修士学位論文

音響セグメントモデルを用いた
音声認識の高精度化に関する研究

東北大学大学院工学研究科
電気・通信工学専攻
林 貴文

目 次

図 目 次

表目次

第1章

序論

1.1 研究の背景

近年の計算機処理能力の向上に伴い、音声認識という用語は一般的なものになりつつある。音声は人間が日常的に用いる最も一般的なコミュニケーション手段の一つであるため、人間とコンピュータの間のインタフェースとして音声を利用したいという要求は古くからあった [1]。インタフェースとして音声が利用できれば、キーボードやマウスなどのようなボタン操作が不要となり、利用者がより快適にコンピュータを扱うことができるようになる可能性がある。この要求に答えるため数多くの研究が成されており、現在では音声認識を組み込んだシステムがいくつか実用化されている。既に商品化されているものの例として、マイクに向かってしゃべった声を、そのまま漢字かな混じりの文章として計算機に入力する機能を持った音声入力ワープロソフトなどがある。また、いくつかの研究機関では大語彙のディクテーションシステムやタスク依存の対話システムなどの開発が行われている (例えば [2])。

しかしながら、これらのシステムには利用者に課せられた制約がある場合も少なくない。現在のシステムの多くは正確な認識結果を得るためには利用者は明瞭な発声をしなくてはならない。このように、未だコンピュータとのインタフェースとして音声を “自然” に利用できていないというのが現状である。利用者がもっと自然で、かつ快適にこれらのシステムを利用できるように、制約の少ないシステムが求められている。制約が少なく、かつ高精度な音声認識システムの構築が可能となればさまざまな応用が期待できる。

1.1.1 大語彙連続音声認識システムの概要

現在は利用者に対する制約が比較的少ない、大語彙の連続音声認識についての研究が主流である。この節では一般的な大語彙連続音声認識システムについて簡単に述べる。 図 1.1

にシステムの構成を示す。大きく分けて、特徴抽出、マッチング、言語処理の 3 つの処理から成る [3]。

- 特徴抽出

  入力音声の波形データを短い時間幅で切り出し、音声の特徴をよく表わすような特徴量ベクトルに変換する。この処理の前後に、雑音に対処する処理を行うこともある。現在、音声認識でよく用いられている特徴量の一つにケプストラム (cepstrum) がある。この処理により、音声波形から認識に有効な特徴パラメータが抽出され、入力音声は特徴量ベクトルの時系列として表わされる。

- マッチング

  入力音声を変換した特徴量ベクトル系列と辞書の類似度を調べ、高い類似度を示した辞書の単語を認識結果の候補としていく。マッチングの際に使用する辞書は、音響モデルによって表現される。辞書の精度、すなわち音響モデルの精度は認識性能に大きく関わる要因の一つである。

- 言語処理

  音響的なマッチングによる類似度に統計的な言語モデルのスコアを加え、評価する。この処理はマッチングと同時に行われることもあり、これにより認識結果の絞り込みがなされる。更に、文法的な知識に基づく処理や意味解析を行って最終的な認識結果を得る場合もある。

高精度な大語彙連続音声認識システムの実現のためには、認識率を向上させるような特徴抽出法の検討や、前述した処理の中で用いる音響モデル、言語モデルの高精度化などが必要である。特に、特徴量に変換された入力音声と直接の比較を行う音響モデルの高精度化は、大語彙連続音声認識システムの性能を向上させる上で、非常に重要な要因の一つであると言える。そこで本研究では、マッチングの際に用いる音響モデルに着目する。

1.1.2 音素に基づく音響モデル

大語彙連続音声認識システムで使用する音響モデルについて述べる。システムで用いる辞書には認識したい全ての単語を登録しておく。そして、登録した単語を表わす音響モデルを用意する必要がある。性能のよい音響モデルを作成するためには多量の学習サンプル用の音声データが必要である。大語彙の認識の場合、単語の数だけ音響モデルを作成することは、必要となる学習サンプルの問題があり現実的ではない。そのため、単語毎に音響モデルを作成するのではなく、単語よりも小さな単位である音素を単位として音素毎に音響モデルを作成するのが一般的である。

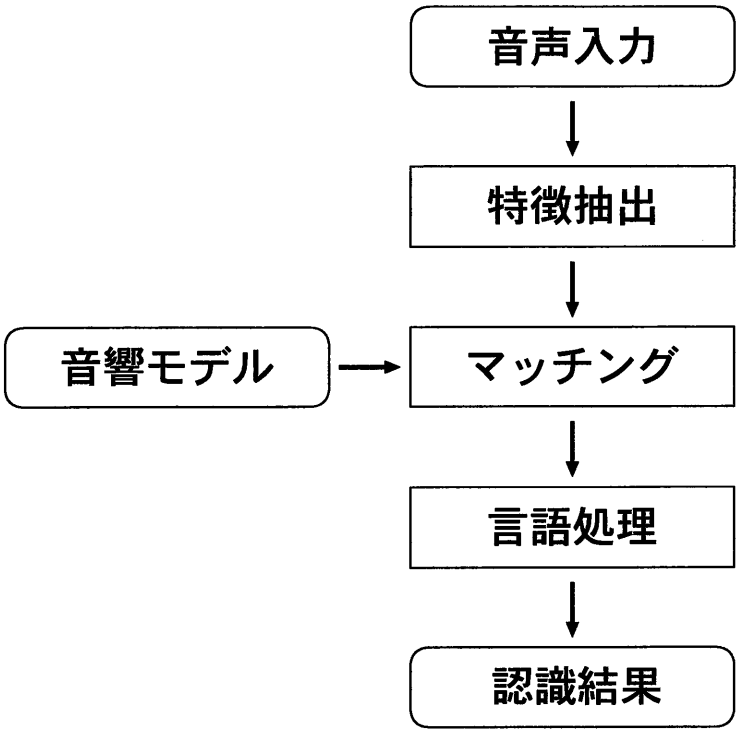

図 1.1: 大語彙連続音声認識システムの概要

日本語の音素は大別すると母音と子音から成り、30 ～ 40 種類程度に定義することが一般的である。音素は単語などに比べ種類が少ないので、認識のために用意するモデルの数が少なくてすみ、精度のよいモデルの学習が比較的容易である。加えて、全ての単語や文は音素系列で表記できるため、多くの大語彙連続音声認識システムでは音素に基づく音響モデルが用いられている。

ただし、音素毎に作成した音響モデルには次のような問題がある。例えば同じ / a / という音素でも、先行音素が / k / であり後続音素が / u / である/ k − a + u / という音素と、先行音素が / o / であり後続音素が / u / である / o − a + u / という音素とでは、先行音素の違いによる音響パターンの変動があることが知られている。このような前後のコンテキストの違いによる音響パターンの変動を調音結合という。音素モデルを用いて認識を行った場合には、調音結合による影響が認識性能の低下の一因になることが知られている。この問題に対処するため、前後の音素の違いを考慮したコンテキスト依存の音素モデルの導入が試みられている。コンテキスト依存の音素モデルについては次章で後述する。

1.1.3 新たな認識単位に基づく音響モデル

一方で、音素という枠組みにとらわれずに、新たな認識単位を獲得して音響モデルを作成しようとする研究が近年行われている。音素という単位は人間が先見的に決定した音声学に基づく単位であり、必ずしも計算機による音声認識処理に適した単位である保証はない。また、前述したように音素に基づく音響モデルには、調音結合による問題を考慮する必要があった。音声データをセグメンテーションして何らかの単位を得る以上、調音結合による影響は必ず生じる問題であるが、音声データの中には、調音結合の影響が強く現われる部分とそうでない部分があると考えられる。このことを考慮して決定した新たな単位を導入することにより、調音結合による悪影響を軽減できる可能性があると考えられる。

以上の理由から、より頑健で高精度な音響モデルを構築するために、音素に代わる新たな認識単位を学習データから獲得しようとする研究が数多く行われている [4]–[10]。これらは大きく分けると、新たな認識単位として、音素との対応がない単位を獲得する方法と、いくつかの音素のまとまりを単位とするような音素との対応のある単位を獲得する方法の2種類に大別される。以下、その2種類の方法について代表的な研究を取り上げ簡単に説明し、それぞれの特徴について述べる。

- 音素との対応がない単位を獲得する方法

  “エルゴディック HMM に基づく音声の自動獲得単位を用いた音声認識” (中川ら、1997) [6]

  獲得する単位は音素と同程度の長さを持つ抽象化単位である。単位の時間長と単位数の条件を与えることで、学習サンプル (孤立単語発声) から単位を獲得する。認識の際に用いる単語辞書は、「同じ単語に対する複数の発声が同じ単位系列で表現される」という制約に基づき、獲得した単位を連結することにより作成する。この手法は音素という枠組みにとらわれずに認識単位を獲得できるという柔軟性がある。しかし、この方法によって獲得した認識単位を用いる場合には、単語を表現する単位系列を得るための計算が必要であり、単語辞書の作成にコストがかかる。さらに、学習サンプルに含まれない単語については単語辞書作成のための計算を行うことができないため、辞書に登録することができない。そのため、大語彙連続音声認識システムで使用するには適さない。

- 音素のまとまりを新たな認識単位とする方法

  “連続音声認識における可変長音声単位の構成法” (新井ら、1999) [10]

  音素認識実験での誤り傾向と、学習サンプル中での2連続ならびに3連続音素の出現頻度を基準として認識単位を決定する手法。この基準により選定した単位として、基本音素 40 種類に、2 音素連結 20 種類、3 音素連結 10 種類を加えた合計 70 種類

のものと、2 音素連結 30 種類、3 音素連結 20 種類を加えた合計 90 種類のものを提案している。認識単位と音素との対応がとれているので、大語彙連続音声認識システムでも比較的容易に使用することができる。しかしこの方法では、単位の長さ (最大で 3 音素の連続としている) やモデルの形状は人手であらかじめ与えておく必要がある。モデルの形状が自動決定できないため、パラメータの共有がないモデルとなっている。そのため、信頼性のあるモデルを得るためには多量の学習サンプルが必要である。

1.2 研究の目的

利用者が快適に使用できる大語彙連続音声認識システムの実現のためには、高精度な音響モデルが必要不可欠である。本研究では、新たな認識単位の自動獲得法を提案し、それに基づく高精度な音響モデルを構築することを目的とする。従来法には、音素との対応がある認識単位を獲得する方法と、対応のない単位を獲得する方法があった。本研究では、大語彙連続音声認識システムを用いたモデルの性能評価を念頭に置いているため、獲得する認識単位は音素との対応がとれたものとする。

ここで従来法の問題点について、もう一度述べる。従来の認識単位獲得手法の多くは、認識単位の長さ、種類、モデルの形状などをあらかじめ与えておく必要がある。そのため、モデルの形状が自動決定できず、適切なパラメータ共有を行うことが困難であるという問題点があった。本研究で提案するアルゴリズムでは、認識時と同じ尺度である学習サンプルの尤度最大という基準の下で、パラメータ共有も含めたモデルの形状、認識単位の長さ、種類を自動決定することができる。提案アルゴリズムでは、モデルの生成と同時にパラメータ共有を行うことができるため、相対的に学習サンプルが増加し、頑健なモデルを生成することができる。

1.3 本論文の構成

本論文の構成は次の通りである。

第1章 序論

研究の背景、目的を述べる。

第2章 音響セグメントモデル

学習サンプルの尤度最大という基準の下で、新たな認識単位を決定し、それに基づくモデルを自動生成するアルゴリズムを提案する。そして提案するアルゴリズムによりモデルを生成する実験を行い、得られた認識単位とモデルの特徴について述べる。

第3章 音声認識実験による ASM の評価

音響モデルの一般的な評価方法である、音素区切り既知の音素認識実験と、音素区切り未知の連続音声認識実験の 2 種類の実験を行い、提案したモデルの性能を評価する。

第4章 ASM を用いた大語彙連続音声認識システム

提案したモデルを大語彙の連続音声認識システムに組み込む方法を提案し、大語彙連続音声認識システムによるモデルの性能評価を行う。

第5章 結論

本研究の成果、今後の課題について述べる。

第2章

音響セグメントモデル

2.1 はじめに

序論では、新たに決定した認識単位を導入することで、音声認識の精度を改善できる可能性があることについて述べた。そこで本章では、モデルに対する学習サンプル全体の尤度が最大になるように学習サンプルを分割することで、学習サンプルから新たな認識単位を自動獲得するアルゴリズムを提案する。本研究で提案する新たな認識単位は、いくつかの音素がまとまったものであり音響セグメントと呼ぶこととする。音響セグメントに基づくモデルを音響セグメントモデル (Acoustic Segment Model: ASM) と呼ぶ。

本章では、まず従来の一般的な音声認識で用いられている認識単位について述べ、その後、提案する認識単位である音響セグメントと、それに基づくモデルである ASM について説明する。そして実際に提案するアルゴリズムを用いて ASM を生成する実験を行い、獲得された音響セグメントと ASM について分析し、その特徴を述べる。

2.2 音声認識の認識単位

語彙数が大きくない場合の音声認識システムでは、認識単位として単語を使用し単語毎に音響モデルを作成する方法が採用されていた。しかし語彙数が多くなった場合、単語モデルの種類が増加し各モデルに与えられる学習サンプルを用意することが困難になるため信頼性の高いモデルを学習することができなくなる。また、大語彙化に伴い個々の単語の音響パターンは必然的に重複する。単語毎に音響モデルを作成することは、個々の単語の要素となっている音響パターンを独立に扱うということになる。よって、異種の単語間に類似した音響パターンがある場合でもそれらの類似点を考慮しないために、過度に冗長なモデルになる。そのため、大語彙のシステムでは、これよりも効率のよい認識単位が必要

表 2.1: 本研究で使用する音素一覧

a	o	u	i	e	j	w	m
n	ng	b	d	g	r	z	dz
h	s	t	c	ts	p	k	q
Q	N	y	a:	o:	u:	i:	e:

とされる。

大語彙連続音声認識システムにおいては音響モデルの認識単位は以下を使用するのが一般的である。

- 音素

  大別すると母音と子音から成る。日本語の場合、ほとんどの音素はローマ字表記したときのアルファベット 1 文字と対応している。

- 音節

  単独で発話の単位となりうる最小の単位。例えば日本語の “頭”(atama) という単語は (a−ta−ma) のように 3 つに区切って発音することができるが、それ以上短く切って発音することはできない。このように、それ自身の中には切れ目がなく、その前後に切れ目の認められる部分を音節という。

音声学的な見地より定義された単位の代表的なものとして上に挙げた音素と音節の 2 種類が存在する。その他にも擬似音素単位、擬似音節単位、擬似半音節単位、音素対などがある [11]。音声認識の分野では、音素を認識単位とする場合が多い。音素は音節に比べて種類が少ないため、音響モデルを作成する際に学習処理の労力が小さくてすむためである。次節では、音声認識の分野で用いられる一般的な音素モデルについて述べる。

2.3 音素モデル

日本語の音素の定義は複数存在するが、本研究では、表 2.1 に示す音素を用いる。/ a: / ～ / e: / は長母音を示す記号である。本研究で使用する音声データである ATR 連続音声データベースの文は、表 2.1 に示した 32 音素の他に / i / と / u / の無声化母音、/ i / の二重母音、/ f /、/ w /、/ t /、/ d / の外来語音の 7 音素を加えた合計 39 音素を用いてラベリングされている。

認識を行うためには、音素毎に音響モデルをあらかじめ作成しておく必要がある。音響モデルとしては次に述べるような確率モデルがよく用いられる。

表 2.2: HMM の定義

HMM $\boldsymbol{M} = (\boldsymbol{S}, \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}, \boldsymbol{F})$

$\boldsymbol{S}$: 状態の有限集合; $\boldsymbol{S} = \{s_i\}$

$\boldsymbol{Y}$: 出力シンボルの集合

$\boldsymbol{A}$: 状態遷移確率の集合; $\boldsymbol{A} = \{a_{ij}\}$
a_{ij} は状態 s_i から状態 s_j への遷移確率. $\sum_j a_{ij} = 1$

$\boldsymbol{B}$: 出力確率の集合; $\boldsymbol{B} = \{b_{ij}(k)\}$
$b_{ij}(k)$は状態 s_i から状態 s_j への遷移の際に
シンボル k を出力する確率

$\boldsymbol{\pi}$: 初期状態確率の集合; $\boldsymbol{\pi} = \{\pi_i\}$
π_i は初期状態が s_i である確率. $\sum_j \pi_j = 1$

$\boldsymbol{F}$: 最終状態の集合

2.3.1 HMM

音素に基づく音響モデルは、統計的な確率モデルである隠れマルコフモデル (Hidden Markov Model: HMM) を用いるのが現在の主流である。HMM とは、いくつかの状態とそれらを結ぶアークからなり、状態遷移のたびに 1 個のシンボルを出力する有限状態オートマトンの一種である。出力シンボル系列は観測できるが、状態遷移系列は直接観測できないことから、隠れマルコフモデルと呼ばれている。HMM は以下のような利点があるため、音響モデルとして広く用いられている。

- 個人差や発声の強さ、速さなどによる音声パターンの変動を統計的に処理することが可能
- 確率モデルであるため理論的展開が容易
- 効率のよいパラメータ推定アルゴリズムが存在

HMM の定義を 表2.2 に示す。表に示した 6 個の組 $\boldsymbol{M} = (\boldsymbol{S}, \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}, \boldsymbol{F})$ によって HMM は表わされる。HMM を音声認識に用いる場合には、以下の 3 つの問題が重要である。

1. 音声データを表わす特徴量ベクトル系列 $\boldsymbol{O} = (o_1, o_2, \ldots, o_T)$ をHMM $\boldsymbol{M}$ が出力する確率 $P(\boldsymbol{O}|\boldsymbol{M})$ の計算

   $P(\boldsymbol{O}|\boldsymbol{M})$ を計算することは、音声データがモデルにどの程度適合するかを知るための問題としてとらえることができる。この確率 $P(\boldsymbol{O}|\boldsymbol{M})$ をモデルに対する音声データの尤度と呼ぶこともある。認識時には、各モデルに対して $P(\boldsymbol{O}|\boldsymbol{M})$ が最大になる $\boldsymbol{M}$ を決定する。各モデルに対する尤度を動的計画法を用いて効率的に計算することのできるアルゴリズムにForward アルゴリズム [12] がある。

2. 音声データ $\boldsymbol{O}$ に対応するHMM $\boldsymbol{M}$ の最適状態遷移系列の決定

   HMM は、ある与えられた音声データに対して対応する状態遷移系列が一意に決定しないため、音声データとモデルの状態とを対応づけることができない。音声認識の分野では、HMM を連結して連続音声データを扱う場合などにおいて、音声データのベクトル系列とHMM の状態との対応づけが必要となることがある。そこで対応をとるために、音声データに対するモデルの最適状態遷移系列を求める。最適状態遷移系列を求めるアルゴリズムの代表的なものとしてViterbi アルゴリズム [12] がある。

3. $P(\boldsymbol{O}|\boldsymbol{M})$ を最大にするためのHMM $\boldsymbol{M}$ のパラメータ $\boldsymbol{A}$ 、$\boldsymbol{B}$ 、$\boldsymbol{\pi}$ の推定

   このパラメータ推定は、音声データ $\boldsymbol{O}$ を最もよく表わすようなHMM $\boldsymbol{M}$ を生成することを意味する。これをモデルの学習と呼ぶ。学習のための効率のよいアルゴリズムの一つとしてBaum-Welch アルゴリズム [12] が知られている。このアルゴリズムはパラメータの反復計算をおこなうことで、 $P(\boldsymbol{O}|\boldsymbol{M})$ を局所的に最大にするアルゴリズムである。ただし、Baum-Welch アルゴリズムではパラメータが最適値に収束する保証がないため、学習の際にモデルによい初期値を与えることが重要である。

代表的なHMM の構造としては図2.1に示すようなergodic HMM や図2.2に示すようなleft-to-right HMM がある。ergodic HMM はすべての状態間での遷移を許す構造になっており、自由度の高いモデルであると言える。音声認識で用いられるHMM としては、図2.2に示すleft-to-right HMM が一般的である。left-to-right HMM は前の状態に逆戻りするような遷移がないため、音声の時間的な変化を表現するのに適している。

2.3.2 HMnet

音声のパターンをよく表現できるモデルであるHMM であるが、単に音素毎に作成したHMM を使用して認識を行った場合には、調音結合の影響による認識性能の低下が起こるという問題がある。この問題に対処するため、前後の音素の違いを考慮したコンテキスト依存の音素モデルの導入が試みられている。コンテキスト依存モデルには、いくつかの種

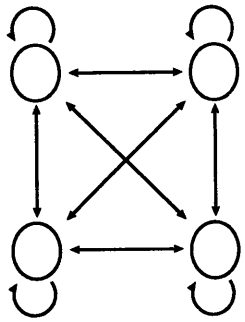

図 2.1: ergodic HMM

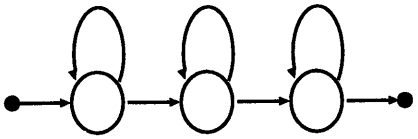

図 2.2: left-to-right HMM

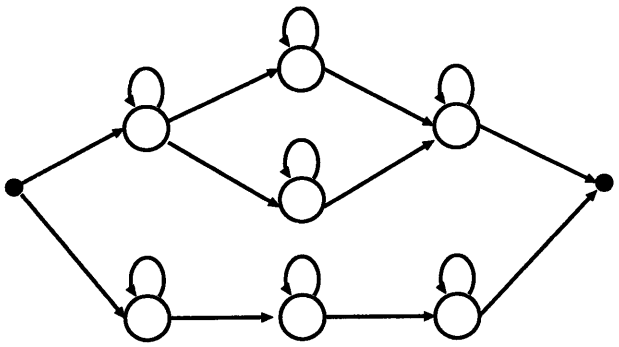

図 2.3: HMnet

類が存在する。代表的な例としては、Triphone モデル [13] や音素決定木に基づくコンテキストモデル [14] などがある。また、モデルの状態数を情報量基準を用いて自動で決定しコンテキストモデルを生成するという研究もある [15]。

状態共有関係をネットワークとして表現した高精度なコンテキスト依存モデルの一つに隠れマルコフ網 (Hidden Markov Network: HMnet) がある。本研究では、このモデルに着目する。その形状を 図2.3 に示す。図2.3 に示すように HMnet は left-to-right HMM と ergodic HMM の中間の形状をしている。HMnet は状態共有関係を有する left-to-right HMM の集合とみなすことができる。以下、HMnet を構成している left-to-right HMM をパスと呼ぶことにする。各パスがそれぞれ、あるコンテキストを表現した音素モデルになっている。コンテキスト依存モデルは単なる音素モデルに比べてモデルの数が大幅に増加するため、使用できる学習サンプルに制限がある場合のモデルのパラメータの学習が問題になる。信頼性が高く頑健なモデルを学習するためには、適切なパラメータの共有が必要となる。真に最適な HMnet を生成するためには、コンテキストの分類や状態共有構造などに関する膨大な組合せ問題を解く必要があり、それを実現するのは困難である。そこで、その近似的解を求め、HMnet を生成するアルゴリズムが提案されている。

HMnet 生成アルゴリズム

HMnet を生成するアルゴリズムには、いくつかの種類が存在する。代表的なものとして逐次状態分割法 (Successive State Splitting: SSS) [16] [17] や、その改良手法である SSS-free [18] がある。その他の例としては、最尤逐次状態分割法 (ML-SSS) [19] や音素決定木に基づいて HMnet を生成するアルゴリズム [20] [21] などがある。

図 2.4 に SSS 及び SSS-free アルゴリズムの流れを示す。どちらの方法も状態を逐次的に増加させていくことで HMnet を自動生成していくアルゴリズムである。これらのアルゴリズムは、時間方向も含めてモデルの構造を自動決定することが可能である。また、これらのアルゴリズムは学習に時間を要するという欠点があるため、高速化の手法が提案されている [18]。SSS では環境要因 (先行音素や後続音素など) が必要であるのに対して、SSS-free では環境要因を必要としない点が両者の大きな相異点である。SSS-free は音響的類似性に基づいて、高精度な音響モデルを生成するアルゴリズムであることが報告されている。以下、高速化 SSS-free アルゴリズムについて説明する。

高速化 SSS-free アルゴリズム

1. 初期モデルの学習
   初期モデルとして、1 状態で出力確率分布が単一ガウス分布 (対角共分散行列) を持つ HMM を用意し、全ての学習サンプルを使って学習する。
2. 分割する状態の選択
   全ての状態の中で分布が最も広がった状態を選び、分割すべき状態とする。分布の広がりを判断する際には、単一ガウス分布の分散の値と推定に用いたサンプル数を乗じたものを基準とする。つまり、分散が大きく、かつ学習サンプルが多く割り当てられている状態を分割することになる。各状態毎に以下の式で示す d_i を計算し、それが最大値となる状態 i を分割する状態として選択する。

$$d_i = n_i \times \sum_{k}^{K} \frac{\sigma_{ik}^2}{\sigma_{Tk}^2}$$

K : パラメータ次数
σ_{ik}^2 : 状態 i の出力確率分布の分散
n_i : 状態 i の推定に用いたサンプル数
σ_{Tk}^2 : 全サンプルの分散 (正規化係数)

3. 分割方向の決定
   選択された状態を 2 つに分割し、2 つの新しい状態を生成する。このとき、新しい状態の出力確率分布は以下のようにして求める。

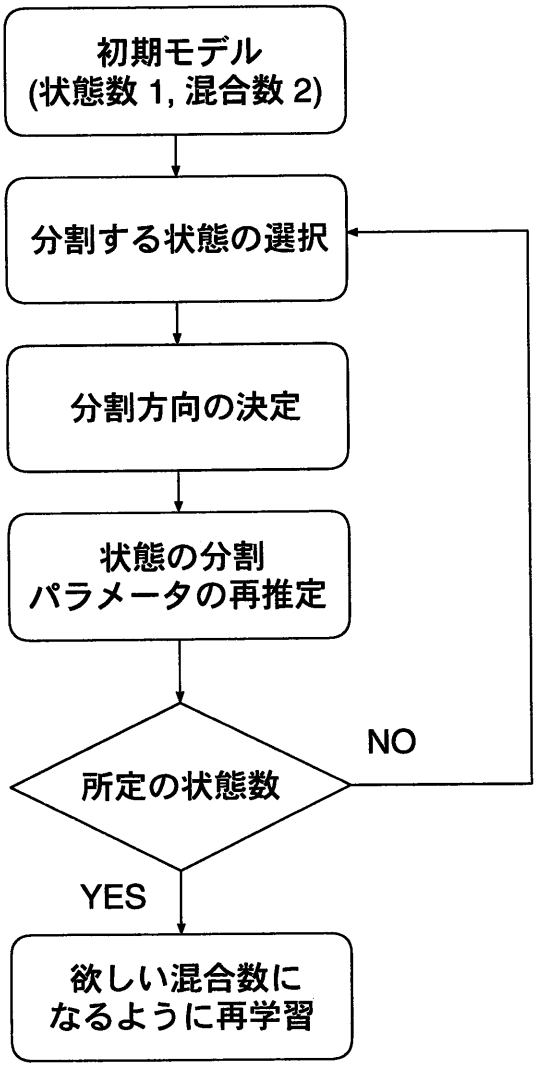

図 2.4: SSS 及び SSS-free アルゴリズム

(a) 選択された状態を通る全ての学習サンプルについて Viterbi アルゴリズムを使ってこの状態が出力するサンプルの部分系列を切り出してくる。

(b) (a) で切り出された全ての学習サンプルの部分系列を用いて、2 混合ガウス分布を出力確率分布に持つ 1 状態 の HMM を学習する。

(c) 得られた 2 つのガウス分布をそれぞれ新しい状態に割り当てる。

このようにして 2 つの新しい状態の出力確率分布を求めた後、分割方向の決定を行う。分割の方向には時間方向 (直列方向) とコンテキスト方向 (並列方向) の 2 通りが考えられる。新しい状態を、時間方向に連結した場合の学習サンプルに対する尤度 P_t と、コンテキスト方向に連結した場合の尤度 P_c を計算し、尤度の高いほうを採用することで、分割方向を決定する。P_t と P_c は、それぞれ以下のようにして計算される。

- 時間方向への分割

  時間方向へ分割する時は、どちらの状態を先に置くかで 2 通りの可能性がある。そこで、2 つの可能性についてそれぞれ尤度を計算し、その高いほうを P_t とする。

- コンテキスト方向への分割

  コンテキスト方向への分割は、パスが 2 つに別れるためにそれぞれの学習サンプルがどちらの状態を通るかを決定する必要がある。各学習サンプル 1 つ 1 つについて、尤度の高いほうの状態を通るように決定する。

$$P_c = \sum_{y_j \in Y_m} \max(P_m(y_j), P_M(y_j))$$

Y_m : 分割前の状態 m を通る学習サンプルの集合

y_j : 状態 m を通る j 番目の学習サンプル

$P_m(y_j)$: y_jを状態 m に割り当てた時の尤度

$P_M(y_j)$: y_jを分割後の状態 M に割り当てた時の尤度

4. 状態の分割及びパラメータの再推定

   前までのステップで決定した分割方向に従って、選択された状態を分割する。分割終了後、最適なパラメータを求めるために HMnet 全体を再学習する。その後、所定の状態数になるまで、2 と 3 を繰り返す。所定の状態数になればアルゴリズムを終了する。アルゴリズム終了後のモデルの出力確率分布は単一ガウス分布になっているので、混合数を増やしたければ再学習を行い所望のモデルを得る。

2.4 音響セグメントモデル

2.4.1 音響セグメント

本研究で提案する音響セグメントという単位は、いくつかの音素がまとまったものと定義する。そのため、音響セグメントは音素以上の長さを持った単位になる。

日本語の連続発声された文データを学習サンプルとして用い、この中に出現する音素系列のうちのいくつかを音響セグメントとして決定する。具体的な例をいくつか挙げると、音響セグメントは / a k /、/ o i /、/ a Q t /、/ o d a / などのような形になる。音響セグメントは、モデルに対する尤度が最も高くなるように、学習サンプルである文データを分割することで決定する。つまり、音響セグメントは、そのモデルを用いたときに学習サンプルに対する認識率が最大になるように決定した単位と言うこともできる。

このような長い単位を用いることの効果としては以下のようなことが考えられる。発声時間の短い子音のような音素は、単独でモデル化するよりも、隣接する音素とまとめて音

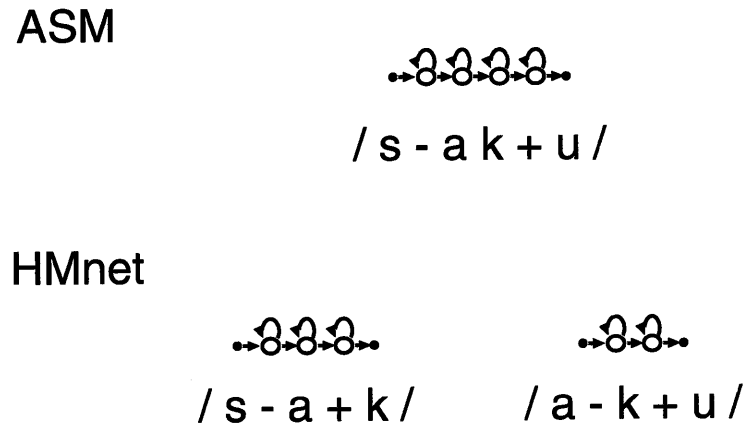

図 2.5: ASM と HMnet の比較

響セグメントのような形でモデル化すれば、より安定したモデルになることが期待できる。以下、音響セグメントに基づくモデルを音響セグメントモデル (Acoustic Segment Model: ASM) と呼ぶ。

2.4.2 ASM の利点

ASM は音素との対応がとれているため、従来の音素ベースの音声認識実験での評価を比較的容易に行うことが可能である。また、ASM を連結することにより単語や文などのモデルを作ることができるため、従来の音素モデルに基づいた認識システムに ASM を容易に組み込むことができる。

また、ASM には従来の音素モデルよりも長い距離のコンテキストを考慮できるという利点がある。例として / a k / という音素系列をモデル化することを考える。ここで音素系列 / a k / の先行音素は / s / で後続音素は / u / であるとする。図 2.5 に ASM と音素 HMnet をコンテキストモデルとして用いたときについて示す。音素 HMnet では前後の 1 音素をコンテキストとして考慮し / s − a + k / と / a − k + u / という 2 つの音素モデルの連結で / a k / を記述する。それに対して ASM では / a k / を 1 つのモデルとして表現する。コンテキストモデルとしてみた場合、音素 HMnet では、/ a / と / k / 共に前後 1 つの音素のみをコンテキストとして考慮している。それに対して ASM では、 / a / と / k / をまとめてモデル化しているため、/ a / はこの場合の音素 HMnet では考慮できない 2 つ後ろの音素 / u / を考慮したものになる。同様に / k / は音素 HMnet では考慮できない 2 つ前の音素 / s / を考慮したものになる。このように 2 音素以上の長さを持つ音響セグメントを用いると、長い距離のコンテキストを扱うことができる。

2.5 ASM 生成アルゴリズム

ASM と音響セグメントを生成するアルゴリズムの概要について述べる。アルゴリズムは大きく分けて、ASM の状態分割を行いモデルを詳細化していくステップと認識単位である音響セグメントを決定するステップの 2 つのステップから成る。以下、提案アルゴリズムを単に ASM 生成アルゴリズムと呼ぶ。

- ASM の詳細化

  詳細化とは、ASM の状態の分割を行い、モデルの状態数を増やすことを意味する。ASM の詳細化には、HMnet 生成アルゴリズムの 1 つである SSS-free を用いる。SSS-free は音素を単位として動作するアルゴリズムであるが、ここでは音響セグメントを単位として SSS-free を動作させる。そのため ASM のモデルの形状は HMnet と同一なものとなる。ASM は音響セグメントを単位として生成した HMnet であると言うこともできる。それゆえ ASM は HMnet と同じく状態共有関係を有する left-to-right HMM (パス) の集合とみなすことができる。SSS-free では認識時の尺度と同じ尤度最大を基準としてモデルが構成され、その結果、学習サンプルの通るパスが決定される。このため、音響的に似た音響セグメントは同一のパスに割り当てられることが期待できる。

- 音響セグメントの決定

  音響セグメントは、学習サンプルである連続発声された文データを、いくつかに分割することにより得られた音素列である。この際には、ASM に対する尤度が最大となるように学習サンプルのセグメンテーションが行われる。具体的には、学習サンプルを最もよく表現するような ASM のパス系列を求め、各パスに対応する音素列を音響セグメントとすることで学習サンプルの分割を行う。

提案するアルゴリズムは、この 2 つのステップを繰り返し行うことにより ASM を生成する。ASM の生成過程に伴って、音響セグメントはその時の ASM に対する尤度が最大になるように動的に変化する。

以下に、ASM 生成アルゴリズムについて更に詳しく述べる。

1. 初期モデルの学習

   初期モデルの形状を 図2.6 に示す。初期モデルは状態数 2 の left-to-right HMM であり、出力確率分布は単一ガウス分布 (対角共分散行列) である。なお、このモデルは、学習サンプルに付与されている音素区切りの時刻でのみ最終状態から開始状態への遷移を許すことにする。この特別な遷移をバックループと呼ぶ。バックループの特別な

条件は、獲得する認識単位である音響セグメントと音素との対応をとるために必要なものである。
連続発声された文データを学習サンプルとしてバックループを考慮した初期モデルの学習を行う。

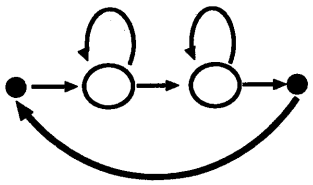

図 2.6: 初期モデル

2. 音響セグメントの獲得

図2.7に、音響セグメントの獲得の例を示す。学習終了後の初期モデルに対し、Viterbiアルゴリズムを用いて学習サンプルの最適状態遷移系列を求める。その結果、図2.7に示すように学習サンプルと初期モデルの状態遷移系列が対応づけられる。そして、開始状態から最終状態までの区間に対応する音素列を 1 つの音響セグメントとして学習サンプルの分割を行う。図2.7の破線をまたぐ遷移は 図2.6に示した初期モデルのバックループに相当する。
この例では、学習サンプルは / a r /、/ a j u r /、/ u g e / という 3 つの音響セグメントに分割される。これは、この学習サンプルを最もよく表現するように初期モデルとの対応をとったことを意味する。

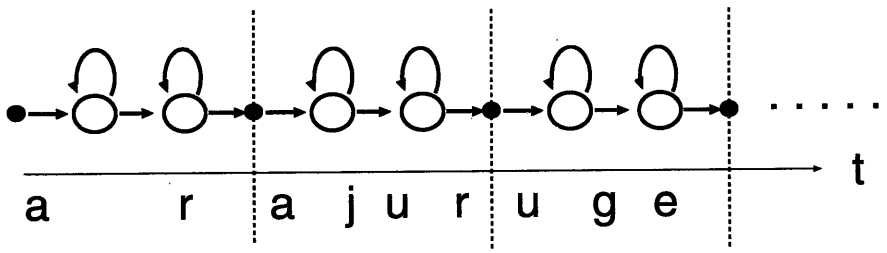

図 2.7: 音響セグメントの獲得

3. ASM の詳細化

ASM の状態の分割を行い状態数を増加させることで、ASM の詳細化を行う。前のステップで獲得した各音響セグメントを学習サンプルとみなすことで SSS-free をそのまま適用して数回の状態分割を行う。この際はバックループは考慮しない。SSS-free は音素を単位とした学習サンプルを用いて動作するアルゴリズムであるが、ここで提案している単位である音響セグメントを単位とした場合でも、特別な変更なしに適用することができる。

図 2.8 に状態の分割についての図を示す。分割した状態を配置する方向には、直列な方向と並列な方向の 2 通りがあるが、学習サンプル全体の尤度が最大となる方向を選択する。ここでの尤度の計算は音響セグメントを単位として行われる。

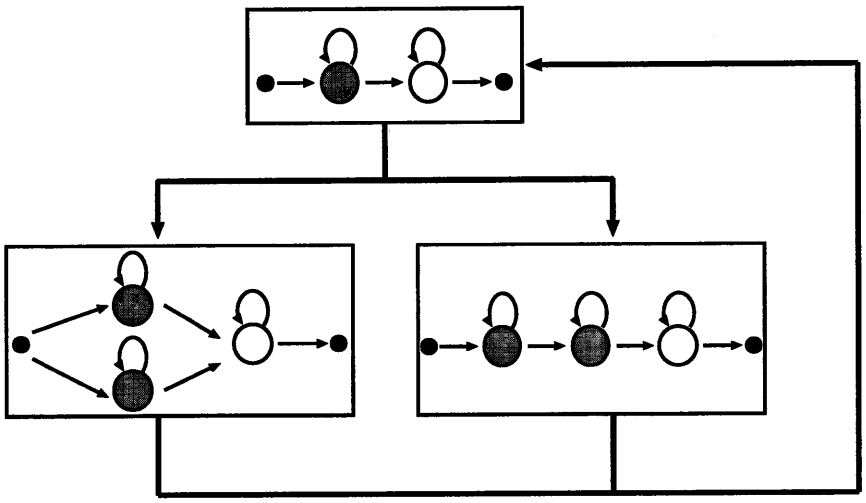

図 2.8: ASM の詳細化

4. 音響セグメントの再構成

前のステップで詳細化した ASM に対する学習サンプルの尤度が最大になるように音響セグメントの再構成を行う。図 2.9 に音響セグメントの再構成の例を示す。バックループを付加した ASM に対して Viterbi アルゴリズムを用いて学習サンプルの最適状態遷移系列を求める。そして 2. 音響セグメントの獲得の場合と同様にして、ASM の開始状態から最終状態までの区間に対応する音素列を 1 つの音響セグメントとすることで、学習サンプルを分割する。この例では、学習サンプルは / a /、/ r a j /、/ u r u /、/ g e / の 4 つの音響セグメントに分割され、各音響セグメントは最大の尤度を示すパスにそれぞれ割り当てられる。ASM が詳細化されたため、2. 音響セグメントの獲得の例とは異なる音響セグメントが得られている。このように、音響セグメントは、ASM の詳細化に伴って、そのときのモデルに対する尤度が最大になるように動的に変化する。

ASM の各パスに割り当てられる音響セグメントは尤度のみを基準として決定されるため、1 つのパスに複数の種類の音響セグメントが混在する場合がある。図 2.9 の例では、異なる種類の音響セグメント / a / と / r a j / が同一のパスに割り当てられている。

所定の状態数になるまで 3. ASM の詳細化と 4. 音響セグメントの再構成を繰り返す。

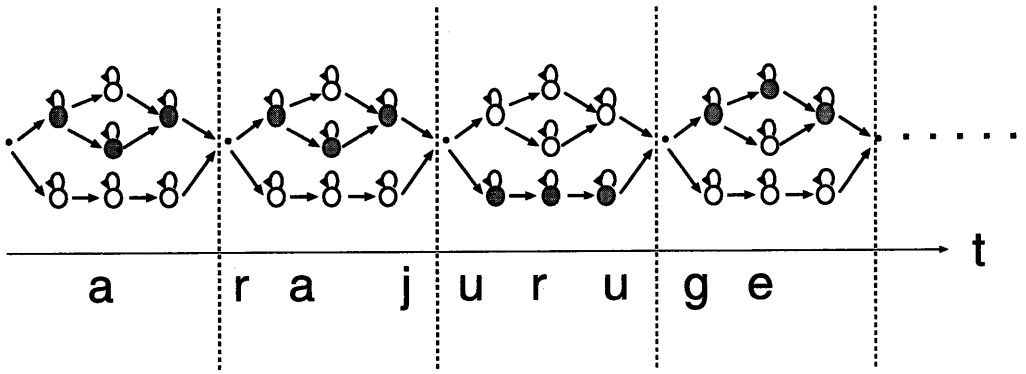

図 2.9: 音響セグメントの再構成

表 2.3: 実験条件

音声データ	ATR 連続音声データベース
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
特徴量	logpow, cep(16), Δlogpow,
	Δcep(16) からなる 34 次ベクトル
話者	男性 4 名, 女性 2 名
学習サンプル	400 文

2.6 音響セグメントと ASM の生成

提案した ASM 生成アルゴリズムにより、音響セグメントと ASM を作成する実験を行う。そして、獲得した新たな認識単位である音響セグメントと、それに基づくモデルである ASM の特徴について述べる。具体的には、学習サンプルから得られた音響セグメントの種類や、ASM のモデル形状について調査する。

2.6.1 実験条件

特定話者 6 名について ASM を生成した。実験条件は 表 2.3 に示す。なお、音響セグメントは状態の分割を 10 回行うごとに再構成した。アルゴリズムの終了条件である ASM の状態数は 500 とし、全話者について、その状態数になるまで分割を行った。

2.6.2 獲得した音響セグメント

表 2.4 に男性 1 の学習サンプルより得られた 261 種類の音響セグメントの一部を示す。最も長い音響セグメントは 4 音素から成る音響セグメント (/ k i s i /) であった。実験の結果、音素と全く同じものである 1 音素から成る音響セグメントは、全ての音素について得られており、ASM は音素モデルを包含したものになった。最も種類が多かったのは 2 音素から成る音響セグメントで全体の約 66 % を占めていた。

次に、話者による音響セグメントの種類の違いについて調査した。 表 2.5 にそれぞれの話者についての音響セグメントの種類を示す。1 音素から成る音響セグメントは、2 つの外来語音素と 1 つの無声化母音を除いた全ての音素について、全話者から得られていた。このことより、音響セグメントは 39 種類の音素をほとんど全て包含したものになってい

表 2.4: 音響セグメントの例 (男性 1)

/ Q t /	/ s i /	/ y o: /	/ e N /
/ N n /	/ s u /	/ e i /	/ r e /
/ y o /	/ N d /	/ h i /	/ y u /
/ ts u /	/ y a /	/ Q p /	/ a o /
/ N ng /	/ o w /	/ a i /	/ i N /
/ r o /	/ c i /	/ o o /	/ a e /
/ u o /	/ Q k /	/ i j /	/ N m /
/ a w /	/ u t /	/ k i /	/ N b /
/ h u /	/ u s /	/ u w /	/ r i /
/ N r /	/ u j /	/ o: o /	/ r u /
/ ts u s /	/ h i s /	/ s i s /	/ y o: o /
/ y o: w /	/ a w a /	/ u Q t /	/ h u s /

表 2.5: 音響セグメントの種類

	1 音素	2 音素	3 音素	4 音素	5 音素	6 音素	7 音素	合計
男性 1	39	173	48	1	0	0	0	261
男性 2	39	256	195	31	2	0	0	523
男性 3	38	244	184	51	13	0	0	530
男性 4	38	279	231	36	2	0	0	586
女性 1	38	288	345	124	38	2	2	837
女性 2	39	165	51	9	0	0	0	264

ると言える。2 ～ 3 音素程度の音響セグメントが多く得られるという傾向は全ての話者に共通していた。

学習サンプルを分割した結果、得られた音響セグメントの種類が最も多い話者は女性 1 で 837 種類、最も種類が少ない話者は男性 1 で 261 種類であった。このように話者により得られる音響セグメントの種類には違いがみられた。これは話者による音響パターンの違いや発声速度の違いによるものと考えられる。最も長い音響セグメントは女性 1 から得られた 7 音素から成るものであった。ただし、このような長い音響セグメントが全体の中で占める割合は極めて小さい。このことについては後述する。

一方で、全ての話者から共通して得られる音響セグメントもあった。2 音素以上から成る音響セグメントの中で、全話者に共通しているものの一例を 表 2.6 に示す。比較的学習サンプルに頻繁に出現するような音素列が全話者に共通な音響セグメントになっている傾

表 2.6: 全話者に共通な音響セグメントの例

/ Q t /	/ s i /	/ y o: /
/ N n /	/ s u /	/ e i /
/ y o /	/ N d /	/ h i /
/ ts u /	/ y a /	/ Q p /
/ N ng /	/ o w /	/ a i /
/ r o /	/ c i /	/ o o /
/ u o /	/ Q k /	/ i j /
/ ts u s /	/ h i s /	/ y o: w /

表 2.7: 各音響セグメントの出現回数

	1 音素	2 音素	3 音素	4 音素	5 音素	6 音素	7 音素
男性 1	17587	2448	101	1	0	0	0
男性 2	14547	3468	389	32	2	0	0
男性 3	14771	3115	495	58	13	0	0
男性 4	12866	4145	479	45	2	0	0
女性 1	9158	4717	1023	222	42	2	2
女性 2	17870	2284	100	11	0	0	0

向が見られた。全話者 6 名に共通の音響セグメントは 144 種類であった。男性 1 に関して言えば、約 55 % の音響セグメントが全話者共通のものだったということになる。

ここまでは、得られた音響セグメントの種類について述べてきたが、次に、学習サンプルに出現した各音響セグメントの個数について述べる。表 2.7 に話者毎の各音響セグメントの出現回数を示す。これによると、学習サンプルの多くは 1 音素から成る音響セグメントとして分割されていることが分かる。1 音素から成る音響セグメントは音素と等価なものであるが、そのような音響セグメントが最も多く出現していた。2 音素以上から成る音響セグメントの占める割合は全体で約 21 % であった。

2.6.3 ASM の特徴

男性 1 の ASM について、特定の音響セグメントに着目してモデルの特徴について述べる。500 状態まで分割した男性 1 の ASM のパスの総数は 7402 本であった。この節では、この中の音響セグメント / o r / と / u m / が割り当てられているパスのみに注目する。

表 2.8 に音響セグメント / o r / が割り当てられている ASM のパスについて示す。学習

サンプルを分割することにより7個の音響セグメント / o r / が得られ、それらは5種類のパスに割り当てられていた。図2.10に音響セグメント / o r / が割り当てられているパスの状態共有関係を示す。5本のパス間で共有している状態が多く見られ、同じ種類の音響セグメントは比較的類似したパスに割り当てられている様子が確認できる。表2.8をみると / o r / 以外の音響セグメント / u r / が1個割り当てられているパスが1本あるが、それ以外のパスは / o r / のみが割り当てられていた。提案アルゴリズムでは、音響的な類似性に基づいて、音響セグメントが割り当てられるパスが決定される。同一の音素列により記述された7個の音響セグメント / o r / 全てが音響的に似ていて、なおかつ、他の音響セグメントとの独立性が強かったため、このような結果になったと考えられる。この例のような傾向は他の音響セグメントのいくつかについても同様に見られ、同一の音素列からなる音響セグメントはまとまってモデル化されていることが確認できた。

表 2.8: 音響セグメント / o r / が割り当てられているパス

パス (状態番号系列)	/ o r /	/ u r /
317 277 288 90 105	1	0
486 90 105	2	1
397 379 71 254 431	1	0
397 379 71 90 105	2	0
317 379 71 90 105	1	0

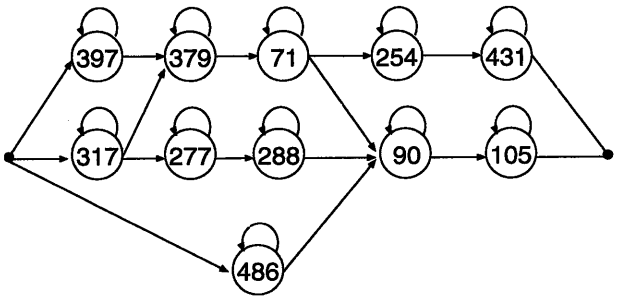

図 2.10: 音響セグメント / o r / に関するパスの状態共有関係

表2.9に音響セグメント / u m / が割り当てられている ASM のパスについて示す。学習サンプルを分割することにより5個の音響セグメント / u m / が得られ、それらは3種類のパスに割り当てられていた。図2.11に音響セグメント / u m / が割り当てられているパスの状態共有関係を示す。3本のパス間で共有している状態がいくつか見られ、図2.10と同様に、同じ種類の音響セグメントは比較的類似したパスに割り当てられている様子が確認できる。しかしこの例で示した3本のパスには 図2.10とは異なり、/ u m / 以外の異種の音響セグメント (/ N m / や / m / など) が数多く混在していた。混在していた音響セグメントは、/ u m / に音響的によく似ていたため、同一のパスに割り当てられたと考えられる。このように ASM では1つのパスに複数の種類の音響セグメントが混在する場合がある。

表 2.9: 音響セグメント / u m / が割り当てられているパス

パス (状態番号系列)	/ u m /	/ u m / 以外
278 14 61	1	63
162 14 295 415	2	3
457 14 295 415	2	5

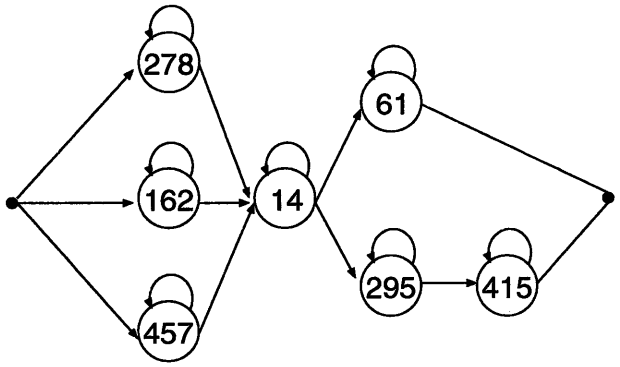

図 2.11: 音響セグメント / u m / に関するパスの状態共有関係

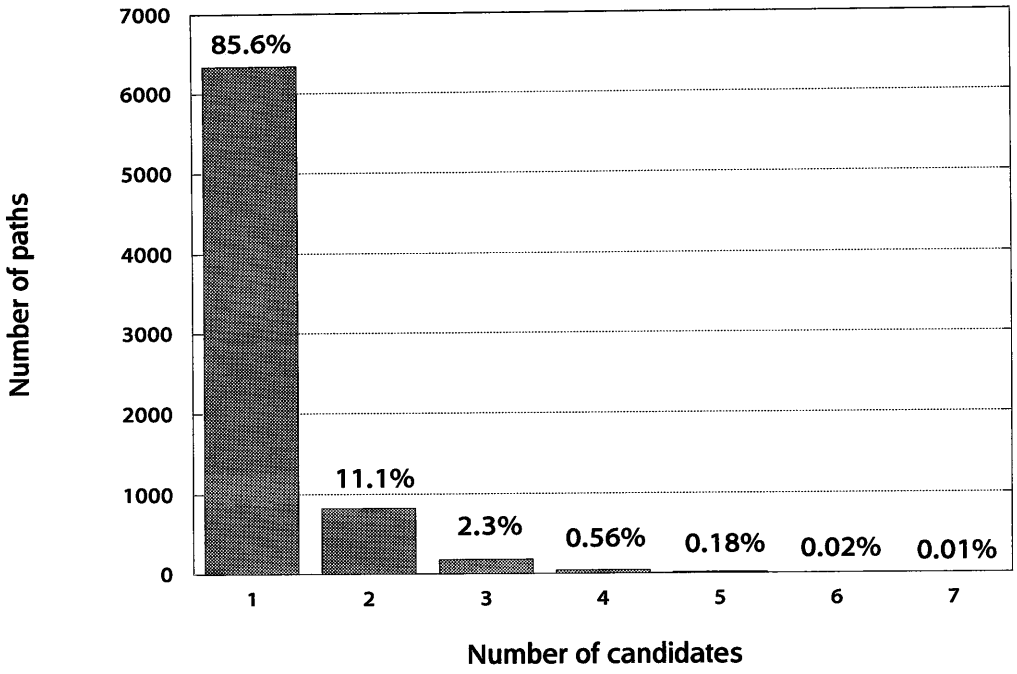

図 2.12: パスに割り当てられている音響セグメントの種類

表 2.9、図 2.11 に示した音響セグメント / u m / が割り当てられたパスは複数の種類の音響セグメントが同一のパスに混在しているため、認識に用いる際に認識結果を一意に定めることができないという問題が生じる。ASM のパスの中で、このように認識結果に曖昧性を持つものがどの程度の割合で出現しているか調査した。

図 2.12 に、男性 1 の ASM の各パスに割り当てられている音響セグメントの種類を示す。最も多いのは複数の種類の音響セグメントが混在することなく 1 種類の音響セグメントのみが割り当てられているパスであり、全体の約 85 % を占めていた。このようなパスは前述したような問題がなく認識に使用することができる。それに対して、2 種類以上の音響セグメントが割り当てられているパスは認識結果を一意に定めることができないが、このようなパスは全体の約 15 % であった。

2.7 まとめ

本章では ASM を生成するアルゴリズムを提案した。提案したアルゴリズムは、学習サンプルの尤度最大という基準の下で、認識単位を自動決定することが可能であった。加えて、各状態間でのパラメータの共有も行うため、相対的に学習サンプルが増え頑健なモデ

ルを学習できる。

提案したアルゴリズムにより ASM を生成する実験を行い、得られた音響セグメント、ASM の特徴について述べた。音響セグメントの種類には話者による相違があったが、全話者に共通して見られる傾向もあった。全話者について不変な音響セグメントが少なからず存在したことがその一例である。

ASM は音響的な類似性に基づいてモデルが生成されるため、同種の音響セグメントは類似したパスに割り当てられており比較的音素との対応があるモデルになった。一方で、ASM が従来の音素モデルと大きく異なるのは、複数の種類の音響セグメントが混在したパスがあるという点である。そのため ASM を用いて音声認識実験を行う場合には、認識結果を一意に決定できない場合があることが分かった。

第3章

音声認識実験による ASM の評価

3.1 はじめに

前章では音響セグメントという認識単位を自動獲得し、それに基づくモデルである ASM を生成するアルゴリズムを提案した。本章では、音声認識の分野での一般的な評価方法である音素認識実験、及び連続音声認識実験により ASM の性能を評価する。ASM はモデルの性質上、複数の認識結果の候補が存在し、認識結果を一意に決めることができないという問題点があった。そのため従来の評価法がそのまま適用できない。そこで本章では ASM の認識結果の曖昧性を考慮した評価を行うことにより、ASM が有する認識性能について議論する。

3.2 ASM を用いた音素認識実験

テストサンプルの音素区切りが既知という条件の下で音素認識実験を行う。認識は音響セグメントの長さとテストサンプルの音素区切りに対応をつけて行うため、音素の挿入や脱落は生じない。この節では ASM と従来の音素モデルである HMnet の音素認識率を比較し、ASM の認識性能を評価する。

3.2.1 実験条件

特定話者 6 名 (男性 1 ～ 4、女性 1 ～ 2) について認識実験を行う。学習のためのサンプルと、テストのためのサンプルは同じ話者の発声によるものである。ASM は前章の 表 2.3 に示す実験条件の下で作成したものを使用する。学習サンプルは 400 文であり、ASM の状態数は 500 であった。認識に使用するテストサンプルは、学習サンプル以外の 103 文を

用いる。

話者によっては ASM が全ての音素モデルを包含していないことがあるため、テストサンプルによっては認識が不可能な音素を含む場合がある。このような事態を回避するために、全音素について、状態数 3 の音素環境非依存 left-to-right HMM を ASM と同時に使用して認識を行った。

3.2.2 平均候補数

前章で述べたように ASM では 1 本のパスに 2 種類以上の音響セグメントが割り当てられている場合があるため、認識結果が一意に定まらないことがある。音素認識実験では、正解音素を含んだ認識結果が 1 つでもあれば認識できたものとして認識率を計算した。

このように ASM の認識率は、曖昧性を残したまま計算するため、どの程度の曖昧性をもって認識が行われたかを示す指標が必要である。そのため、正解を含む認識結果については以下の式で与えられる平均候補数 PP をあわせて計算する。この値は認識結果として得られたパスを通る学習サンプル数で重み付けした 1 音素あたりの平均候補数である。PP が大きければ、それだけ認識結果の曖昧さが大きいことを意味し、PP が 1 に近ければ近いほど曖昧性がなく認識結果が一意に定まっていることを意味する。

$$PP = 2^H, \quad H = -\frac{1}{n}\sum_{i=1}^{n}\log_2 p_i$$

ここで p_i は i 番目の正解を含む音素区間に占める正解の割合、n は正解音素数である。

3.2.3 実験結果及び考察

音素認識実験の結果を 表3.1 に示す。比較のため、ASM と同じ状態数である 500 状態まで分割した音素 HMnet を用いたときの結果もあわせて示す。この実験では、コンテキストによるモデルの制限は行わなかった。ASM は、全ての話者について、平均候補数をそれほど大きくすることなく音素 HMnet よりも高い認識率を示した。最も認識率の向上がみられた話者は男性 3 で約 9 % 高い値を示した。平均で約 3.5 % の音素認識率の向上がみられた。ASM では 2 音素以上の長さの音響セグメントに関するパスを用いた部分で認識率の改善がみられる例があり、長い区間をモデル化することの有効性が確認できた。

認識結果が一意に決定しない ASM のパスを認識に用いた場合について述べる。このようなパスは正解音素を比較的多く含んでおり、ASM の認識率の向上の一因になっていた。ASM では、音響的に似ていて区別するのが困難であるような音素列については一意に認識結果を決めず曖昧性を残したまま評価することになるため、このことが認識率の向上に寄与したと考えられる。この評価は ASM の性能を過大評価したものであるが、認識時の曖昧性を示す指標である平均候補数 PP は平均で 1.146 と小さな値を示した。平均候補数

表 3.1: 音素認識実験結果

話者	音素 HMnet	ASM (*PP*)
男性 1	87.52%	91.08% (1.121)
男性 2	78.21%	81.05% (1.241)
男性 3	75.27%	84.34% (1.178)
男性 4	81.06%	82.40% (1.125)
女性 1	84.49%	85.33% (1.091)
女性 2	86.50%	89.84% (1.125)
平均	82.17%	85.67% (1.146)

の増加は 14.6 % 程度に抑えることができているため、過大評価の影響はあまり大きくないと考える。

3.3 ASM を用いた連続音声認識実験

この節では、連続音声認識実験を行い ASM の性能を評価する。前節の音素認識実験とは異なりテストサンプルの音素区切りは未知という条件の下で連続発声された文の認識を行う。

3.3.1 コンテキストを考慮した認識

男性話者 4 名について、ASM を用いた連続音声認識実験を行う。テストサンプルに対して最尤となる ASM のパス系列を Viterbi アルゴリズムによって求める。この実験では言語モデルは使用せず、以下に述べるような方法で学習サンプルから得られるコンテキスト情報のみを考慮して ASM のパスの接続を制限した。コンテキストとしては音響セグメントの前後の 1 音素までを含むことにした。

図 3.1 に認識時の ASM のパス接続の例を示す。図 3.1 の左側のパスには 2 種類の音響セグメントが割り当てられており、その後半部分に / a + u / 及び / k + u / のコンテキストを持つ。このようなパスは、前半部分に / a − u / または / k − u / のコンテキストを持つ音響セグメントが割り当てられているパスのみとの接続が可能であるものとした。 図 3.1 の右下に示すような共通のコンテキストを持たないパスとの接続は考慮しない。

しかし、このようにしてパスの接続を制限すると学習サンプルに含まれないコンテキストを持つテストサンプルについては正しく認識することができないという問題が生じる。そこで学習サンプルに出現しないコンテキストについても認識可能にするため、全てのモ

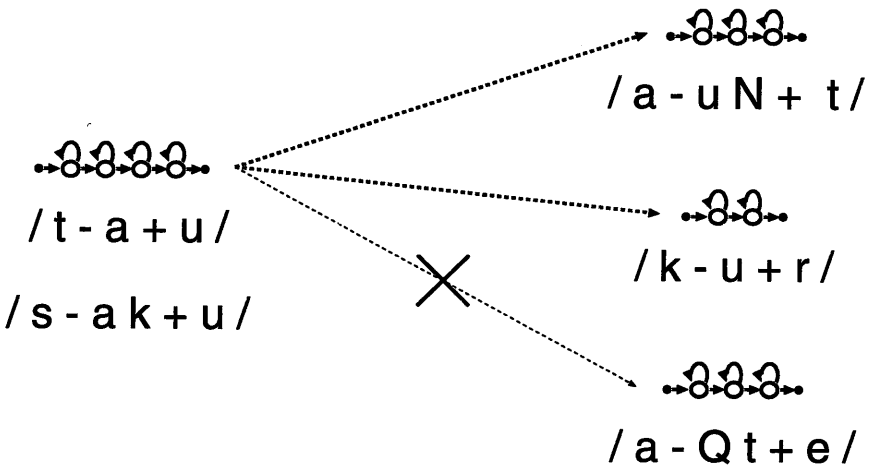

図 3.1: ASM の接続

デルとの接続を許した状態数 3 のコンテキスト非依存の音素 left-to-right HMM を ASM に加えて使用した。

3.3.2 評価方法

評価基準には以下の式で与えられる音素認識精度 $d(\mathbf{r},\mathbf{x})$ を用いた。$\mathbf{r}$ は正解音素系列、$\mathbf{x}$ は認識結果である音素系列である。

$$d(\mathbf{r},\mathbf{x}) = \frac{N-S-D-I}{N} \times 100$$

N : 全音素数
S : 置換誤りの音素数
D : 脱落誤りの音素数
I : 挿入誤りの音素数

ASM では、複数の音響セグメントが割り当てられているパスがある場合は認識結果を一意に定めることができない。そこで今回は、すべての可能性について列挙し、次式の $D(\mathbf{r},\mathbf{X})$ を音素認識精度の上限と定義して計算した。

$$D(\mathbf{r},\mathbf{X}) = \max_i d(\mathbf{r},\mathbf{x}_i), \quad \mathbf{X} = \{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$$

ここで、$\mathbf{x}_i$ は i 番目の認識結果の候補である音素系列であり、$\mathbf{X}$ はその集合、n は認識結果の全候補数である。

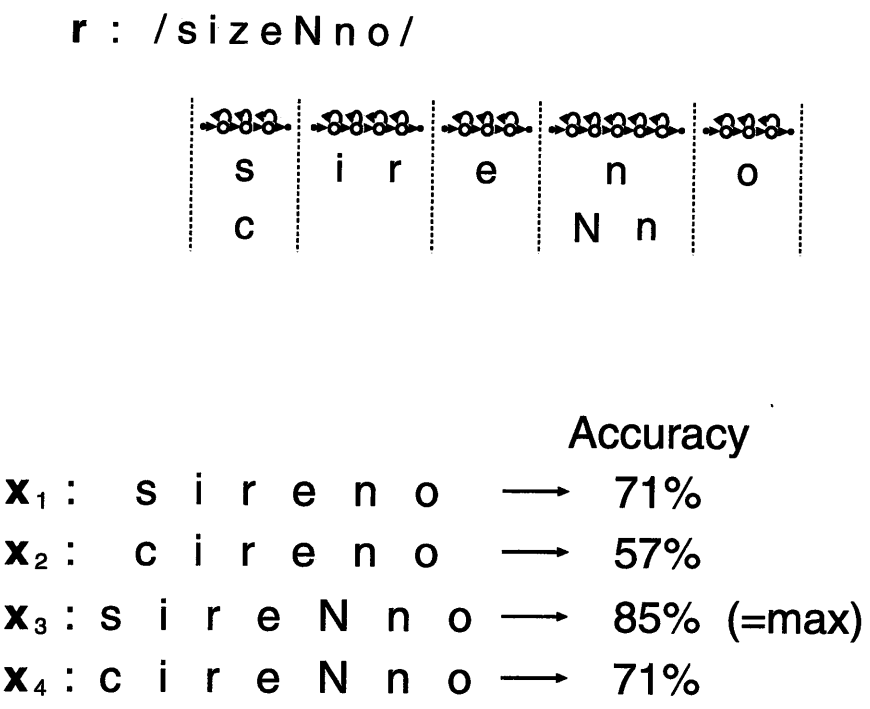

図 3.2: ASM の認識結果の評価例

図 3.2 に音素認識精度の上限を評価する具体的な例を示す。この例で正解音素系列は $\mathbf{r}$: / s i z e N n o / である。図 3.2 に示したような ASM のパス系列は 1 番目と 4 番目に 2 種類の音響セグメントがあるため、認識結果の候補が $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ の 4 通り存在する。評価の際は、$\mathbf{r}$ と $\mathbf{x}_1$ から $\mathbf{x}_4$ までのそれぞれの音素認識精度を計算する。そして最もよい値を評価値とする。図 3.2 の例の場合、$\mathbf{x}_3$ が 85 % と最も高い音素認識精度を示す。この値を評価値とすることで ASM の認識性能の上限値の評価を行う。

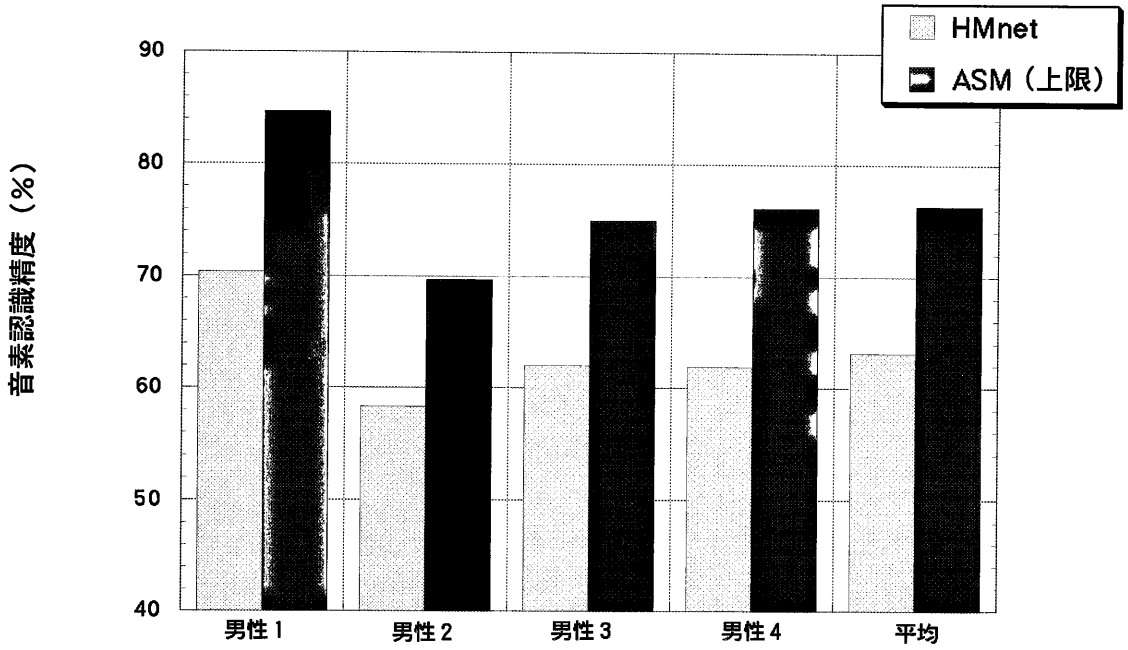

図 3.3: 男性 1 から 男性 4 の音素認識精度

3.3.3 実験結果及び考察

図 3.3 に男性 1 から 男性 4 の連続音声認識実験の結果を示す。比較のため、SSS-free で作成した状態数 500 の 音素 HMnet の結果もあわせて示す。音素 HMnet も ASM と同様にコンテキストを考慮してモデルの接続を制限した。全ての話者において、ASM の音素認識精度の上限値は音素 HMnet に比べて高い値を示した。平均で約 13 % 程度 ASM の上限値のほうが高かった。この実験結果より、ASM の認識結果は音素 HMnet よりも正解を数多く含んでいることが分かった。

表 3.2 から 表 3.5 に各話者ごとの音素認識精度について示す。ASM を用いた場合、全ての話者において挿入誤りが特に大きく減少していた。音素 HMnet に比べて 10 % 以上も挿入誤りが減っていた。音素 HMnet では発声時間の短い音素 (促音 / Q / など) は少ない状態数から成るパスとしてモデル化されることが多いが、そのような短いパスが挿入誤りの原因となっている傾向がみられた。ASM では発声時間の短い音素は多くの場合、前後の音素とまとまって 2 音素以上から成る音響セグメントとしてモデル化される。そのため ASM では挿入誤りの原因となるような短いパスの生成が抑えられたと考えられる。

表 3.2: 男性 1 の音素認識精度

	音素 HMnet	ASM (上限)
認識精度	70.45%	84.64%
置換誤り	5.34%	3.36%
脱落誤り	1.12%	0.72%
挿入誤り	23.08%	11.26%

表 3.3: 男性 2 の音素認識精度

	音素 HMnet	ASM (上限)
認識精度	58.34%	69.72%
置換誤り	11.67%	11.21%
脱落誤り	3.12%	2.65%
挿入誤り	26.86%	16.39%

表 3.4: 男性 3 の音素認識精度

	音素 HMnet	ASM (上限)
認識精度	62.05%	75.05%
置換誤り	9.11%	8.46%
脱落誤り	2.54%	1.83%
挿入誤り	26.29%	14.63%

表 3.5: 男性 4 の音素認識精度

	音素 HMnet	ASM (上限)
認識精度	61.91%	76.18%
置換誤り	8.99%	8.68%
脱落誤り	2.37%	2.41%
挿入誤り	26.71%	12.70%

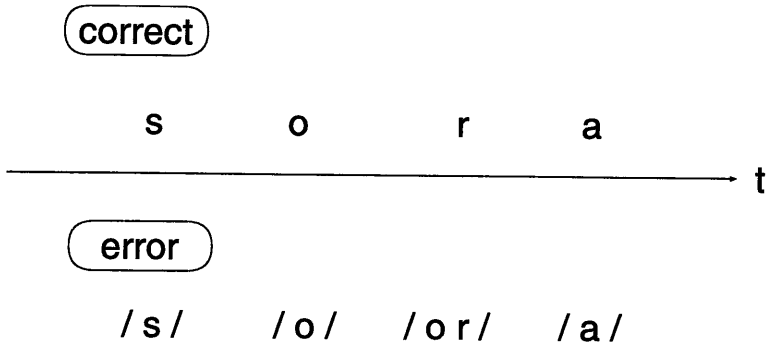

図 3.4: 挿入誤りの例 1

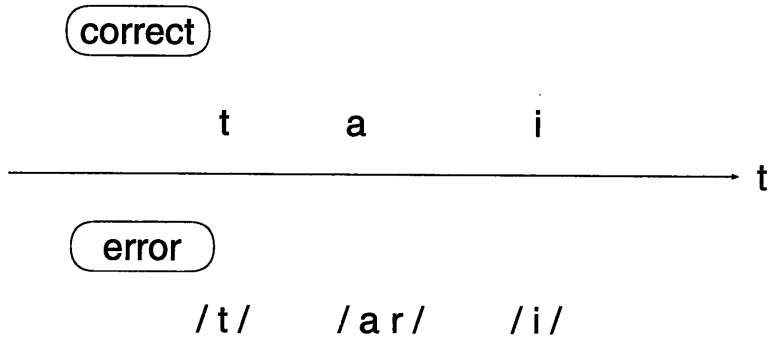

図 3.5: 挿入誤りの例 2

実験結果の例を挙げて、挿入誤りの傾向について考察する。ASM は 音素 HMnet に比べ挿入誤りが減少したが、それでも平均で約 13 % もの挿入誤りがあった。図 3.4 と 図 3.5 に ASM の挿入誤りの例を示す。図 3.4 の例では、音響セグメント / o / が音素 / o / の前半部分にマッチして、音響セグメント / o r / の前半部分が音素 / o / の残りの部分にマッチしてしまったため、挿入が起こったものと考えられる。図 3.5 の例では、音響セグメント / a r / の音響的な特徴が音素 / a / に非常によく似ていたため、音響セグメント / a r / と音素 / a / の部分がマッチしてしまい、/ r / が挿入されたものと考えられる。

3.4 まとめ

本章では、音声認識実験により ASM の性能を評価した。ASM は認識結果が一意に決まらない場合があるため、音素認識実験では平均候補数を導入して評価した。また、連続音声認識実験では ASM の有する性能の上限を評価した。

この 2 種類の実験結果より、ASM の認識結果は音素 HMnet に比べ、正解を多く含んでいることが分かった。このことは ASM を用いることによって認識精度を改善できる可能

性があることを示唆するものである。

第4章

ASM を用いた大語彙連続音声認識システム

4.1 はじめに

前章では ASM の性能を 2 種類の音声認識実験により評価した。その結果、ASM の認識結果は従来の音素モデルに比べて、正解を多く含んでいることが分った。このことより、ASM を実装した大語彙連続音声認識システムは高い認識性能を示すことが期待できる。

本章では、研究用に一般に公開されている大語彙連続音声認識システムを利用した実験を行い、ASM の性能を評価する。評価を行うために、ASM を大語彙連続音声認識システムに組み込む方法を提案する。前章では ASM の有する認識結果の曖昧性を残したままの評価であったが、本章で提案する方法では ASM の曖昧性を排除したものになっている。そのため、上限値ではない ASM の性能を評価することができる。

4.2 大語彙連続音声認識システム

本研究では大語彙連続音声認識に関する研究・開発の共通のプラットフォームとして開発された「日本語ディクテーション基本ソフトウェア」を使用する。この開発プロジェクトは、情報処理振興事業協会 (IPA) の「独創的先進的情報技術に係わる研究開発」の支援を受けている [22]。このソフトウェアは認識エンジン、日本語音響モデル、日本語言語モデル、日本語形態素解析・読み付与ツールなどから構成される。

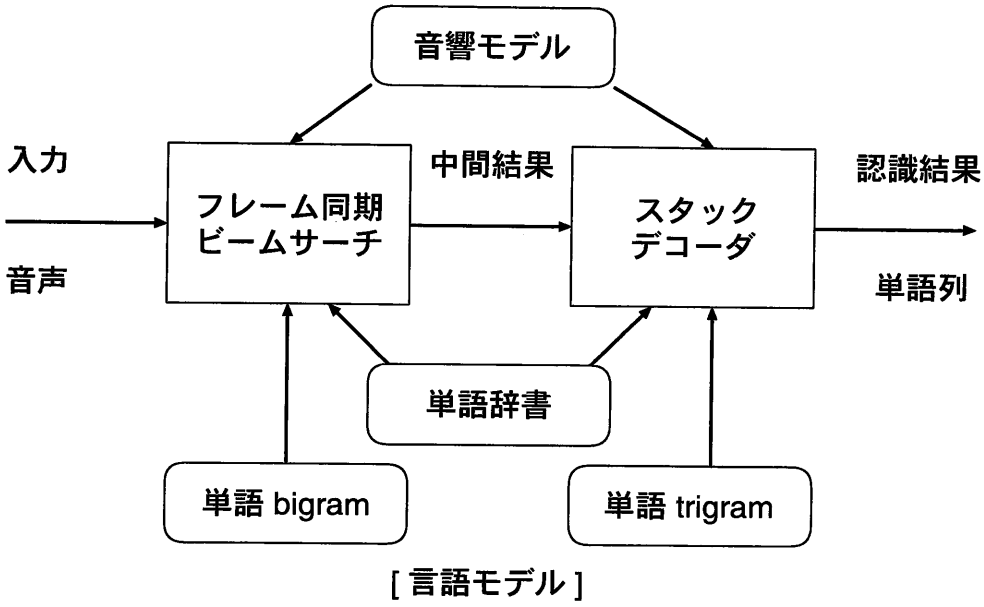

図 4.1: システムの構成図

4.2.1　システムの構成

大語彙連続音声認識システムの構成図を 図 4.1 に示す。認識エンジン JULIUS [23] (以下 JULIUS) の仕様に基づいて音響モデルと言語モデルが統合されている。種々のタイプのモデルを扱うことが可能なため、それらの評価に用いることができる。JULIUS は 2 パス探索を行う。入力音声は特徴量ベクトルに変換されて JULIUS に入力される。そしてまず、第 1 パスで簡易なモデルである単語 bigram により単語候補を絞った上で、中間結果を第 2 パスにわたす。第 2 パスでは高精度なモデルである単語 trigram を用いて再探索、再評価を行う。そして最終的な認識結果を出力する。音響モデルと単語辞書は、第 1 パス、第 2 パスの両方で使用される。

4.2.2　音響モデル

日本音響学会の音声データベース委員会で策定された 43 音素に基づいた音響モデルを使用する。音響モデルは、男性用モデル、女性用モデル、性別非依存モデルが用意されており、出力確率が混合ガウス分布 (対角共分散) である状態数 3 の left-to-right HMM である。モデルの種類は音素環境非依存モデルと音素環境依存モデルがある。音響モデルの学習には、日本音響学会の音素バランス文からなる研究用連続音声データベース (ASJ-PB) の全部と、毎日新聞記事読み上げ音声コーパス (ASJ-JNAS) [24] のうち 100 名分を利用している。合計で男女とも、約 130 名の話者による 2 万文のデータである。

表 4.1: システムで採用する音素一覧

a	o	u	i	e	N	w	y	f	s
sh	z	p	py	t	k	ky	b	by	d
dy	g	gy	r	ry	ts	ch	m	my	n
ny	h	hy	j	q	a:	o:	u:	i:	e:
sp	silB	silE							

表 4.1 に本研究で使用する大語彙連続音声認識システムで採用している音素の一覧を示す。表中で、a: ～ o: は長母音を、q は促音を表わす。silB は文頭、silE は文末、sp は文中 (単語間) のポーズ (無音区間) に対応している。

4.2.3 言語モデル

設定した語彙に基づいて作成した n-gram 言語モデル (単語 bigram と単語 trigram) を使用する。いずれもバックオフ平滑化を行っており、これらの言語モデルは CMU-Cambridge SLM ツールキット [25] により作成されている。ポーズに対応づけた句読点も通常の単語と同様に扱われており、句読点の出現確率の推定によりポーズの出現位置の推定を代用している。

4.2.4 単語辞書

単語辞書は毎日新聞記事 CD-ROM の、1991 年 1 月から 1994 年 9 月までの 45ヶ月分において、高頻度な 5000 単語により構成されている。単語辞書は音響モデルと言語モデルの両方と整合をとっている。

表 4.2 に単語辞書ファイルの形式を示す。単語名には品詞番号も付与されている。音素記号は 表 4.1 の音素の定義に基づいて作成した音響モデルでカバーされており、語彙中の全ての単語には言語モデルによって生起確率が定義される。出力シンボルとは認識結果として出力する文字列であり、単語の読みを示す。複数の読み方や音素表記を持つ単語については、それらを併記した形で一つの単語エントリとなっている。表 4.2 の例の中では “課長” という単語は音素記号列で表記すると / k a ch o: / と/ k a ch o u / の 2 通りがある。このような単語については、両方を並列に扱うことになる。

表 4.2: 単語辞書ファイル

単語名	出力シンボル	音素記号列
課税+1	[カゼイ]	k a z e i
課題+1	[カダイ]	k a d a i
課長+1	[カチョウ]	k a ch o:
課長+1	[カチョウ]	k a ch o u
過ぎ+過ぎる+102	[スギ]	s u g i
過ぎ+過ぎる+114	[スギ]	s u g i

4.3 ASM による単語辞書作成法

前節で説明した大語彙連続音声認識システムを用いて ASM を評価することを考える。システムの音響モデルとして ASM を組み込むためには、ASM のパス系列で単語を記述した単語辞書を作成する必要がある。この節では、ASM を用いた単語辞書の作成法を提案する。

ASM を用いた単語辞書の作成は以下のように行う。単語を記述することのできる ASM のパス系列の中から、その単語を最も少ないパスの数で記述するパスの組合せを採用する。言い換えるならば、2 音素以上から成る長い音響セグメントが割り当てられたパスを優先して用いることになる。長い音響セグメントが割り当てられた ASM のパスは長い距離のコンテキストをうまくモデル化したものになっていることが期待できると考え、このようなパスを優先して用いることにした。また、単語辞書作成時には、前後の 1 音素のコンテキストを考慮してパスを選択する。ASM のパスに存在しないコンテキストについては音素環境非依存の音素 HMM を代用する。単語の始端と終端の音素については、それぞれ先行、後続のコンテキストが不明であるため、音素環境非依存の音素 HMM を用いることにする。

図 4.2 に ASM を用いた単語辞書の作成法の例を示す。ここでは / n i N sh i k i / という単語の ASM によるモデル化の例を示している。単語の始端の / n / と終端の / i / はコンテキストが不明であるため音素 HMM を用い、その他の部分についてはコンテキストを考慮して ASM のパスを選択する。先行音素が / n / であり後続音素が / sh / である音響セグメント / i N / が割り当てられたパスと、先行音素が / N / であり後続音素が / i / である音響セグメント / sh i k / が割り当てられたパスが例では選ばれている。このようにして ASM のパス系列で表記した単語モデルを、全語彙について作成する。

単語辞書を作成する上で、以下に示す 2 つの条件を考慮して単語辞書作成に使用する ASM のパスを選別した。

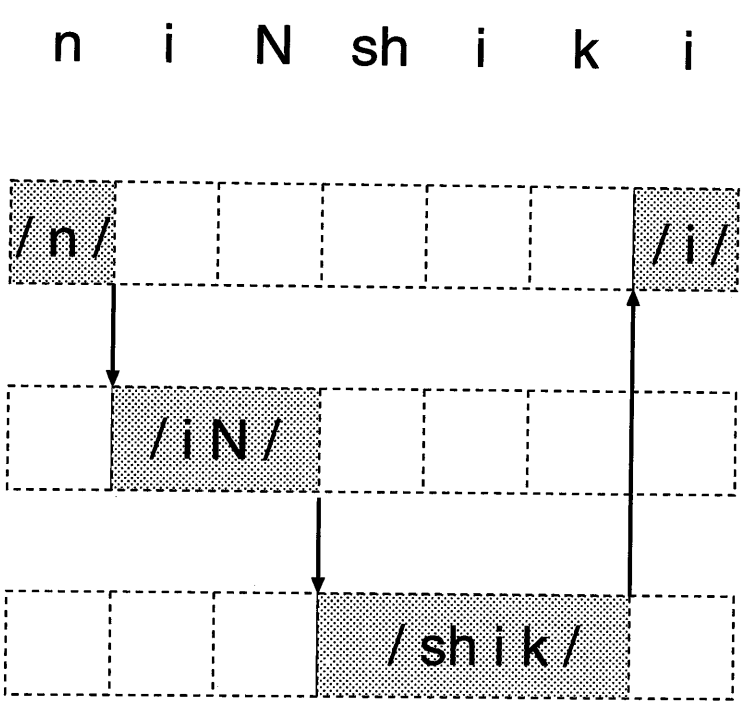

図 4.2: ASM による単語辞書作成法

- 音響セグメント占有率

  ASM では、パスに複数の種類の音響セグメントが割り当てられることがあるが、ここでは、同一の種類の音響セグメントのみが比較的よくまとまって割り当てられたパスに注目する。このようなパスは、音素との対応を持ちつつ、音響的なまとまりをよく反映したものになっていると考えられるため、積極的に認識時に用いることにする。

  表 4.3 に示した 2 種類のパスを例に挙げて説明する。パスに割り当てられている音響セグメントの中で、その数が最も多いものの割合をそのパスの音響セグメント占有率として定義する。パス 1 には 3 種類、合計で 10 個の音響セグメントが割り当てられており、音響セグメント / a m / の数が最も多い。そこで、パス 1 は音響セグメント / a m / を示すパスであると一意に定める。この場合の音響セグメント占有率は全体の中で / a m / が占める割合である 80 % となる。同様の基準で、パス 2 は音響セグメント / a N / を示すパスであると決める。パス 2 の音響セグメント占有率は 40 % である。単語辞書作成時には、ASM の中で音響セグメント占有率の高いパスのみ用いる。パス 2 のような音響セグメント占有率の低いパスは複数の種類の音響セグメントが同じような割合で混在している場合であるが、このようなパスは使用しないことにした。

表 4.3: 音響セグメント占有率

	/a m/	/a n/	/a N/	/a/
パス 1 (80%)	**8**	1	1	0
パス 2 (40%)	3	2	**4**	1

- 学習サンプル数

  得られる音響セグメントの中には、非常に学習サンプルに特化したものが含まれる場合があると考えられる。学習サンプルに特化した音響セグメントを持つパスは、音響セグメント占有率は高い場合が多いが、認識に用いた場合にテストサンプルとうまくマッチせず認識性能の低下を招く可能性がある。そのため、このようなパスは単語辞書作成時には使用しない。

  学習サンプルに特化した音響セグメントはその音響セグメントだけが割り当てられたパスとして ASM の中に出現する傾向がある。つまり、学習サンプルが少ないパスとして存在していることが多い。そこで、学習サンプルが少ないパスは単語辞書作成時に使用しない。このことにより、学習サンプルに特化したものが排除されることが期待できる。

4.4 大語彙連続音声認識システムによる ASM の性能評価

提案した方法により作成した ASM の単語辞書を使用することにより、 ASM を大語彙連続音声認識システムに組み込む。そして認識実験を行い、ASM の性能を評価する。

4.4.1 実験条件

特定話者 8 名 (男性 5 名、女性 3 名) について認識実験を行う。前章と同様に、学習のためのサンプルと、テストのためのサンプルは同じ話者の発声によるものである。単語辞書の語彙数は 5000 語であるが、実験で使用する ATR 連続音声データベースの 503 文には単語辞書に登録されている 5000 語以外の単語も存在する。このような単語が含まれた文は、今回の実験では完全に認識することができない。そこで認識実験で用いるテストサンプルにはなるべく単語辞書外の単語が含まれないような文を選んだ。503 文の中から、辞書に含まれる 5000 語の単語のみで構成されている文 12 文と、辞書に未登録の単語が 1 個

である文 54 文の計 66 文を選び、認識用のテストサンプルとした。テストサンプルの未知語率は約 7.4 % である。それ以外の 437 文は 2 個以上の未知語を含む文であるが、これらは ASM を生成するための学習サンプルとして用いた。

学習に用いる ATR 連続音声データベースの音素の定義は 表 4.1 に示すものと若干異なるため、ここでは 表 4.1 に示す本研究で使用する大語彙連続音声認識システムで採用している音素の定義に合わせた。ATR 連続音声データベースの音素の大部分は 表 4.1 の音素と対応がとれるため機械的に書き換えが可能である。対応がとれない無声化母音と二重母音については区別せず単なる母音として扱った。

言語モデルは学習用のコーパスとして毎日新聞の記事データ 75ヶ月分 (91 年 1 月～94 年 9 月、95 年 1 月～97 年 6 月) を用いて作成した単語 bigram と単語 trigram を使用した。n-gram エントリの閾値は単語 bigram と単語 trigram 両方とも 1 とした。

辞書の作成には、音響セグメント占有率 80 % 以上であり、かつ学習サンプル数が 3 個以上である ASM のパスのみを使用した。なお、ASM と 音素 HMnet は SSS-free に基づいて作成されているため、コンテキストにより一意にパスを決めることができず、1 つの単語を表記するパス系列が複数存在することがある。このように複数存在する場合は、JULIUS で使用することができる語彙数の最大は 30000 語程度なので、一単語につき最大 6 エントリまで許すこととした。パス系列が 6 種類以上存在する場合は、学習サンプルの多いパスが含まれるものを優先して用いた。

評価には単語認識精度を用いた。その式を以下に示す。式の形は前章の音素認識精度と同じであるが、単語単位でマッチングを行うところが異なる。

$$\text{単語認識精度} = \frac{N - S - D - I}{N} \times 100$$

N : 全単語数
S : 置換誤りの単語数
D : 脱落誤りの単語数
I : 挿入誤りの単語数

なお、単語認識精度を計算する際には、単語の漢字表記や品詞が正解と異なっていても、読みが同じであればそれは認識できたものとして扱った。

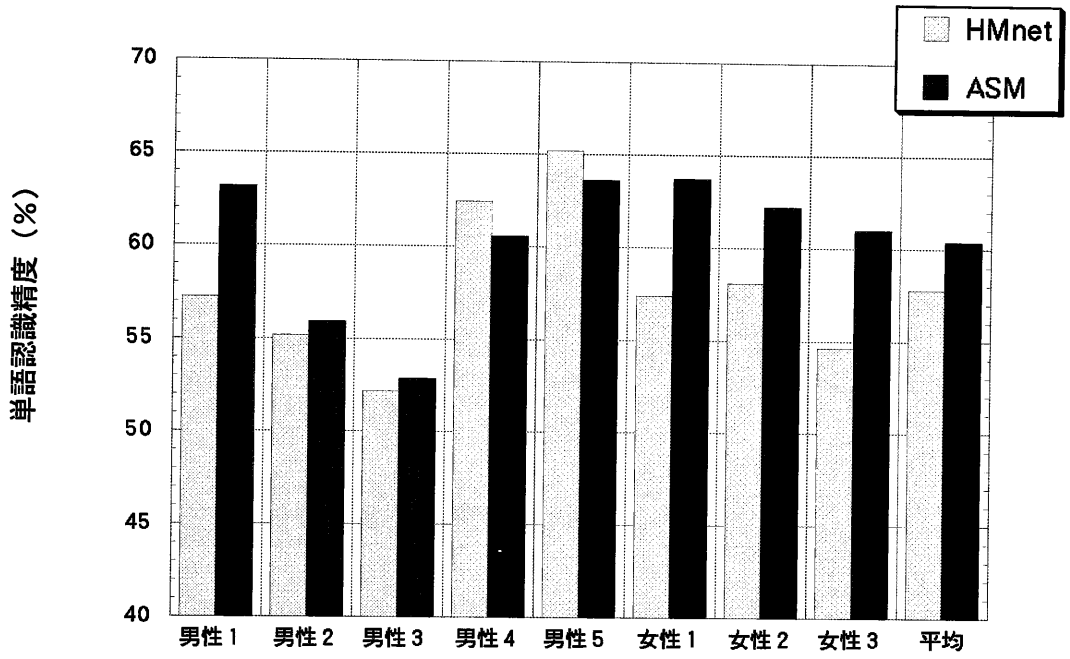

図 4.3: 全話者の単語認識精度

4.4.2 実験結果及び考察

提案した方法で ASM による単語辞書を作成し、認識実験を行った。比較のため音素 HMnet による単語辞書も作成し、実験を行った。8 名の話者 (男性 1～5、女性 1～3) について認識実験を行った。図 4.3 に全話者の単語認識精度を示す。男性 1～3 と女性話者全てにおいて ASM は音素 HMnet よりも高い性能を示した。ASM の効果が大きくみられた話者では、音素 HMnet に比べ 6 % 以上高い値を示した。平均で約 2.6 % の単語認識精度の向上がみられ、ASM の有効性が確認できた。

表 4.4 から 表 4.11 に、それぞれの話者についての単語認識精度を示す。全体的に、ASM は音素 HMnet に比べ挿入誤りを減らすことができた。男性 4 と男性 5 については ASM のほうが音素 HMnet よりも認識精度が悪かった。この 2 人の話者はもともと音素認識率が比較的高い話者であるため、ASM の効果がそれほど表われなかったものと考えられる。

表 4.4: 男性 1 の単語認識精度

	音素 HMnet	ASM
認識精度	57.22%	63.16%
置換誤り	29.40%	25.53%
脱落誤り	1.97%	1.84%
挿入誤り	11.42%	9.47%

表 4.5: 男性 2 の単語認識精度

	音素 HMnet	ASM
認識精度	55.18%	55.92%
置換誤り	32.50%	32.64%
脱落誤り	4.19%	4.55%
挿入誤り	8.13%	6.89%

表 4.6: 男性 3 の単語認識精度

	音素 HMnet	ASM
認識精度	52.16%	52.88%
置換誤り	34.51%	35.99%
脱落誤り	2.75%	2.75%
挿入誤り	10.59%	8.38%

表 4.7: 男性 4 の単語認識精度

	音素 HMnet	ASM
認識精度	62.43%	60.55%
置換誤り	26.70%	30.67%
脱落誤り	2.75%	2.75%
挿入誤り	8.12%	6.03%

表 4.8: 男性 5 の単語認識精度

	音素 HMnet	ASM
認識精度	65.18%	63.61%
置換誤り	25.00%	24.87%
脱落誤り	1.57%	2.36%
挿入誤り	8.25%	9.16%

表 4.9: 女性 1 の単語認識精度

	音素 HMnet	ASM
認識精度	57.39%	63.70%
置換誤り	31.24%	27.39%
脱落誤り	2.35%	2.62%
挿入誤り	9.02%	6.29%

表 4.10: 女性 2 の単語認識精度

	音素 HMnet	ASM
認識精度	58.08%	62.20%
置換誤り	28.52%	26.12%
脱落誤り	1.97%	2.89%
挿入誤り	11.43%	8.79%

表 4.11: 女性 3 の単語認識精度

	音素 HMnet	ASM
認識精度	54.64%	60.99%
置換誤り	32.94%	29.19%
脱落誤り	3.66%	2.88%
挿入誤り	8.76%	6.94%

4.4.3 ASM による改善例

図4.4から 図4.6に ASM による改善例を示す。図4.4の例では、音素 HMnet では “放送” と誤認識していた単語が ASM では “恐れ” と正しく認識された。それに伴い言語モデルのスコアが変化し、後続する単語 “が ／ ある” も正しく認識された。これは ASM による単語辞書 “恐れ” に含まれる音響セグメント / o r / の部分の効果と考えられる。図 4.5 の例では、音素 HMnet では “人 ／ 支援” と 2 単語に誤認識していた部分が ASM では “人間” と正しく認識された。この結果、置換誤りと挿入誤りが減少した。これは ASM による単語辞書 “人間” に含まれる音響セグメント / N g / の部分の効果と考えられる。図 4.6 の例では、音素 HMnet では “生き方” という単語に誤認識していた部分が ASM では “機 ／ から” という 2 単語に正しく認識された。この結果、置換誤りと脱落誤りが減少した。これは ASM による単語辞書 “から” に含まれる音響セグメント / a r / の部分の効果と考えられる。ここで示した改善例は、複数の音素をまとめてモデル化したことが有効にはたらいた例だと考えることができる。

4.5 まとめ

本章では、ASM を大語彙連続音声認識システムに実装する方法を提案した。提案した方法では、ASM の持つ認識結果の曖昧性を排除して ASM をシステムに組み込むため、従来の音素モデルを用いた場合と同様の条件の下で ASM の性能評価を行うことができた。現在の研究の主流となっている大語彙連続音声認識システムにより ASM の評価を行ったことは意義があると考える。大語彙連続音声認識システムを用いた実験では、ASM は音素 HMnet に比べ高い認識性能を示し、ASM の有効性を確認することができた。

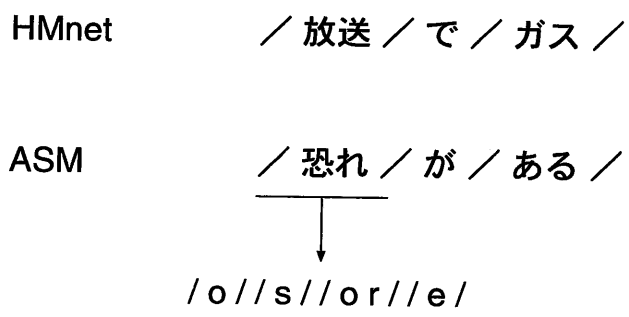

図 4.4: ASM による改善例 1

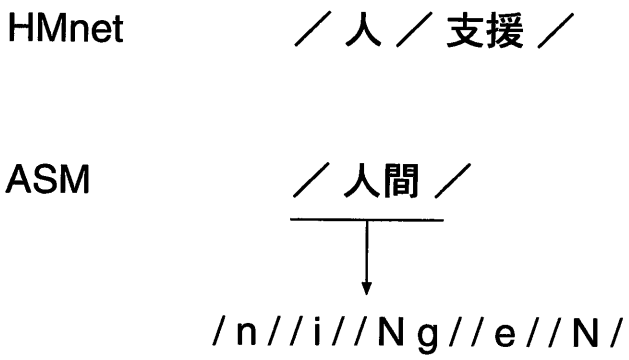

図 4.5: ASM による改善例 2

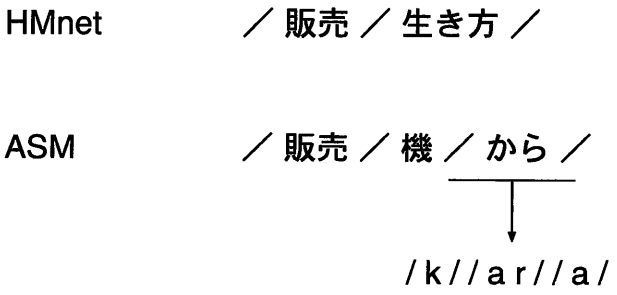

図 4.6: ASM による改善例 3

第5章

結論

5.1 本研究の成果

本研究の成果について述べる。

第 2 章では、新たな認識単位である音響セグメントの獲得と、その単位に基づくモデルである ASM の生成を行うアルゴリズムを提案した。そして提案アルゴリズムにより得られた音響セグメントと ASM の特徴について調査した。ASM では同一のパスに異種の音響セグメントが混在することがあり、認識結果が一意に決定しない可能性があるという問題があることが分かった。

第 3 章では、音声認識実験を行い、ASM の性能を評価した。ASM は認識結果が一意に決まらない場合があるため、認識実験では ASM の有する認識結果の曖昧性を考慮して評価を行った。その結果、ASM は音素 HMnet の認識結果に比べ、正解を多く含むことが分かった。

第 4 章では、ASM を大語彙連続音声認識システムに実装する方法を提案し、認識実験により ASM の性能を評価した。システムでの認識実験は ASM の有する認識結果の曖昧性を排除して行った。8 名の話者について行った認識実験では、ASM は音素 HMnet に比べて平均で約 2.6 % 高い単語認識精度を示し、提案したモデルの有効性が確認できた。

5.2 今後の課題

今後の課題としては以下のことが挙げられる。

第 4 章で行った認識実験では、単語間のコンテキストを考慮していない。単語間のコンテキストを用いて認識を行ったほうが高い認識性能が得られることが広く知られている [26]。そこで、ASM についても単語間のコンテキストを利用した認識方法を検討する必要があ

る。提案したアルゴリズムでは、単語間にまたがる音響セグメントが生成される可能性があり、その扱いについても考慮したほうがよいであろう。

また、本研究では特定話者についての実験を行ったが、現在は不特定話者の大語彙連続音声認識システムについての研究が盛んである。そこで今後は、ASM の話者による違いについて更に分析し、不特定話者のシステムに ASM を適用する方法についても検討する必要がある。

謝辞

本研究を進めるにあたり、多大な御指導とともにこの研究の機会を与えて下さった東北大学大学院工学研究科教授 阿曽弘具氏に心から感謝致します。

本論文をまとめる際に、貴重な御意見を頂きました東北大学電気通信研究所教授 矢野雅文氏に深く感謝致します。

音声認識の分野におきましては、東北大学大型計算機センター教授 牧野正三氏、宇都宮大学工学部助手 森大毅氏、東北大学大型計算機センター助手 鈴木基之氏に御指導をして頂いたことを深く感謝致します。

日々の研究におきましては、昼夜を問わず御指導、及び計算機環境の整備をして頂きました東北大学大学院工学研究科助教授 大町真一郎氏、ならびに阿曽研究室の皆様に感謝致します。

参考文献

[1] 鳥脇 純一郎, "認識工学," テレビジョン学会, 1993.

[2] 平沢 純一, 川端豪, "わかってうなずくコンピュータの試作," 信学技報, NLC97-54, SP97-87, pp.79–86 1997.

[3] J. L. Gauvain and L. Lamel, "Large Vocabulary Continuous Speech Recognition: From Laboratory Systems towards Real-World Application," 信学論, vol.J79-D-II, no.12, pp.2005–2021 1996.

[4] Michiel Bacchiani, Mari Ostendorf, Yoshinori Sagisaka and Kuldip Paliwal, "Unsupervised Learning of Non-uniform Segmental Units for Acoustic Modeling in Speech Recognition," 日本音響学会平成 8 年度春期研究発表会講演論文集, 1-5-15, pp.37–38 1996.

[5] 深田 俊明, Michiel Bacchiani, Mari Ostendorf, 匂坂 芳典, "音響的セグメント単位を用いた自由発話音声認識," 日本音響学会平成 8 年度春期研究発表会講演論文集, 1-5-16, pp.39–40 1996.

[6] 中川 聖一, 斎藤 稔, "エルゴディック HMM に基づく音声の自動獲得単位を用いた音声認識," 信学技報, SP97-9, pp. 55–62 1997.

[7] 向當 一洋, 谷口 秀次, 小泉 卓也, "サブワード単位離散単語認識システムの話者依存性の改善," 信学技報, SP98-47, pp.15–21 1998.

[8] 児島 宏明, 田中 和世, "区分線形セグメントラティスを用いた単語モデルの自動生成," 信学技報, SP95-106, pp.93–98 1995.

[9] 深田 俊明, 匂坂 芳典, "発音ネットワークに基づく発音辞書の自動生成," 音声言語情報処理, 14-3, pp.15–22, 1996.

[10] 新井 寿典, 樽松 明, "連続音声認識における可変長音声単位の構成法," 信学技報, SP98-118, pp.1–7 1999.

[11] L. Rabiner, B. Juang, 古井 貞熙監訳, "音声認識の基礎 (上)(下)," NTT アドバンステクノロジ株式会社, 1994.

[12] 中川 聖一, "確率モデルによる音声認識," 電子情報通信学会, 1988.

[13] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modeling," Proc. ARPA Human Language Technology Workshop, pp.307–312, 1994.

[14] M. D. Huang, X. D. Huang and F. Alleva, "Predicting Unseen Triphone with Senones," IEEE Transactions on Speech and Audio Processing, vol.4, no.6, pp.412–419, 1996.

[15] 篠田 浩一, 渡辺 隆夫, "情報量基準を用いた状態クラスタリングによる音響モデルの作成," 音声言語情報処理, 14-11, pp.75–81 1996.

[16] 鷹見 淳一, 嵯峨山 茂樹, "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, vol.J76-D-II, no.10, pp.2155–2164 1993.

[17] J. Takami and S. Sagayama, "A Successive State Splitting Algorithm For Efficient Allophone Modeling," Proc. ICASSP'92, pp.573–576, 1992.

[18] M. Suzuki, S. Makino, A. Ito, H. Aso and H. Shimodaira, "A New HMnet Construction Algorithm Requiring No Contextual Factors," IEICE Trans. Inf. & Syst., vol.E78-D, no.6, pp.662–668 1995.

[19] M. Ostendorf and H. Singer, "HMM Topology Design using Maximum Likelihood Successive State Splitting," Computer Speech and Language, pp.17–41 1997.

[20] 堀 貴明, 加藤 正治, 伊藤 彰則, 好田 正紀, "音素決定木に基づく逐次状態分割法による HM-Net の性能改善の検討," 音声言語情報処理, 14-12, pp.83–90 1996.

[21] 堀 貴明, 加藤 正治, 伊藤 彰則, 好田 正紀, "状態クラスタリングによる HM-Net の構造決定法の検討," 信学技報, NLC97-42, SP97-75, pp.47–52 1997.

[22] 河原 達也, 李 晃伸, 伊藤 克亘, 伊藤 彰則, 宇津呂 武仁, 小林 哲則, 清水 徹, 田本 真詞, 荒井 和博, 峯松信明, 山本 幹雄, 竹沢 寿幸, 武田 一哉, 松岡 達雄, 鹿野 清宏, "大語彙日本語連続音声認識研究基盤の整備–評価用連続音声認識プログラムの開発–", 情報処理学会研究報告, 97-SLP-18-1, 1997.

[23] 情報処理振興事業協会 (IPA), “日本語ディクテーション基本ソフトウェア–1998 年度版–”, 1999.

[24] 日本音響学会, “新聞記事読み上げ音声コーパス (JNAS)”, 1997.

[25] Cambridge University, “The CMU-Cambridge Statistical Language Modeling Toolkit v2”, http://svr-www.eng.cam.ac.uk/ prc14/toolkit.html, 1997.

[26] 李 晃伸, 河原 達也, “大語彙連続音声認識エンジン Julius における単語間 triphone の扱いの改善,” 日本音響学会講演論文集, 2-1-1, pp.55–56 1999.

研究業績

査読付国際会議

1. "Speech recognition using acoustic segment model based on Hidden Markov Network,"
   Takafumi Hayashi, Hiroki Mori, Motoyuki Suzuki, Shozo Makino, Hirotomo Aso,
   International Conference on Speech Processing, vol.1 pp.299–304 (1999-8).

学会発表等

1. "共分散行列の推定誤差を考慮した音素認識,"
   林 貴文, 森 大毅, 鈴木 基之, 大町 真一郎, 牧野 正三, 阿曽 弘具,
   東北大学電気通信研究所第297回音響工学研究会, 297-2 (1998-5).

2. "逐次状態分割法による音声認識単位の自動獲得,"
   林 貴文, 森 大毅, 鈴木 基之, 牧野 正三, 阿曽 弘具,
   日本音響学会1999年春季研究発表会, 1-1-7 (1999-3).

3. "HMnetに基づく音響セグメントモデルを用いた連続音声認識,"
   林 貴文, 森 大毅, 鈴木 基之, 牧野 正三, 阿曽 弘具,
   日本音響学会1999年秋季研究発表会, 1-1-9 (1999-9).