第 3 章 音素ラベリングの評価実験を行なった結果を示す。

第 4 章 本研究の成果と今後の課題について述べる。

第 2 章

HMnet を用いた音素ラベリング

2.1 はじめに

現在よく用いられている音素ラベリングの手法は，HMM(Hidden Markov Model) を用いる方法である。HMM を用いる方法は，発声の音響的変動などを統計的に扱える，多量のサンプルから自動学習できる，音素のセグメンテーションをする必要がないなどの利点がある。HMM の中でも，最近では，HMnet(Hidden Markov Network) を用いる方法が有効であると考えられているので，本章では，まず HMnet について簡単に説明し，また HMnet を用いた音素ラベリングの方法を述べる。

2.2 HMnet(Hidden Markov Network)

HMM の形状としては，図 2.1 のような left-to-right 型が一般的である。この型の利点は，前の状態へ戻る遷移がないために，時間的な変化をよく表現できる点である。

ところが，多量のサンプルから音素ごとに HMM を学習する時に，すべてのサンプルが必ず同じ状態を遷移していくために，各状態の出力分布が拡がってしまいあまりに特徴の違うサンプルを 1 つの HMM で表現するには無理が生じる。これに対して，図 2.2 のようなすべての状態間の遷移を許した ergodic HMM と呼ばれる型は，サンプルの特徴によって通る状態が変るので，left-to-right HMM のように必要以上に出力分布が拡がることはない。しかし，HMM が小規模な場合は各状態が ergodic な結合をしているためにサンプルの時間的遷移をうまく表現できず，またそれを表現するために大規模なモデル構成にすると，パラメータの推定が大変にまってしまい，あまり一般的ではない。

そこで，両者の利点を取り入れた型として，図 2.3 のような HMnet が考えだされた。この型は，left-to-right HMM と同じように時間的な変化をよく表現でき，またすべてのサンプルが同じ状態を遷移していくこともないので，出力分布が拡がりすぎることもない。

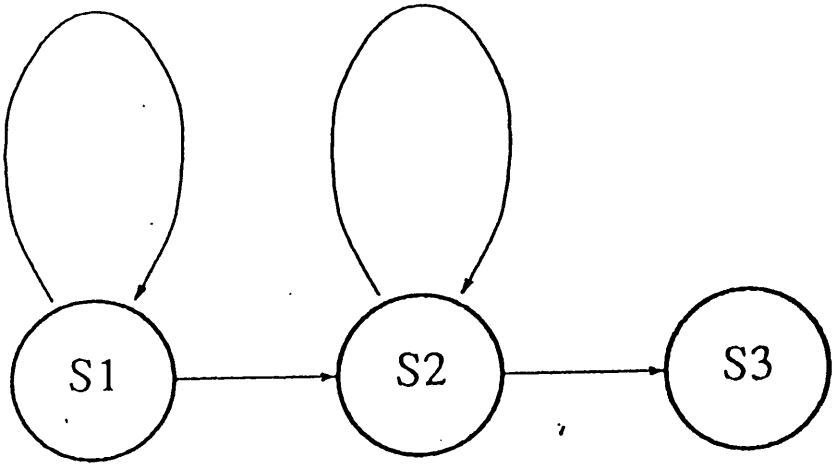

図 2.1: left-to-right HMM

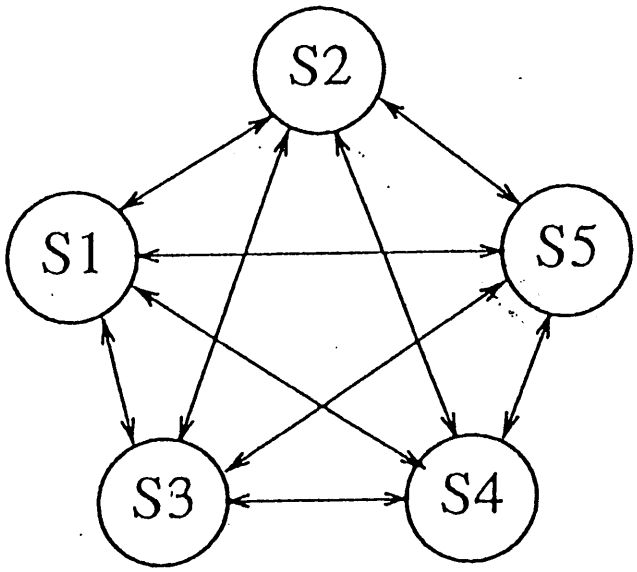

図 2.2: ergodic HMM

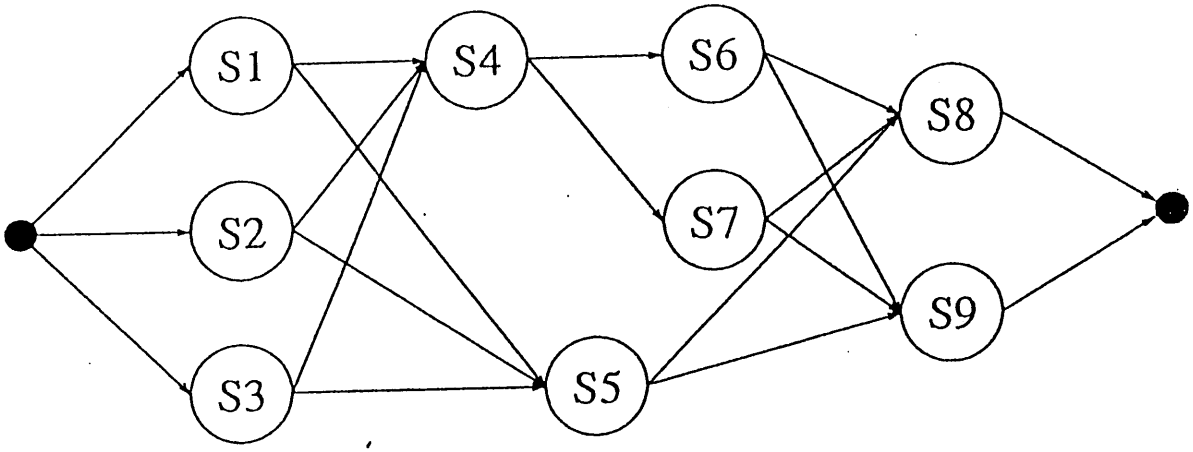

図 2.3: HMnet

HMnet を構成するアルゴリズムのに逐次状態分割法 (Successive State Splitting : SSS)[参考文献 3,4] やそれを改良した SSS-free[参考文献 5] がある。ここでは，アルゴリズムの詳細は省き，逐次状態分割法によって得られた HMnet の例を図 2.4 に示す。これは，HMnet のなかの音素 /g/ に対応する部分だけ抜き出したものである。図 2.4 に示されているように，各状態は受理する先行音素と後続音素のリストを持つ (* は，すべての音素を受理)。そこで，音素の認識時に認識すべき音素の前後の音素を仮定してやれば，それに対応するパスがただ 1 本に決まる。例えば先行音素が/a/ 後続音素が /e/ であると仮定した場合，このときの /g/ のモデルは，#30 - #19 - #4 というパスになり，通常の left-to-right HMM とみなせる。

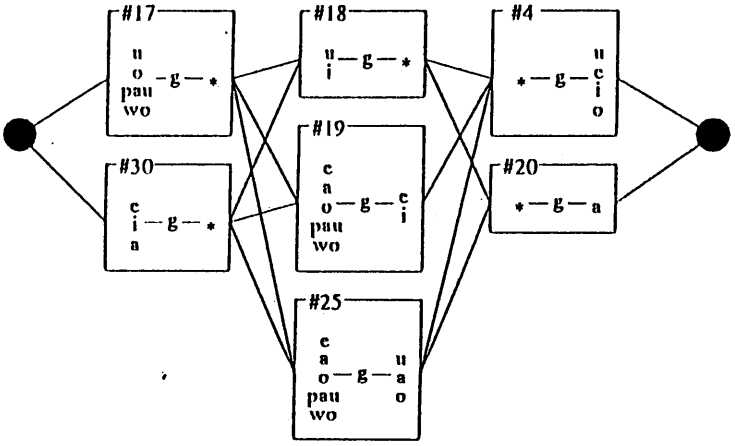

図 2.4: 逐次状態分割法によって構成された HMnet の例

2.3 HMnet を用いた音素ラベリング

2.2 節で述べたようにモデルとして HMnet を用いる利点は多いので，今回は SSS-free で構成された HMnet を使い音素ラベリングを行なう。
ところで，1.2 節でも述べたように連続音声認識において，音素コンテキストの違いによる音響的変動が大きいという問題があったが，HMnet を用いることによってこの問題は回避できる。その理由は，HMnet を構成するアルゴリズムである SSS や SSS-free は，もともと効率のよいコンテキスト依存モデル (前後の音素を考慮したモデル) を構成するために提案された手法だからである。また，音素区切が明確でないという問題には，尤度最大となる音素ラベル系列を求め，その結果として音素区切を決める方法で対処する。それを効果的に行なうのが以下に述べる Viterbi アルゴリズムである。
未知入力 $X = x_1, x_2, ..., x_t, ..., x_T$ に対して，次のように認識する。すべての音素 p について，以下の 4 ステップにしたがって最大尤度をとる系列を計算する。図 2.5 に Viterbi アルゴリズムを示す。

Step 1—Initialization

$$Q_1(p,1) = \log b_p(x_1|1), \qquad 1 \leq p \leq N \tag{2.1}$$

$$\Psi_1(p,j) = (0,0), \qquad 1 \leq p \leq N, 1 \leq j \leq p_s \tag{2.2}$$

Step 2—Recursion

$$\text{For } 2 \leq t \leq T, \quad 1 \leq p \leq N, \quad 1 \leq j \leq p_s \tag{2.3}$$

$$Q_t(p,j) = \left\{ \begin{array}{ll} \max_i \{Q_{t-1}(p,i) + \log a_p(i,j)\} & j > 1 \\ \max \left\{ \begin{array}{l} \max_r \{Q_{t-1}(r,E_r) + \log T_r(r,p)\} \\ Q_{t-1}(p,1) + \log a_p(1,1) \end{array} \right. & j = 1 \end{array} \right\} + \log b_p(x_t|j) \tag{2.4}$$

$\underline{j > 1}$

$$\Psi_t(p,j) = (p, \text{argmax}_i \{Q_{t-1}(p,i) + \log a_p(i,j)\}) \tag{2.5}$$

$\underline{j = 1}$

IF $\max_r \{Q_{t-1}(r,E_r) + \log T_r(r,p)\} > Q_{t-1}(p,1) + \log a_p(1,1)$

$$\Psi_t(p,j) = (\text{argmax}_r \{Q_{t-1}(r,E_r) + \log T_r(r,p)\}, E_r) \tag{2.6}$$

ELSE

$$\Psi_t(p,j) = (p,1) \tag{2.7}$$

Step 3—Termination

$$P^* = \max_p Q_T(p,E_p) \tag{2.8}$$

$$(p,j)_T^* = (\text{argmax}_p Q_T(p,E_p), E_p) \tag{2.9}$$

Step 4—Path (state sequence) backtracking

$$\text{For } \quad t = T-1, T-2, ..., 1 \tag{2.10}$$

$$(p,j)_t^* = \Psi_{t+1}((p,j)_{t+1}^*) \tag{2.11}$$

ここで,

$a_p(i,j)$：音素 p のモデルで状態 i から状態 jへの遷移確率

$b_p(x_t|j)$：音素 p のモデルで状態 jがフレーム目のベクトル x_tを出力する確率

$T_r(r,p)$：音素 rから音素 p への遷移確率

E_p: 音素 p のモデルの最終状態番号

N：音素モデルの数

p_s：モデル p の状態数

最終的に $\max_p Q_p(E_p,T)$ となる $Q_p(E_p,T)$ からバックポインタをたどることで，最適な音素系列を得る。モデル r の最終状態からモデル p の初期状態への遷移確率 $T_r(r,p)$ は学習文章中に現われた音素の接続に対して 1 に，それ以外は 0 にした。

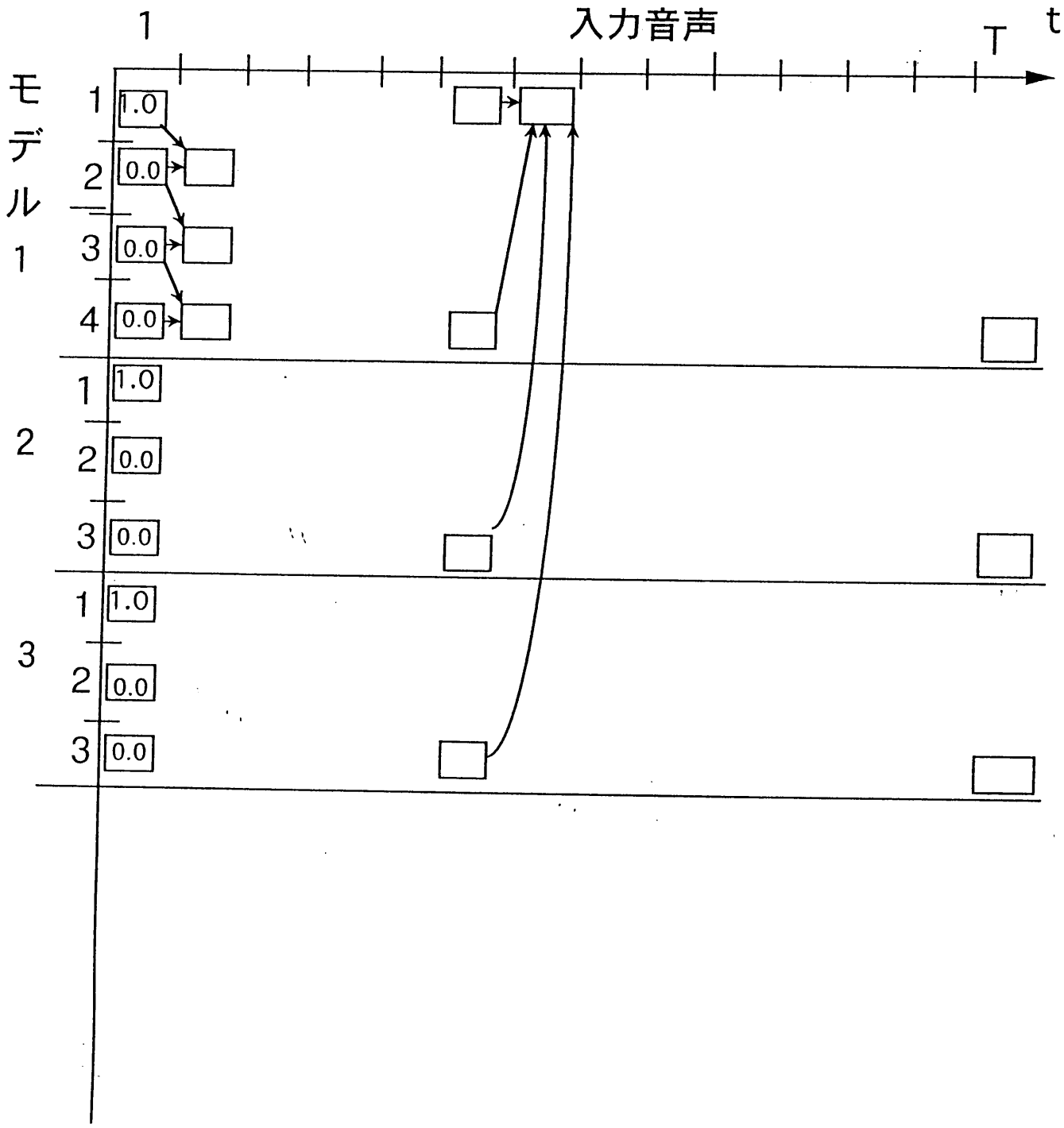

図 2.5: Viterbi アルゴリズム

第 3 章

評価実験

3.1 はじめに

2 章で述べた方法で実際に音素ラベリングを行なった。今回，HMnet の学習，また実験には，音響学会が提供する音声データを用いた。実験は特定話者，不特定話者に対して行ない，次節以降にその結果を示す。

3.2 特定話者での実験

実験は 2 人の話者に対して，それぞれ，自分が発声した 103 文章で学習した HMnet を用いて行なった。実験条件は以下のとおりであり，表 3.1、表 3.2，表 3.3 に全音素の正解率 ($\frac{[\text{正解音素系列中の総音素数}]-[\text{置換誤り数}]-[\text{挿入誤り数}]-[\text{脱落誤り数}]}{[\text{正解音素系列中の総音素数}]}$)，挿入率，脱落率を示す。

実験条件

特徴量

logpow，cep(16)， Δlogpow，Δ cep(16) からなる 34 次元ベクトル

分析条件

サンプリング周波数:12KHz
16bit 量子化
20ms ハミング窓
フレーム周期：5ms

学習サンプル

男性 1 名が発声した 103 文章
女性 1 名が発声した 103 文章

テストサンプル

話者：男性 1 名 (MAT0001)、女性 1 名 (MAT1001)
学習サンプル中になかった 10 文章

全音素の正解率は表 3.1 のとうりだが，特定話者での実験にしては低すぎると思われる。この一番大きな原因は，表 3.2 に見られるように挿入率がかなり高いことである。また脱落率は，平均で 5%以内におさまっており許容範囲と思われる。

状態数	100	200
MAT0001	60.1(%)	63.0(%)
MAT1001	62.5(%)	67.1(%)
mean	61.3(%)	65.1(%)

表 3.1: 全音素の正解率

状態数	100	200
MAT0001	15.0(%)	13.6(%)
MAT1001	16.4(%)	12.6(%)
mean	15.7(%)	13.1(%)

表 3.2: 挿入率

状態数	100	200
MAT0001	4.9(%)	5.3(%)
MAT1001	2.1(%)	3.1(%)
mean	3.5(%)	4.2(%)

表 3.3: 脱落率

3.3 不特定話者での実験

実験条件は以下のとうりである。

実験条件
特徴量，分析条件は特定話者での実験と同じ。

学習サンプル
　　話者以外の 36(人)*50 文章

テストサンプル
　　話者：男性 2 名、女性 2 名
　　学習サンプル中にあった 10 文章 (1)
　　学習サンプル中になかった 10 文章 (2)

表3.4 からわかるように，話者によって正解率にかなりばらつきがある。例えば，MAT0001 と MAT1001 では約 20%の差がある。また，テストサンプルとして，学習サンプル中にあった文章を用いた場合 (1) の方が，学習サンプル中になかった文章を用いた場合 (2) よりも正解率がよいと思っていたのだが，状態数が 100 と 200 の HMnet では (2) の方がよい。さらに気になるのは，(1) は状態数の増加とともに順調に正解率がよくなっているが，(2) では状態数 200 と状態数 300 で正解率はほとんど変わっていないことである。状態数をもっと増して検討する必要がある。

(1)

状態数	100	200	300
MAT0001	31.5(%)	40.6(%)	48.4(%)
MAT1001	50.3(%)	59.6(%)	65.0(%)
TSU0001	41.4(%)	49.5(%)	53.1(%)
TSU1001	45.8(%)	53.2(%)	56.4(%)
mean	42.3(%)	50.7(%)	55.7(%)

(2)

状態数	100	200	300
MAT0001	37.6(%)	41.9(%)	44.8(%)
MAT1001	53.9(%)	63.1(%)	63.9(%)
TSU0001	43.7(%)	54.5(%)	54.1(%)
TSU1001	43.8(%)	49.6(%)	48.8(%)
mean	44.8(%)	52.3(%)	52.9(%)

表 3.4: 全音素の正解率

(1)

状態数	100	200	300
MAT0001	22.2(%)	14.3(%)	11.9(%)
MAT1001	13.7(%)	8.5(%)	9.3(%)
TSU0001	14.9(%)	11.3(%)	12.3(%)
TSU1001	12.9(%)	7.2(%)	9.6(%)
mean	15.9(%)	10.3(%)	10.5(%)

(2)

状態数	100	200	300
MAT0001	18.3(%)	12.4(%)	13.3(%)
MAT1001	15.4(%)	8.5(%)	9.5(%)
TSU0001	13.7(%)	7.1(%)	8.8(%)
TSU1001	10.9(%)	6.8(%)	9.4(%)
mean	14.6(%)	8.7(%)	10.3(%)

表 3.5: 挿入率

(1)

状態数	100	200	300
MAT0001	6.8(%)	8.3(%)	6.4(%)
MAT1001	3.9(%)	5.6(%)	3.5(%)
TSU0001	6.5(%)	5.3(%)	5.3(%)
TSU1001	7.8(%)	8.8(%)	7.6(%)
mean	6.3(%)	7.0(%)	5.7(%)

(2)

状態数	100	200	300
MAT0001	6.6(%)	8.5(%)	5.4(%)
MAT1001	4.4(%)	6.4(%)	4.6(%)
TSU0001	7.1(%)	7.3(%)	6.6(%)
TSU1001	11.6(%)	12.8(%)	11.8(%)
mean	7.4(%)	8.8(%)	7.1(%)

表 3.6: 脱落率

参考までに，2 つほど音素ラベリングの結果を以下に示す。これは，(2) の実験で状態数 200 の HMnet を用いて，MAT1001 が発声した文章をラベリングしたものの一部である。

- 55 歳だって うれしいときは うれしいのだ。

  k o: z j u: j o k o s a i d a t e t o u r e s i t a k j a o a r e s i n o d a

- セミが 鳴き 夕日 を あびた 絹雲が 淡い さんご色に 染まっていた。

  s e m i r a n a c i k i Q p j u e s i j o a b i t a t e n a o m o b a w a i s a w a k o i r o i s o m a t e i t a

(注)
o: のように :のついているものは (オー) とのばす。
j は y のかわり。つまり，(ヤ) は ja と表記。

3.4 まとめ

音素ラベリングの評価実験を特定話者と不特定話者について行なった。どちらの実験でも挿入誤りが多いのが目立ったが，これはモデルに継続時間制御のパラメータを導入することである程度対処できると思う。
不特定話者の実験では，話者による正解率のばらつきがかなり大きいことがわかり，話者適応の必要性を強く感じた。また，今回，状態数 300 までしか実験しなかった理由は，これ以上状態数を増すと記憶容量が大きくなりすぎて無理があったためである。

第 4 章

結論

HMnet を用いて音素ラベリングを行なったが，問題点が多く，あまりよい成果が得られなかった。3 章ではふれなかったが，2.3 節で述べたアルゴリズムには，重大な問題が残っている。1 文を音素ラベル系列に変換するのに，HMnet の状態数が 100 のもので 1 分弱，状態数 300 のものだと 30 分もかかってしまう。タスクを限定しモデル数を減らせばこの問題はとりあえず回避できるが，自然発話などを対象とした連続音声認識システムの構築を考えた場合，あまりタスクを限定することはできない。日本語では音素は 24 種類あり，前後の音素を考慮した場合の組み合せは，$24^3 = 13824$ 種類である。そのうち実際に日本語にでてくる組み合せは 3000 種類あまりといわれているから，最低それくらいのモデル数は必要と思われる。今回用いた状態数 300 の HMnet のモデル数が約 3000 ぐらいである。
今後の課題としては，音素認識部の計算時間の大幅短縮，話者適応の導入，言語処理部の作成などがあげられる。

謝辞

本研究を進めるにあたり，数多くの御指導とともにこの研究の機会を与えてくださりました東北大学工学部通信工学科阿曽弘具教授に心から感謝致します。また音声認識の分野におきましては，東北大学情報科学研究科牧野正三教授，阿曽研究室鈴木基之氏に終始御指導戴きましたことを深く感謝いたします。
日々の研究におきましては，阿曽研究室成富敬助手，後藤英昭助手，森大毅氏をはじめとする阿曽研究室の皆様に数多くの御意見，御討論戴いたことに深くお礼申し上げます。

参考文献

1. 中川聖一:「確率モデルによる音声認識」
   電気通信情報学会 (1988)
2. 今井 聖:「音声認識」
   共立出版 (1995)
3. J.Takami and S.Sagayama: 「A successive state splitting algorithm for effecient allphone modeling」
   Ploc.ICASSP'92,pp.1835-1842(1993)
4. 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,J76-D-II,No10,pp. 2125-2164(1993)
5. 鈴木基之:「音響類似性に基ずく隠れマルコフ網を用いた連続音声認識に関する研究」
   修士学位論文, 東北大学大学院工学研究科 (1995)