モーラ情報を用いた単語音声認識の研究

妹尾 貴宏

概要

現在，単語音声認識では特徴パラメータとしてケプストラムが使用されている．ケプストラムは，低次にフォルマント情報を，高次にピッチ情報を，それぞれ含んでいる．従来の音声認識では，入力パラメータとして，フォルマント情報が使用されている．ピッチ情報は，ピッチ周波数を安定して抽出するのが困難なため，通常使用されていない．しかし，最近の研究でピッチ周波数とモーラ情報には依存関係があることが知られている．本研究では，ピッチ情報の代わりにモーラ情報を使用することによりフォルマントにおけるピッチの影響を分離できると仮定し，モーラ情報を使用して単語音声認識を行った．その結果，多くの話者に対し認識率が向上した．また，モーラ情報を使用した認識結果は，従来の音声認識で有効な手法として知られる triphone モデルの認識結果とほぼ同等の結果が得られた．その結果，モーラ情報の有効性が認められた．

目 次

図 目 次

表 目 次

1　はじめに

現在の単語音声認識では，特徴パラメータとしてケプストラムやメルケプストラムなどが使用されている．このケプストラムは，低次にフォルマント情報を，高次にピッチ情報をそれぞれ含んでいる．従来の音声認識では，入力パラメータとして，フォルマント情報が使用されるが，ピッチ情報は，ピッチ周波数を安定して抽出するのは困難であるため，通常，使用されない．また，ケプストラムにおいて，フォルマントはピッチ周波数の影響を受けやすいことが知られている．

ところで，最近の研究でピッチ周波数と単語のモーラ数およびモーラ位置の間に依存関係が存在することが知られている．この依存関係を利用することにより，音声合成 [10] や音素ラベリング [11] の研究において，その品質や精度が向上することが確認されている．単語音声認識の研究においても，単語のモーラ数およびモーラ位置を考慮して音素 HMM を作成して認識を行った場合に，認識率の向上することが確認されている．このことから，モーラ情報を使用することによって，フォルマントにおけるピッチの影響を分離できると考えられ，HMM の精度の向上に役立つと考えられる．

また，従来の研究で，単語音声認識において語頭および語尾情報のみを考慮しただけでも，かなり認識率が向上することが知られている．そして，前後の音素環境を考慮した triphone は，現在の音声認識において最も有効な手法として知られている．

そこで，本研究では単語音声認識において，モーラ情報がどの程度有効であるかを，語頭語尾情報や triphone モデルと比較して実験を行い明らかにした．

実験の結果では，モーラ情報を用いることによって認識率は向上した．その割合は，full-covariance において，従来有効な手法である triphone とほぼ同程度であった．このことから，モーラ情報の有効性が認められた．

以下，本論文では，第 2 章で音声分析，第 3 章で HMM による単語音声認識，第 4 章でモーラ情報とピッチ周波数，第 5 章でモーラ情報用いた単語音声認識，第 6 章で考察，第 7 章でまとめについて述べる．

2　音声分析

音声は音声生成のモデルのパラメータによって効率よく表現され，音声の音響音韻的 (acoustic-phonetic) 性質はこのパラメータによって特徴づけられる．音声生成のモデルのパラメータのように音声の音響音韻的な性質を特徴づけるパラメータを音響パラメータと呼んでいる．

音声の音響音韻的な性質は音響パラメータによって特徴づけられるが，音声を構成する言語音の音韻識別 (phonemic discrimination) には，音響パラメータの全データが必要になるわけではない．音韻識別に必要な部分を特徴パラメータ (feature parameter) と呼んでいる [4].

従来経験的に，音声情報はそのスペクトル (正確には電力スペクトル) によって特徴づけられることが知られてきた．その一つの表現がフォルマント (formant) 構造である．音声波形のスペクトルが複数個の共鳴 (共振) 周波数の存在によって特徴づけられることは古くから知られており，その共振を周波数の低い方から順番に「第 1 フォルマント，第 2 フォルマント，…」と名付けられている．

波形とスペクトルの関係は原理的にはフーリエ変換で記述できる．従来その処理は，(アナログ技術的に) 帯域フィルタ (BPF) 群による周波数分析によって近似的に実現されてきた．最近になって，計算機による高速フーリエ変換 (FFT) の技術が実用化され，ディジタル化された波形からそのスペクトルを FFT によって直接求めることがスペクトル分析の主流となっている [3].

2.1　音声分析の原理

音声は，さまざまな音素に対応する言語音から構成されていて，信号の性質が常に変化している非定常信号 (non-stationary signal) であるが，100 分の 1 秒程度の短時間区間ではいちおう定常的な信号とみなすことができるので，音声信号のスペクトル分析において定常過程 (stationary process) に対するスペクトル推定 (spectral estimation) の方法を利用することができる．

音声情報処理においては，音声波に含まれる位相情報は通常聴覚に作用しないため，位相情報を取り除いた周波数スペクトルの領域で処理を行なうのが普通である．音声スペクトルは，前述したように時間的に変化するので，音声波をほぼ定常と見なせる区間の長さを有する時間窓を，一定の周期でずらしながら乗じて波形を切り出し，それぞれについて短時間スペクトル (特徴ベクトル) を求める．こうして切り出した音声区間のこ

とをフレーム (frame)，この長さをフレーム長，フレームをずらす周期をフレーム周期と呼ぶ.

時間窓は，切り出し区間の両端に急激な変化が起こらず徐々に 0 になるようにすることが必要である．これにより音声のスペクトルに，窓関数のフーリエ変換の畳込み，すなわち重み付き移動平均が施される．このため窓関数としては，次のような性質をもつことが望ましい.

- 周波数分解能が高い.
- 畳込みによって他の周波数成分から生ずるスペクトルの洩れが少ない.

この両者を完全に満たすことはできないが，音声分析には次のいずれかの窓関数が用いられることが多い.

$$\text{ハミング}\,(Hamming)\,\text{窓}\quad : W_H(n) = \quad 0.54 - 0.46\cos(\frac{2n\pi}{N-1}) \tag{1}$$

$$\text{ハニング}\,(Hanning)\,\text{窓}\quad : W_N(n) = \quad 0.5 - 0.5\cos(\frac{2n\pi}{N-1}) \tag{2}$$

波形はこれらの窓を乗ずると，窓関数の両端に近い波形が圧縮されるので，等価的にフレーム長が 40%程度短くなると同時に，周波数分解能が 40%程度悪くなる．このことを考慮して，フレーム長を 20〜30ms 程度にし，周期をフレーム長の半分以下にして，分析区間を一部重複させる.

スペクトルの分析方法は，パラメトリック分析法 (PA:parametric analysis) とノンパラメトリック分析法 (NPA:non-parametric analysis) に分けることができる．PA の代表的なものに，線系予測分析があり，NPA の代表的なものにケプストラム分析がある [2].

2.2 ケプストラム分析

もし音声の言語情報が声道の形状による共振特性によって担われていると仮定すれば，分析によって抽出する特性はまずそのスペクトル包絡である.

時間 (波形) 的には，音声波形は音源波形と声道共振系のインパルス応答との畳込み (convolution) で表される．したがって，周波数次元では両者の特性の積で表される.

音源と共振系の特性を分離して抽出する方法は逆畳込み (de-convolution) と呼ばれる.

その方法の一つが以下に述べるケプストラム (cepstrum) 分析である．その処理の基本を図 1 に示す.

音声波形のフーリエ変換の結果は，音源と共振系のスペクトル特性の積で表される.

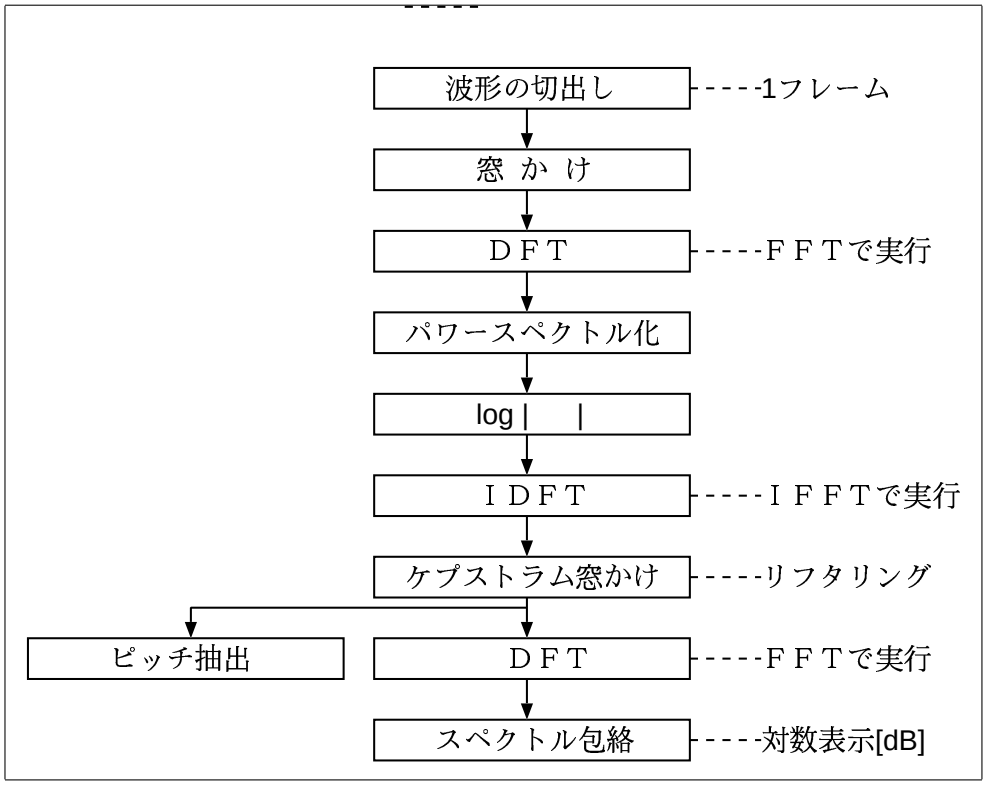

図 1: ケプストラム分析の手順

そこで，以下の操作を行なう．

1. スペクトルの対数をとることによって積を和に分解する．

2. フーリエ逆変換をとる．

もし音源の線スペクトルの基本周波数が低く (音源ピッチが低く) 高調波間隔が狭ければ，それを波形と見なしたときの周波数は高いから，その逆変換スペクトルの周波数は高い．一方，スペクトル包絡を形成するフォルマント間隔が互いにあまり近接していなければ，それを波形と見なしたときの包絡 (波形) のスペクトルは，低周波成分をもつ．したがって，逆変換された次元で，低域成分は包絡に，高域成分は音源パルス列に対応する．そこで適当な周波数分離を行なえば，音源成分と包絡成分とを分離できる．

このような原理による分析 (音源スペクトルとスペクトル包絡の分離) 法をケプルトラム分析という．

この逆変換次元を，周波数の逆次元量ではあるが時間そのものではないという意味で “quefrency”(ケフレンシイ)，変換された結果を “cepstrum”(ケプストラム)，ケフレンシイによる分離操作を “liftering”(リフタリング) という．

高ケフレンシイ成分として分離されたケプストラム成分のピーク値からピッチ周期が，さらにその成分の逆変換によって音源ピッチパルス列を再生することができる．

ケプストラム分析では，音源パルスの周期が長く，フォルマントが相互によく分離しているときはよい結果が得られるが，音源ピッチが高く，フォルマントが互いに近接しているときは両成分の分離が不完全となり，誤差が生ずる [3].

2.3 メルケプストラム

人間の聴覚特性は，音の大きさに対してはほぼ対数的な特性で，周波数分解能は低い周波数では細かく，高い周波数では粗いメル尺度で特徴づけられる．前者に対してはケプストラムで既に表現されているので，更に周波数軸をメル尺度に非線形変換したものをメルケプストラム (mel-cepstrum) とよんでおり，音声認識に有効な特徴パラメータとなっている [1].

周波数軸の伸縮する手法としては，全域通過フィルタ

$$H_\alpha(z) = \frac{(z^{-1} - \alpha)}{1 - \alpha z^{-1}} \tag{3}$$

を用いた手法が提案されている．この全域通過フィルタで近似的に周波数軸を，次式のようにメル尺度に変換できる．

$$\omega_{new} = \omega + 2\tan^{-1}(\frac{\alpha \sin\omega}{1 - \alpha\cos\omega}) \tag{4}$$

ここで，ω は正規化角周波数で，ω_{new} は変換された正規化角周波数である．また α を変えることによって周波数軸の伸縮の度合を任意に設定することができるので，α を周波数伸縮パラメータ (frequency warping parameter) と呼ぶ．サンプリング周波数が，6.67kHz, 8kHz，10kHz では，それぞれ $\alpha = 0.28, 0.31, 035$ とすれば，メル尺度をよく近似できる [4][5].

2.4 音源情報

音源情報の特徴パラメータには，有声音 (パルス音源)/無声音 (雑音源) の区別 (声帯振動の有無)，有声音の場合の基本周期 (基本周波数：ピッチ)，音源の振幅の 3 種類がある．ピッチ周波数はほぼ声の高さに対応する物理量である．正確なピッチ周波数の抽出 (ピッチ抽出) は，次の理由により確立されていない．

- 特に語頭や語尾において，声帯振動が完全な周期性をもたない.
- 音声波 (スペクトル) 中の声道特性を声帯信号を正確に分離するのが難しい.
- 基本周波数の変化範囲が広い

従来，音声の特性は，周波数スペクトルの形状によって，よりよく特徴づけられ，また，正確なピッチ周波数の抽出は困難なため，音源情報の特徴パラメータは，通常音声認識には用いられない.

3　HMMによる単語音声認識

単語ごとに区切って発声した音声を認識することを孤立単語音声認識 (isolated word speech recognition) といい，通常，これを簡単に単語音声認識 (spoken word recognition) あるいは単語認識 (word recognition) と呼ぶ.

単語音声の認識には，DP マッチング (dynamic programming matching) による方法，セグメンテーションと音素ラベリングに基づく方法，HMM による方法，ニューラルネットワークによる方法などが利用される.

単語数が数百程度の小語彙の単語音声の認識に対しては，DP マッチング法が効果的ではあるが，認識対象の単語の数が数千語以上になると，単語音声の音響パラメータの時系列に対する DP マッチングによる単語音声認識の方法は計算量が膨大になることから実際的ではなくなる.

大語彙単語に対する音声認識は，音声信号の音素レベルの記号化をもとにして行なわれる．音素レベルの記号化はセグメンテーションと音韻識別によって行なうか，音韻スポッティングによって行なう．音韻の HMM を接続した単語モデルに基づく単語音声認識においては一種の音韻スポッティング行なっていることになる．[4]

本論文では，二千語以上の単語の認識を想定して実験を行なうため，音素 HMM を用いて単語音声認識を行なう．以下，この章では HMM を用いた単語音声認識について述べる.

3.1　HMM法

3.1.1　HMMの理論

（1）マルコフモデル

確率的な生起事象の系列 (過程) を考える．各事象間に関連 (相関) のある場合を考え，これをマルコフモデル (Markov model) またはマルコフチェインという．その相関の程度によって単純マルコフモデルと多重マルコフモデルがある.

単純マルコフモデルというのは，個々の事象の生起が，事象自身によっての決まる確率に支配されるときであり，多重マルコフモデルというのは，それまでに生起した (過去の) 事象と関連で決まる確率に支配される場合である.

（2）隠れマルコフモデル

マルコフモデルの確率的な自由度をより拡大したモデルとして隠れマルコフモデル (HMM:Hidden Marcov Model) がある．このモデルでは，状態 (内部状態) と出力シンボルの 2 過程を考え，状態が確率的に遷移するとともに，それに応じてシンボルを確率的に出力すると考える．そのとき，外部からは状態の遷移は直接的には観測できず，出力シンボルのみが観測可能であるとする．この意味で隠れマルコフモデルという [3].

HMM はパラメータとして状態遷移確率，シンボル出力確率，初期状態確率を持つ．そして，シンボル出力確率の計算方法によって離散型 HMM と連続分布型 HMM に別れる．以下では，離散型 HMM について述べる [1].

T	： 観測系列の長さ
$o_1, o_2, \cdots, o_T$	： 観測系列
N	： 状態数
L	： 観測シンボルの数
$S = s$	： 状態集合
s_t	： 時刻 t の時の状態 (番号)
i, j	： 状態番号
$v = v_1, v_2, \cdots, v_L$	： 出力可能なシンボル集合

と定義すると，このオートマトンは，状態遷移確率 A，シンボル確率 B，初期状態確率 π は，以下のように示される．

$$A = \{a_{ij} | a_{ij} = P(s_{t+1} = j | s_t = i)\} \qquad (1 \leq i, j \leq N) \tag{5}$$

$$B = \{b_{ij}(o_t) | b_{ij}(o_t) = P(o_t | s_{t-1} = i, s_t = j)\} \quad (1 \leq i, j \leq N, 1 \leq t \leq T) \tag{6}$$

$$\pi = \{\pi_i | \pi_i = P(s_0 = i)\} \qquad (1 \leq i \leq N) \tag{7}$$

これらのパラメータを用いて，HMM を次のように略記する．

$$\lambda = (A, B, \pi) \tag{8}$$

観測系列 O が

$$o_1, o_2, \cdots, o_T \qquad (o_t = v_k, 1 \leq k \leq L, 1 \leq t \leq T)$$

という系列を生成する過程は次のようになる．

1. 初期状態確率を π にしたがって決定する．

2. 次に遷移する状態 ($s_{t+1} = j$) を現在の状態 ($s_t = i$) と状態遷移確率 a_{ij} にしたがって決定する．

3. 状態遷移する際に出力するシンボルをシンボル出力確率 $b_{ij}(o_t)$ にしたがって決定する．

4. 2. に戻る．

HMM には，ある状態から全ての状態に遷移できる全遷移型 (Ergodic) モデルや，状態遷移が一定方向に進む left-to-right モデルがある．図 2 に簡単な HMM(left-to-right モデル) の例を示す．

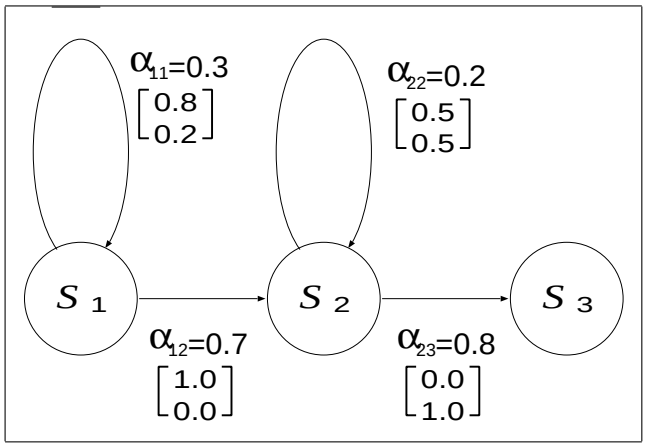

図 2: 3 状態 left-to-right モデル

この HMM は三つの状態で構成され，2 種類のシンボル a と b からなる．初期状態確率は $\pi_1 = 1.0, \pi_2 = 0, \pi_3 = 0$，最終状態を S_3 とし，図のような遷移のみを行なうものとする．a_{ij} は，状態 S_i から S_j への遷移確率を示し，[ ] 内の数字に上段はラベル a の出力確率，下段はラベル b の出力確率を表す．状態 S_1 を例にとれば，状態 S_1 から状態 S_1 自身に 0.3 の確率で遷移し，遷移の際に 0.8 の確率で a を出力し，0.2 の確率で b を出力する．他の状態，遷移についても同様である．ここで，出力シンボルが”aab” であった場合，状態遷移系列は $S_1 - S_1 - S_2 - S_3$ と $S_1 - S_2 - S_2 - S_3$ の 2 種類で，それぞれの確率は，

$$0.3 \times 0.8 \times 0.7 \times 1.0 \times 0.8 \times 1.0 \quad = \quad 0.1344$$

$$0.7 \times 1.0 \times 0.2 \times 0.5 \times 0.8 \times 1.0 \quad = \quad 0.056$$

である．

よって，この HMM が”aab” を出力する確率は二つの合計，

$$0.1344 + 0.056 = 0.1904$$

となる．

HMM では状態系列に意味を持たないが，最尤の経路を推定することはできる．この例では，”aab” を出力する可能性が最も高い状態系列は，前述の計算から容易に $S_1-S_1-S_2-S_3$ とわかる (3.2.1 参照)．

3.1.2 HMM の種類

HMM にはスペクトルパターンの表現方法により，離散型 HMM，連続分布型 HMM に大別される．また，離散型 HMM と連続分布型 HMM の中間的な性質を持った半連続型 HMM などもある．以下，離散型 HMM，連続分布型 HMM，半連続分布型 HMM の特徴について述べる [2]．

- **離散型 HMM(Descrete HMM)**

  出現されるスペクトルパターンは，有限個のシンボルの組合せで表現される．出力確率は，スペクトルパターンのクラスタ化 (ベクトル量子化) によって，代表スペクトルパターン (符号ベクトル) を生成し，各符号ベクトルの出現確率の組合せによって表現する．図 2 は 2 個のシンボルを出力する離散型 HMM である．

- **連続分布型 HMM(Continuous HMM)**

  出現するスペクトルパターンは，連続値で表現される．出力確率は，単一ガウス分布 (正規分布)，または混合ガウス分布で表現される．パラメータの自由度を減らすために無相関ガウス分布 (Digonal) が用いられることが多い．

- **半連続分布型 HMM(Semi-continuous HMM)**

  連続分布モデルと離散分布モデルの中間の性質を持つ．これは，連続分布モデルにおける混合ガウス分布を，すべてのモデルのすべての状態で共通にし，各分布の重みだけを変え るようにしたものである．結び混合分布モデル (tied-mixture model) とも呼ばれる．離散分布モデルにおける各符号ベクトルに確率分布を持たせたものということもできる．

3.2 HMM の基本アルゴリズム

音声認識を実行するためのアルゴリムには，以下のようなものが必要である．

1. 単語の隠れマルコフモデルの作成アルゴリズム

2. 事後確率計算のアルゴリズム

これらのアルゴリズムは，マルコフモデルとしての数学的な基礎が確立されており，それを使って学習サンプルから厳密に計算 (推定) することができる．

3.2.1 Viterbi アルゴリズム

Viterbi アルゴリズムはモデル λ において最適な状態系列 (最短経路) $S = s_1, s_2, \cdots, s_T$ と，この経路上での確率を求めるアルゴリズムである．

モデル λ において観測系列 $O = o_1, o_2, \cdots, o_T$ に対する最適な状態系列 $s = s_1, s_2, \cdots, s_T$ を求めるために，時刻 t で状態 i に至るまでの最適状態確率 $\delta_t(i)$ を定義する．

$$\delta_t(i) = \max_{s_1, s2, \cdots, s_{t-1}} p(s_1, s_2, \cdots, s_t = i, o_1, o_2, \cdots, o_t|\lambda) \tag{9}$$

時刻 $t+1$ における最適状態の確率は次のように導出できる．

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \cdot b_{ij}(o_{t+1}) \tag{10}$$

時刻 t 状態 i において生成確率を最大にする経路 (状態遷移) を $\Psi_t(j)$，最適経路の生成確率を p^*，最適経路上の最終状態を s_T^* とすると最適経路，およびその生成確率は以下の手順で求まる．

1. 初期化：

$$\delta_0 = \pi_i \Psi_0(i) = 0 \qquad (1 < i < N) \tag{11}$$

2. 繰返し：

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij} b_{ij}(o_t)] \tag{12}$$

$$\Psi_t = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij} b_{ij}] \qquad (1 < t < T), (1 < j < N) \tag{13}$$

3. 最終チェック：

$$p^* = \max_{1\le i\le N}[\delta_T(i)] \tag{14}$$
$$s_T^* = \arg\max_{1\le i\le N}[\delta_T(i)] \tag{15}$$

4. 経路 (back) トレース：

$$s_t^* = \Psi_{t+1}(s_{t+1}) \qquad (t = T-1, \cdots, 1) \tag{16}$$

4. で求めた $s_0^*, s_1^*, \cdots, s_T^*$ が最適経路となる．

3.2.2 Baum-Welch アルゴリズム

観測系列の生成確率を最大にするモデル λ のパラメータの局所的最適値を求める方法として，Baum-Welch アルゴリズム (パラメータ再推定法) がある．

モデル λ が観測系列 $O = o_1, o_2, \cdots, o_T$ を生成する場合において，時刻 t で状態 i から状態 j に遷移する確率 $\xi_t(i,j)$ を次のように定義する．

$$\xi_t(i,j) = P(s_{t-1} = i, s_t = j|O, \lambda) \tag{17}$$
$$= \frac{\alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{P(O|\lambda)} \qquad (1 \le t \le T) \tag{18}$$
$$\tag{19}$$

ここで，シンボル生成過程で，時刻 t で状態 j にいる確率 $\gamma_t(j)$ を定義する．

$$\gamma_t(j) = P(s_t = j|O, \lambda) \tag{20}$$
$$= \sum_{i=1}^{N} \xi_t(i,j) \qquad (1 \le t \le T) \tag{21}$$

この $\gamma_t(i)$ と $\xi_t(i,j)$ からモデル λ の再推定 $(\lambda \to \bar{\lambda})$ を次のように行なう．

1. 初期状態確率

$$\bar{\pi_i} = \gamma_0(i) = \frac{\alpha_0(i)\beta_0(i)}{P(O|\lambda)} \qquad (1 \le i \le N) \tag{22}$$

2. 状態遷移確率

$$\bar{a_{ij}} = \frac{\sum_{t=1}^{T}\xi_t(i,j)}{\sum_{t=1}^{T}\gamma_{t-1}(i)} = \frac{\sum_{t=1}^{T}\alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{\sum_{t=1}^{T}\alpha_{t-1}(i)\beta_{t-1}(i)} \tag{23}$$

3. シンボル出力確率

$$\bar{b}_{ij}(O_t) = \frac{\sum_{t\in(o_t=v_k)} \xi_t(i,j)}{\sum_{t=1}^{T} \xi_t(i,j)} = \frac{\sum_{t\in(o_t=v_k)} \alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)} \tag{24}$$

再推定された $\bar{\lambda}$ の評価は次のようになる．

1. $\bar{\lambda} = \lambda$ → (局所的な) 収束状態

2. $P(O|\bar{\lambda}) > P(O|\lambda)$ → シンボル系列 O を出力するより最適なモデル λ を推定

Baum-Welch アルゴリズムは，学習データの尤度を最大にするようにパラメータを学習する．ただし，基本的には gradient 学習によるパラメータ収束の学習方法であるため，local maxmum の方向にしか学習は進まない．そのため Baum-Welch アルゴリズムは初期値が重要になる．

3.2.3 連結学習

音素 HMM を学習するためには，大量の音声データを用いる必要があるが，その音声データに，逐一，人手によるラベル付けを行なうことは非常に困難である．そこでラベル付けされていない音声データベースを用いて，音素の学習を行なう方法が，連結学習である．ただし，音声データベース中の各単語には発声内容を示すテキストが添付されている．連結学習では，このテキストに基づいて音素モデルを連結し，文モデルを合成する．連結は図 3 のように行なわれる．音素モデルの連結においては，音素モデルの最終状態が次の音素の初期状態となるように連結を行なう．また，連結した文モデルの始終端には無音モデルを付け加える．その後，学習データごとにモデルを作り，Baum-Welch アルゴリズムによって HMM のパラメータの学習を行なう．最後に，Viterbi アルゴリズムを用いて各音素の境界を決定し，文モデルを分解して，各音素ごとの HMM を推定する．連結学習を行なうためには，初期モデルがきわめて重要である．通常は，ラベル付けされた音声データを用いて初期モデルを構成する．その後，大量の単語発声の音声データから HMM のパラメータを学習する [6].

3.3 単語音声認識の原理

現在の音声認識において，HMM の基本単位として，単語よりも短い音声単位 (サブワード単位) が広く用いられている．サブワード単位は，単語を単位とする場合と比較して次のような利点を持っている．

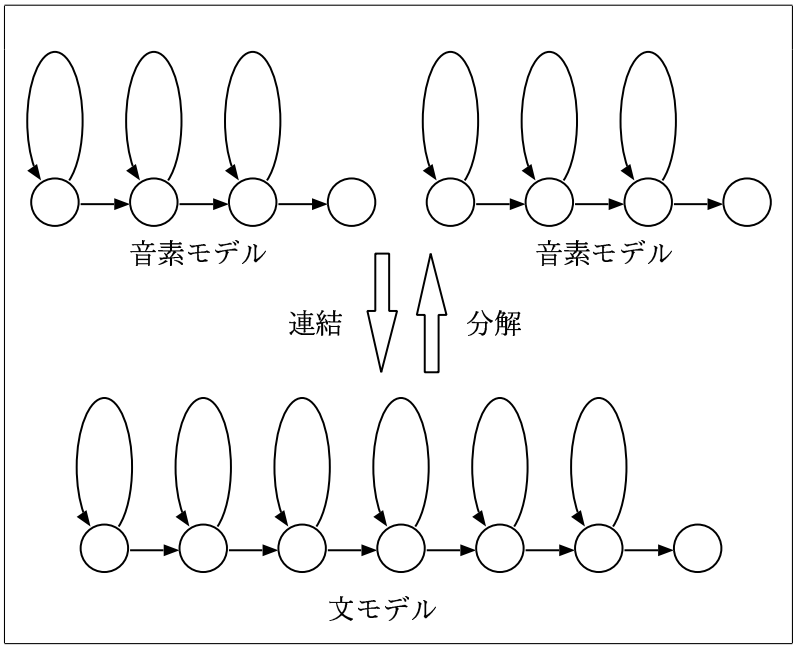

図 3: 音素モデルの連結による文モデルの合成

- 少ない HMM で単語 HMM を作成することができる
- 学習データに依存しない単語に対する単語 HMM を表現できる．

サブワード単位を音素とした場合の，HMM を用いた単語音声認識の基本構成を図 4 に示す．音声認識に対する言語音の単位として音素をとる．各音素はそれぞれ HMM によって表現され，単語は音素 HMM を連結した形の単語 HMM で表される．

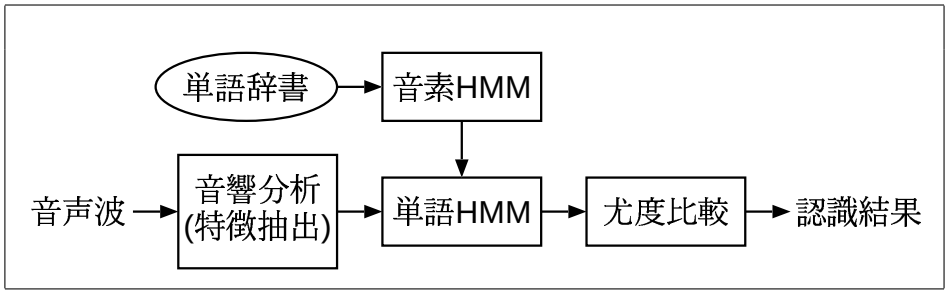

図 4: HMM を用いた単語音声認識の基本構成

3.3.1 音響分析 (特徴抽出)

音声認識の入力パラメータは，音声波から一定周期ごとに，短時間スペクトル (密度) を抽出して，これを用いることが多い．現在，音声認識に用いられている短時間スペクトル分析の手法としては，帯域フィルタ群を用いる方法，FFT を用いて直接的にスペクトルを計算する方法，LPC 分析に基づく方法などがある．現在では，FFT あるいは LPC によるスペクトルをケプストラムに変換して用いることが多い．

ここでは，ハードウェアによる実時間分析の実現が容易であることから帯域フィルタ群による特徴抽出方法について述べる．

3.3.2 帯域フィルタ群による特徴抽出方法

人の聴覚は，音の高さに関して，メル (mel) 尺度と呼ばれる対数に近い非線形の特性を示し，低い周波数では細かく，高い周波数では粗い周波数分解能をもつ．このため，各帯域フィルタを対数周波数軸上あるいはメルスケール上に等間隔に配置することが多い．FFT によるスペクトルを元に，メルスケールの帯域フィルタ群出力を抽出する手順を図 5 に示す．

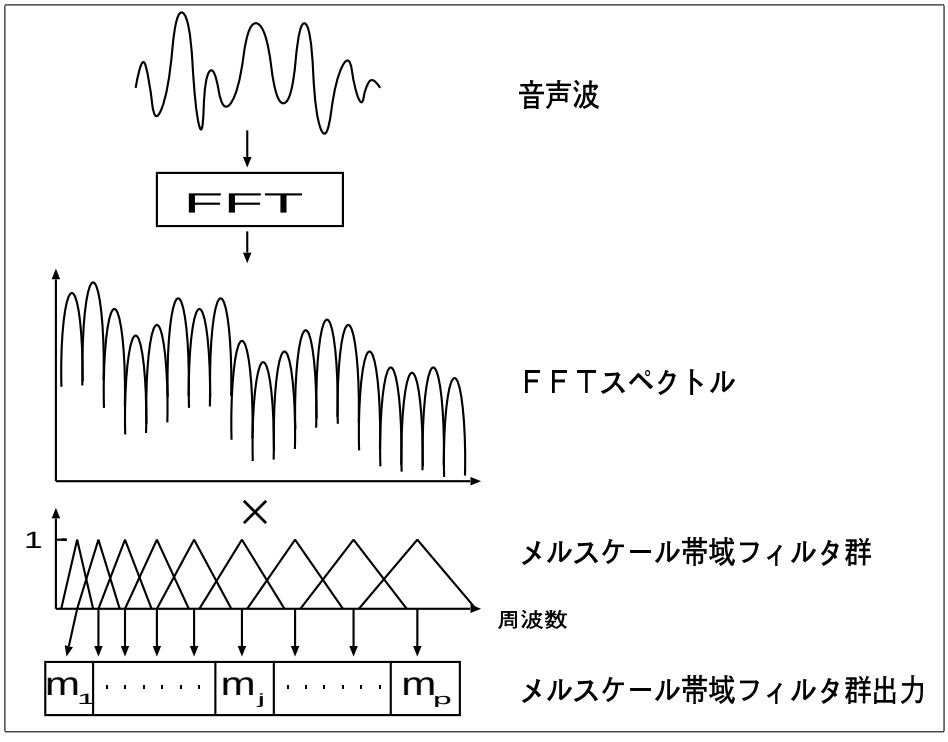

図 5: FFT に基づくメルスケール帯域フィルタ群分析の手順

メルスケールの帯域フィルタ群出力を対数変換し，逆フーリエ変換することによって求められたパラメータをメル周波数ケプストラム係数 (MFCC:Mel Frequency Cepstrum Coefficient) と呼び，音声認識の分野において特徴パラメータとして広く用いられている．

3.3.3 その他のパラメータ

短時間 (対数) パワーも重要な特徴であるが，声の平均的大きさには意味がないので，数百 ms〜数 s の平均パワーで正規化した値を用いる．

Δ(デルタ) ケプストラム (Δcepstrum) 法は，スペクトルの動的特徴 (スペクトル変化) が音声知覚において重要な役割を果たしていることに着目して考案された方法である．この方法の有効性は多くの実験によって確認され，広く用いられている．

また，対数パワーの変化率，すなわち Δ 対数パワーも広く用いられている [2].

3.3.4 音素モデル

音素モデルは，現在音声認識のための音響モデル単位として，最も広く用いられている．音素モデルは通常，音素環境によってコンテキストを依存させ，コンテキスト依存モデルとして用いられる．コンテキスト依存モデルとしては，前後の音素環境を考慮した triphone モデルを用いることにより，あらゆる音素の前後パターンの渡り部分をモデル化できる．しかし，triphone を含めたコンテキスト依存モデルでは，モデル数が膨大になることやモデルの構造が複雑になるため計算量が増加する．

以下，本論文では，コンテキストに依存していない音素モデルを，基本的な音素のみのモデルという意味で基本モデルと呼び，前後の音素環境を考慮した音素モデルを triphone モデルと呼ぶ．

3.3.5 単語辞書

サブワード単位を用いて音声認識を行なうためには，単語とサブワードとの対応関係を定義しておく必要がある．単語辞書とは，単語とサブワード系列との対応関係を定義したものである．

主なサブワード単位には，音素，音節，半音節などがある．

4　モーラ情報とピッチ周波数

4.1　モーラ

モーラとは，日本語の場合，仮名文字単位に相当し，音節とはやや異なっている [2]. 俳句や和歌で 5，7，5，7，7 などと数えるときの音の単位で，伸ばす音「ー」(長母音) や詰まる音「ッ」(促音)，跳ねる音「ン」(撥音) なども 1 モーラと考える [3]．また，単語のモーラの総数 (仮名文字の総数) を単語のモーラ数と呼び，単語のモーラの位置を単語のモーラ位置と呼ぶ．なお，本論文では，単語のモーラ数とモーラ位置をあわせて，モーラ情報と表現する．

単語「実験」におけるモーラ数とモーラ位置の例を表 1 に示す.「実験」は単語の仮名文字数が 4 であるため，モーラ数は 4 となる．モーラ位置は,「ジ」は 1,「ッ」は 2,「ケ」は 3,「ン」は 4 となる．

表 1: 単語「実験」におけるモーラ数とモーラ位置

単語	ジ	ッ	ケ	ン
ローマ字表記	z i	q	k e	n
モーラ数	4			
モーラ位置	1	2	3	4

4.2　モーラ情報とピッチ周波数の依存関係

特定話者の単語の発声において，単語のモーラ数およびモーラ位置が決まればピッチ周波数が，ほぼ決まることが知られている [10]．図 6-8 は参考文献 [10] から引用したもので，単一話者のナレータが発声した 4，5，6 モーラ語の地名 (固有名詞) のピッチ周波数の平均値と分散を示している．なお，このピッチ周波数の解析には，xwave+[20] を使用している．図の横軸は時間を示し，縦軸は周波数を示している．横軸の時間は，モーラ位置で正規化されている．図中の縦線がピッチ周波数の分散を示し，■はピッチ周波数の平均値を示している．

図 6-8 より，ピッチ周波数は単語に関係なく単語のモーラ数およびモーラ位置で決定できることがわかる．また，固有名詞だけでなく，普通名詞においても図 6-8 と同様の傾向があり，ピッチ周波数は単語のモーラ数およびモーラ位置から，ある程度決定できることが報告されている [16].

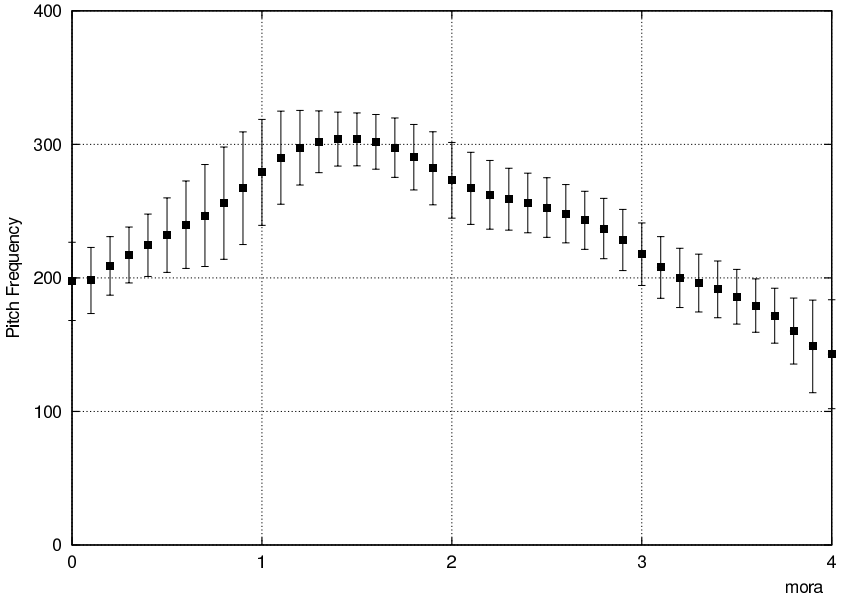

図 6: 4 モーラ語 1,500 件のピッチ周波数平均値と分散

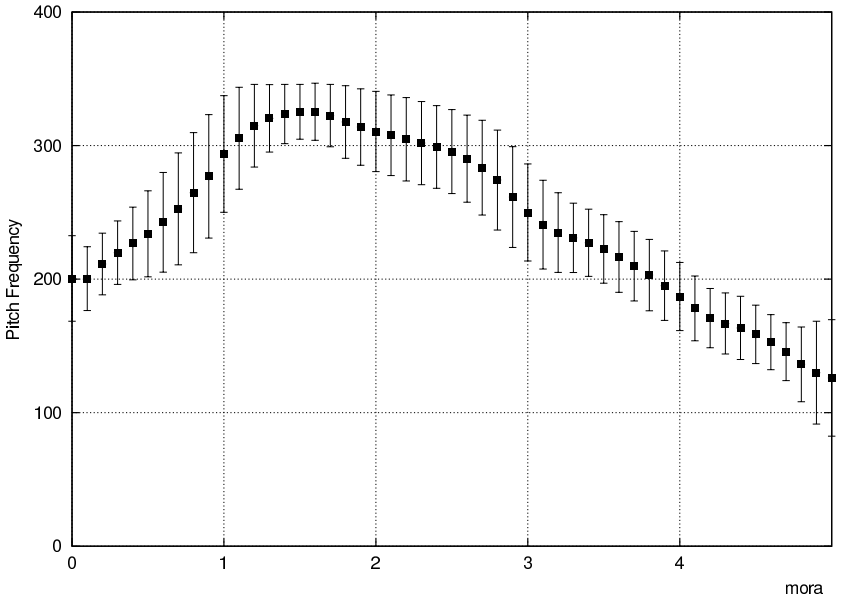

図 7: 5 モーラ語 2,800 件のピッチ周波数平均値と分散

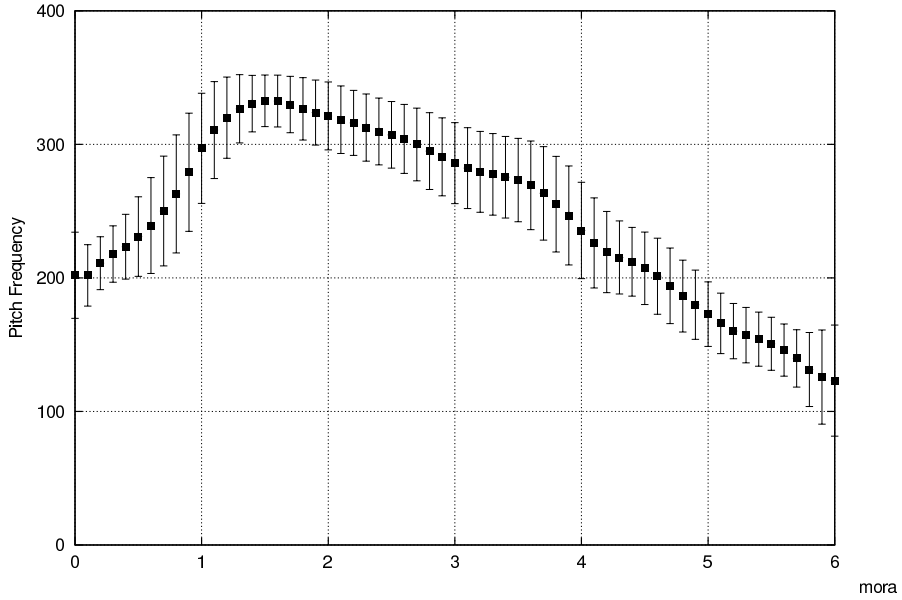

図 8: 6 モーラ語 2,200 件のピッチ周波数平均値と分散

4.3 モーラ情報を用いた音素モデルの提案

ケプストラム分析において，音源ピッチが低く，フォルマントが相互によく分離しているときはよい結果を得られるが，音源ピッチが高く，フォルマントが互いに近接しているときは両成分の分離が不完全となり，誤差が生ずることは前述した (2.2 節参照)．すなわち，フォルマントはピッチ周波数によって影響を受けやすく，精度の高い音素モデルを構築するためには，ピッチ周波数を考慮することも必要であると考えられる．しかしながら，安定したピッチ周波数の抽出は困難である (2.4 節参照)．

そこで，モーラ情報とピッチ周波数の依存関係に着目し，ピッチ周波数を直接考慮して音素モデルを構築するのではなく，単語のモーラ数とモーラ位置を考慮して音素モデルを構築する．これによって，より精度の高い音素モデルを構築することができると考えられる．以下，本論文では，単語のモーラ数およびモーラ位置を考慮した音素モデルをモーラモデルと呼ぶ．

4.4 語頭語尾情報を用いた音素モデル

従来の音声認識において，特に単語音声認識では，単語の語頭と語尾を考慮しただけでも認識率が向上することが知られている．

声帯振動が語頭や語尾において完全な周期性を持たないため，ケプストラム分析では，語頭や語尾において，フォルマントとピッチの分離が不完全となる．このため，同じ音素でも，語頭，語尾，それ以外 (語中) では，ケプストラムのパラメータは大きく異なる．

そこで，語頭，語尾，語中をそれぞれ分けることによって，より精度の高い音素モデルを構築することができ，これによって単語音声認識の認識率は向上する．

本論文では，単語の語頭と語尾を考慮した音素モデルを語頭語尾モデルと呼ぶ．

5　モーラ情報を用いた単語音声認識

5.1　音素ラベルの分類

学習および評価を行うデータには，音声波形データファイル (以下，波形ファイル) と音声ラベルファイル (以下，ラベルファイル) を含むデータベースを使用する．モーラ情報および語頭語尾情報を使用して分類する場合は，データベース中の全ラベルファイルの母音と撥音を分類する．これは，促音は無声音でありピッチは存在しないため，本研究では分類しない．同様に，子音もピッチの影響が小さいため，分類しない．triphone はすべての音素を，前後の音素環境を考慮して分類する．以下に，モーラ情報，語頭語尾情報，前後の音素環境を考慮したラベルの分類方法を述べる．また，具体的な分類例を表 2 に示す.

モーラ情報を用いた音素ラベルの分類

モーラ情報を使用したラベルファイルは母音と撥音の音素ラベルの前後に単語のモーラ数およびモーラ位置情報をそれぞれ付け加えることにより，母音と撥音を分類する．音声ラベルが apaato(アパート) である場合について，母音と撥音の分類例を示す．アパートという単語は，4 モーラ語であるため，分類する音素の前方に 4 をつけ，後方に各々のモーラ位置をつけて分類する．1 番目，3 番目，5 番目の音素 a は，4a1，4a2，4a3 という音素にそれぞれ置き換えられ，異なる音素ラベルとして扱う.

語頭語尾情報を用いた音素ラベルの分類

語頭語尾情報を使用したラベルファイルは母音と撥音の音素ラベルの後に，単語の語頭，語中および語尾情報をそれぞれ付け加えることにより分類する．ここで，1 モーラ語は語頭および語尾が存在しないので，語中モデルとして扱う．音声ラベルが apaato(アパート) である場合について，母音と発音の分類例を示す．1 番目の音素 a は aa に置き換えられ語頭モデルとして扱い，7 番目の音素 o は oz に置き換え語尾モデルとして扱い，それ以外の音素 a は語中モデルとして扱う.

triphone の音素ラベルの分類

triphone モデルのラベルファイルは，焦点となる音素の前に前音素を，後ろに後音素を-と+でつなぐことによって分類する．音声ラベルが apaato(アパート) である場合について，分類例を示す．1 番目の音素 a は前音素はなし，後音素は p なので a+p と置き換えられる．また，3 番目の音素 a は前音素は p，後音素は a なので p-a+a と置き換えられる．

表 2: 音素ラベルの分類例

・単語：アパート (音素列：a p a a t o)

基本音素	a	p	a	a	t	o
モーラ情報	4a1	p	4a2	4a3	t	4o4
語頭語尾情報	aa	p	a	a	t	oz
triphone	a+p	a-p+a	p-a+a	a-a+t	a-t+o	t-o

・単語：きまり (音素列：k i m a r i)

基本音素	k	i	m	a	r	i
モーラ情報	k	3i1	m	3a2	r	3i3
語頭語尾情報	k	ia	m	a	r	iz
triphone	k+i	k-i+m	i-m+a	m-a+r	a-r+i	r-i

・単語：実験 (音素列：j i q k e ng)

基本音素	zh	i	q	k	e	ng
モーラ情報	zh	4i1	q	k	4e3	4ng4
語頭語尾情報	zh	ia	q	k	e	ngz
triphone	zh+i	zh-i+q	i-q+k	q-k+e	k-e+ng	e-ng

5.2 音素 HMM の構築

5.2.1 音素 HMM の種類と初期モデル

本研究では，音素 HMM の作成にあたり，次の 2 点を考慮する．

1. 半連続分布 HMM の使用

   母音と撥音をモーラ情報や語頭語尾情報を使用して分類する場合，音素数の増加に伴い，作成される音素 HMM の数は増加する．これにより，総学習データ数が一定である場合，1 つあたりの音素 HMM の学習データが減少し，HMM パラメータの

信頼度が低下する．これを防ぐために本研究では，ガウス分布を全 HMM において共通にした半連続分布 HMM[7] を使用する．

2. 初期モデルの作成方法

HMM の学習は初期モデルが重要である．そこで，モーラ情報や語頭語尾情報を考慮したモデルや triphone モデルの初期モデルは基本的な音素 HMM を学習したものを，複製することによって，作成する．このため，モーラ情報や語頭語尾情報を考慮したモデルや triphone モデルの初期モデルは，同じ音素であれば基本的な音素 HMM と同じ出力確率の分布と遷移確率をもつ．

5.2.2 音素 HMM の作成手順

図 9 に音素 HMM の作成手順を示す．以下，基本モデル，モーラモデル，語頭語尾モデル，triphone モデルの作成方法について述べる．

A 基本モデルの作成

学習データに，従来使用されている基本的な音素ラベルをもつラベルファイルと波形ファイルを使用する．この学習データから Viterbi alignment を使用して初期モデルを作成する．この初期モデルを，Baum-Welch アルゴリズムを使用して再推定し，連結学習を行って連続分布の音素 HMM を作成する．

作成した連続分布の音素 HMM から，すべての音素 HMM の混合ガウス分布を共通にした半連続分布の音素 HMM の初期モデルを作成する．学習データを使用して，連結学習を行って，半連続分布の音素 HMM を作成する．

B モーラモデルの作成

A で作成した基本モデルのうち，母音と撥音の音素 HMM を複製することによって，モーラ情報を考慮した音素 HMM の初期モデルを作成する．そして，モーラ情報を用いた音素 HMM は，学習データにモーラ情報を使用したラベルファイルと波形ファイルを使用して，連結学習を行うことによって作成する．

C 語頭語尾モデルの作成

A で作成した基本モデルのうち，母音と撥音の音素 HMM を複製することによって，語頭語尾情報を考慮した音素 HMM の初期モデルを作成する．そして，語頭語

尾情報を用いた音素 HMM は，学習データに語頭語尾情報を使用したラベルファイルと波形ファイルを使用して，連結学習を行うことによって作成する．

D triphone モデルの作成

triphone モデルの初期モデルは，A で作成した基本モデルを複製することによって作成する．そして triphone モデルは，学習データに triphone ラベルファイルと波形ファイルを使用して，連結学習を行なうことによって作成する．

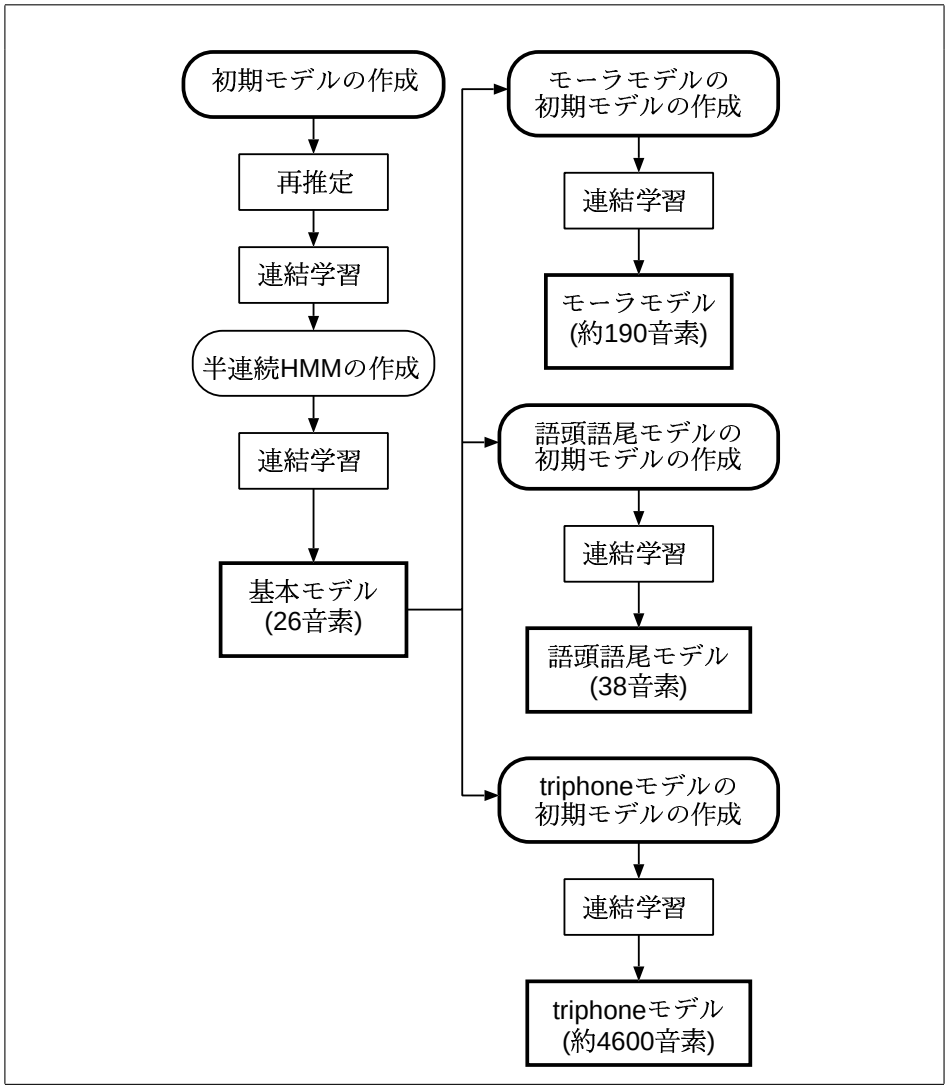

図 9: 音素 HMM の作成手順

5.3 評価実験

認識ツール

単語音声認識を行うツールとして，HTK Ver3.0[19] を使用する．

音声データ

本手法の有効性を評価するために用いた音声データは，ATR の単語発話データベース Aset である．このデータベースには，20 話者 (男性 10 話者，女性 10 話者) が読み上げた音声データが収録されている．話者ごとに 1 モーラから 7 モーラまでの使用頻度の高い単語，5240 単語の音声波形データが含まれている．また，この音声波形データには，人手によって付与された音素境界位置情報を付与してある．

実験には，このデータベースを奇数番と偶数番に分け，奇数を学習データ，偶数を評価データとして使用する．実験に用いた音声データの詳細を表 3 に示す．

表 3: ATR 単語発話データベース Aset

データベース	ATR 単語発話データベース Aset	
話者	6 話者 (男性 3 話者，女性 3 話者)	
学習データ	単語数	約 2620 単語/話者
	音素数	約 15500
	母音数	約 8000
評価データ	単語数	約 2620 単語/話者
	音素数	約 15500
	母音数	約 8000

実験に使用する単語辞書

認識時に必要である単語辞書は，評価データの音素ラベルによって作成する．単語辞書には，評価データの発話単語と対応する音素列が示されている．単語辞書に登録されている単語数は，評価データ数とほぼ一致する．

音響分析条件

特徴パラメータには，帯域フィルタバンク分析によって抽出したメルケプストラム (以下，MFCC とする．) を使用する．帯域フィルタバンクの分析次数は 32 次で，MFCC の次数は 16 次である．

特徴ベクトルは，MFCC(16 次) の他に，ΔMFCC(MFCC の一次差分，16 次)，対数パワー (1 次)，Δ 対数パワー (対数パワーの一次差分，1 次) の計 34 次とする．

HMM の混合ガウス分布の共分散行列には，Diagonal-covariance(以下，Diagonal) と Full-covariance(以下，Full) の 2 種類を使用し実験を行う．連続型 HMM の混合分布数は，母音，撥音，無音を 10mixture とし，その他の音素を 4mixture とする．半連続型 HMM の混合分布数は，いずれの場合も 256mixture とする．

表 4: 音響分析条件

標本周波数	16kHz		
分析窓	Hamming 窓		
分析窓長	20ms		
フレーム周期	5ms		
特徴ベクトル	16 次 MFCC+16 次 ΔMFCC+ 対数パワー+Δ 対数パワー (計 34 次)		
分析次数	32 次		
音響モデル	３ループ４状態 半連続分布型		
共分散行列	Diagonal-covariance，Full-covariance		
連続型 HMM の 混合分布数	Diagonl	母音，撥音，無音	10mixture
		その他の音素	4mixture
	Full	母音，撥音，無音	10mixture
		その他の音素	4mixture
半連続型 HMM の 混合分布数	256mixture		

5.4 実験結果

表 5 に Diagonal の HMM を用いた単語音声認識の実験結果を，表 6 に Full の HMM を用いた単語音声認識の実験結果を示す．表中の上段は誤り率を示し，下段の括弧内は誤

り数を示している．実験の結果から，モーラ情報を用いることによって，認識率は基本モデルよりも向上した．以下，Diagonal と Full の HMM それぞれの場合について，誤り率および改善率の結果について述べる．

表 5: 単語音声認識結果 (Diagonal-covariance)

話者	誤り率			
	基本モデル	モーラモデル	語頭語尾モデル	triphone モデル
mau	3.74% (98)	3.09% (81)	3.70% (97)	1.87% (49)
mmy	6.41% (168)	4.77% (125)	5.61% (147)	3.74% (98)
mnm	5.73% (150)	4.35% (114)	5.42% (142)	3.32% (87)
faf	5.00% (131)	4.08% (107)	5.27% (138)	5.38% (141)
fms	5.53% (145)	4.62% (121)	3.78% (99)	4.55% (119)
ftk	5.04% (132)	4.20% (110)	4.89% (128)	3.55% (93)
平均	5.24%	4.19%	4.78%	3.73%

表 6: 単語音声認識結果 (Full-covariance)

話者	誤り率			
	基本モデル	モーラモデル	語頭語尾モデル	triphone モデル
mau	3.78% (99)	1.91% (50)	2.37% (62)	1.49% (39)
mmy	3.02% (79)	2.60% (68)	2.79% (73)	2.25% (59)
mnm	3.24% (85)	2.71% (71)	3.32% (87)	2.48% (65)
faf	2.25% (59)	1.60% (42)	2.29% (60)	2.52% (66)
fms	3.13% (82)	2.10% (55)	2.60% (68)	2.29% (60)
ftk	2.82% (74)	2.29% (60)	2.48% (65)	2.29% (60)
平均	3.04%	2.20%	2.64%	2.22%

5.4.1 Diagonal の HMM における誤り率・改善率

Diagonal の HMM を用いた単語音声認識では，基本モデルと比較して，モーラ情報を使用することによって，誤り率は実験で使用したデータベース 6 話者平均で 1.05%減少した．また，改善率は 20.1%であった．Diagonal で一番良い結果は triphone モデルであった．triphone モデルの誤り率は，基本モデルと比較して 1.51%減少し，改善率は 28.8%であった．

5.4.2 Full の HMM における誤り率・改善率

Full の HMM を用いた単語音声認識では，基本モデルと比較して，モーラ情報を使用することによって，誤り率は実験で使用したデータベース 6 話者平均で 0.84%減少した．また，改善率は 27.6%であった．Full で一番良い結果はモーラモデルであった．triphone モデルと比較すると，誤り率の差は非常に小さく，モーラモデルと triphone モデルは同等の結果であった．

6　考察

6.1　男性話者と女性話者の比較

男性話者と女性話者の認識率について考察する．表 7 は，各モデルの男性 3 話者の平均の認識率と，女性 3 話者の平均の認識率を，Diagonal と Full について示している．

表 7: 男性話者と女性話者の平均認識率

話者	HMM	基本モデル	モーラモデル	語頭語尾モデル	triphone モデル
男性	Diagonal	5.29%	4.07%	4.91%	2.98%
	Full	3.35%	2.40%	2.82%	2.07%
女性	Diagonal	5.19%	4.30%	4.64%	4.49%
	Full	2.74%	2.00%	2.46%	2.37%

表 7 から，Diagonal と Full とも，男性話者の認識率は，triphone モデルが一番良いのに対し，女性話者の認識率は，モーラモデルが一番良かった．

この理由として，フォルマントにおけるピッチの影響は，女性話者の方が男性話者より大きいことが知られる．このため，モーラ情報を考慮することによって，ピッチの影響を分離できたからであると考えている．

6.2　認識単語の解析

6.2.1　基本モデルとの比較

基本モデルの認識結果と比較すると，モーラ情報を使用することによって認識できるようになった単語の多くは，連続母音を含む単語であった．特に，長母音を含む単語を短母音に誤認識してしまう単語に効果が見られた．この理由として，長母音をモーラ位置で異なる音素として認識することができるようになったことが大きな要因であると考えられる．表 8 に改善された具体的な単語例を示す．

また，モーラ情報を使用しても改善されなかった単語の例を表 9 に示す．改善されなかった単語の多くは，子音の誤認識であった．本研究では，母音と撥音のみにモーラ情報を使用し，子音は従来使用されているものと同じであるため，これらの単語は改善されなかったと考えられる．

表 8: 改善された単語例

改善された単語			誤認識していた単語	
単語	音素列		単語	音素列
交通	k o o ts u u	→	こつ	k o ts u
仕入れる	sh i i r e r u	→	知れる	sh i r e r u
招待	sh o o t a i	→	所帯	sh o t a i
葬式	s o o sh i k i	→	組織	s o sh i k i
報道	h o o d o o	→	歩道	h o d o o
誘拐	y u u k a i	→	愉快	y u k a i

表 9: 改善されなかった単語例

正しい単語			誤認識した単語	
単語	音素列		単語	音素列
一生	i q sh y o o	→	衣装	i sh y o o
応じる	o u zh i r u	→	閉じる	t o zh i r u
回覧	k a i r a ng	→	階段	k a i d a ng
行政	g y o o s e i	→	強制	k y o o s e i
資料	sh i r y o o	→	市場	sh i zh y o o
逮捕	t a i h o	→	太鼓	t a i k o
光り	h i k a r i	→	二人	h u t a r i
労働	r o o d o o	→	堂々	d o o d o o

6.2.2 triphone モデルとの比較

表 7 の結果から，モーラモデルと triphone モデルの認識結果は，男性話者と女性話者で大きく異なる．そこで，男性話者と女性話者に分けて考察する．

男性話者の認識結果は，モーラモデルより triphone モデルの方が良かった．そこで，男性話者の認識結果について，triphone モデルと比較した場合のモーラモデルの誤り単語について考察する．表 10 に，男性話者のモーラモデルの誤り単語の例を示す．モーラモデルの男性話者の認識結果では，子音による誤りが多く見られた．このため，男性話者では triphone の方が良い結果となっていた．この理由は，モーラ情報では調音結合の影響を考慮していないためであると考えている．

また，女性話者の認識結果は，モーラモデルの方が triphone モデルより良かった．そこで，女性話者の認識結果について，モーラモデルと比較した場合の triphone モデルの

表 10: モーラモデルの誤り単語例 (男性話者)

正しい単語			誤認識した単語	
単語	音素列		単語	音素列
いっそ	i q s o	→	基礎	k i s o
音	o t o	→	こと	k o t o
記録	k i r o k u	→	気力	k i r y o k u
審議	sh i ng g i	→	心理	sh i ng r i
受験	zh u k e ng	→	実験	zh i q k e ng
適応	t e k i o o	→	提供	t e i k y o o
プラス	p u r a s u	→	クラス	k u r a s u
ペン	p e ng	→	へん	h e ng

誤り単語について考察する．表 11 に，女性話者の triphone モデルの誤り単語の例を示す．triphone モデルの女性話者の認識結果では，子音による置換誤りと短母音を長母音に誤認識する誤りが多く見られた．このため，女性話者では triphone モデルの認識率は低下していた．子音の置換誤りは，音素モデルの増加によって学習データが不足したことが原因であると考えられる．また，母音の誤りは，6.1 節で述べたピッチの影響によるものと考えている．

表 11: Triphone モデルの誤り単語例 (女性話者)

正しい単語			誤認識した単語	
単語	音素列		単語	音素列
一杯	i q p a i	→	一体	i q t a i
奇麗	k i r e i	→	規定	k i t e i
系統	k e i t o o	→	傾向	k e i k o o
芸術	g e i zh y u ts u	→	現実	g e ng zh i ts u
幸運	k o o u ng	→	公園	k o o e ng
散歩	s a ng p o	→	参考	s a ng k o o
芝居	sh i b a i	→	市街	sh i g a i
電報	d e ng p o o	→	電灯	d e ng t o o
不要	h u y o o	→	不良	h u r y o o
良い	y o i	→	用意	y o o i

6.3　各モーラ数の認識率

1 モーラ語や 2 モーラ語は,「柿」や「牡蛎」などのように，アクセント位置によって意味が異なるものが多く存在する．このため，1 モーラ語や 2 モーラ語は，4 節で説明したピッチ周波数とモーラ情報の関係が成り立つとは限らない．そこで，各モーラ語の認識率について考察を行なった．表 12 は各モーラ語における 6 話者の平均の認識率を，モデルごとに示している．

表 12: 各モーラ語の認識率

モーラ数	基本モデル	モーラモデル	語頭語尾モデル	triphone モデル
1 モーラ語	5.8%	6.4%	51.9%	11.5%
2 モーラ語	4.59%	3.70%	4.67%	5.85%
3 モーラ語	3.37%	2.99%	2.14%	2.11%
4 モーラ語	2.45%	1.00%	1.44%	0.99%
5 モーラ語	0%	0.5%	0%	0%
6 モーラ語	0%	2.2%	0%	0%
7 モーラ語	0%	0%	0%	0%

表 12 から，基本モデルの認識率と比較して，モーラ情報の効果は，特に，2，3，4 モーラ語に見られた．3 モーラ語については，語頭語尾モデルと triphone モデルの認識率の方がモーラモデルの認識率よりも良い結果となっていた．

またモーラモデルの 5 モーラ語以上の認識率が低下していた．モーラ数が増加すれば，音素 HMM の数も増加し，一つあたりの音素 HMM の学習データ数が減少する．このため，HMM の信頼性が低下してしてしまうことが原因であると考えている．

6.4　モーラ数と語頭語尾情報を考慮した音素モデル

6.4.1　モーラモデルと語頭語尾モデルの中間モデルの検討

表 12 の結果から，モーラモデルと語頭語尾モデルの誤り率は，3 モーラ語以上の単語では，語頭語尾モデルの方が良かった．また，1，2，4 モーラ語の誤り率はモーラモデルの方が良かった．

モーラ情報は，モーラ位置を考慮しているため，語頭語尾情報も考慮していると考えることができる．そこで，本節で検討する中間モデルは，モーラモデルのように，モーラ位置およびモーラ数を考慮して作成するのではなく，モーラ数と語頭語尾を考慮して

作成する．以下，この中間モデルをモーラ語頭語尾モデルとし，このモデルを使用して単語音声認識を行なう．実験結果について，他のモデルの認識結果と比較し，モーラ語頭語尾モデルの有効性を考察する．

6.4.2 認識実験の結果と考察

単語音声認識実験は 5.3 節で示した条件下で行なった．実験の結果を表 13 に示す．

表 13: モーラ数と語頭語尾を考慮した認識結果

話者	Diagonal	Full
mau	3.05% (80)	1.76% (46)
mmy	4.69% (123)	2.40% (63)
mnm	4.35% (114)	2.71% (71)
faf	4.05% (106)	1.45% (38)
fms	4.50% (118)	2.14% (56)
ftk	3.93% (104)	2.18% (57)
平均	4.10%	2.11%

実験の結果から，モーラ数と語頭語尾を考慮することにより，Diagonal の HMM を用いた場合，データベース 6 話者の平均で，誤り率は 4.10%であり，改善率は 21.7%であった．Full の HMM を用いた場合，データベース 6 話者の平均で，誤り率は 2.11%であった．また，改善率は 30.8%であった．

この結果から，モーラ語頭語尾モデルの認識実験では，モーラモデルの認識結果と比較して，やや良い結果が得られた．

6.5 triphone におけるモーラ情報の有効性

モーラ情報の有効性については前述した．そこで，従来有効な手法である triphone にモーラ情報を用いた場合，認識率は向上すると考えられる．ここでは，triphone モデルにモーラ情報を考慮した場合について考察を行なった．

6.5.1 triphone モーラモデルの作成

triphone にモーラ情報を考慮したモデルの初期モデルは，5.2.2 節で作成したモーラモデルを複製することによって作成する．そして，triphone にモーラ情報を考慮したモデ

表 14: triphone モーラモデルの認識結果

話者	Diagonal	Full
mau	2.90% (76)	1.64% (43)
mmy	4.89% (128)	2.67% (70)
mnm	4.66% (122)	2.40% (63)
faf	5.73% (150)	4.66% (122)
fms	5.15% (135)	3.36% (88)
ftk	3.89% (102)	2.21% (58)
平均	4.54%	2.82%

ルは，学習データに，triphone ラベルにモーラ情報を考慮したラベルファイルと波形ファイルを使用して，連結学習を行なうことによって作成する．以下，この音素 HMM を，triphone モーラモデルとする．

6.5.2 認識実験の結果と考察

単語音声認識は 5.3 節で示した条件下で行なった．実験の結果を表 14 に示す．

表 14 から，triphone モーラモデルの誤り率は，triphone モデルの誤り率より，悪くなった．これは，前後の音素環境にモーラ情報も考慮に入れることで，さらに音素 HMM の数が約 6800 種類に増加する．その結果，音素 HMM の精度と信頼度は，ともに低下する，いわゆる，過学習が引き起こされる．このため，triphone にモーラ情報を考慮すると認識率が低下したと考えられる．

6.6 不特定話者認識おけるモーラ情報の有効性

不特定話者に対する音声認識は，認識システムの実用化において重要である．そこで，不特定話者単語音声認識におけるモーラ情報の有効性について考察を行なった．

6.6.1 実験条件

音響分析条件は基本的に表 4 に示した条件と同様である．また，音響分析条件の変更点を表 15 に示す．認識データは mau を使用し，学習データには，mau 以外の男性話者 9 名を使用する．

音声認識の音素モデルには，基本モデル，モーラモデル，triphone モデルを用いる．また，学習データが増加するため，過学習が起こりにくいと考え，triphone モーラモデルも実験に用いる．

表 15: 音響分析条件の変更点

共分散行列	Diagonal-covariance
連続型 HMM の混合分布数	すべての音素：10mixture
半連続型 HMM の混合分布数	1024mixture

6.6.2 結果と考察

実験結果を表 16 に示す．

表 16: 不特定話者単語音声認識の実験結果

話者	基本モデル	モーラモデル	triphone モデル	モーラ tri モデル
mau	12.41% (650)	9.76% (511)	7.39% (387)	6.87%(360)

表 16 から，基本モデルと比較して，モーラモデルの認識率は向上し，不特定話者認識においても，その有効性が認められた．しかしながら，認識率は，triphone モデルが一番良い結果であった．この理由として，音韻間の調音の影響は，話者によって大きく異なることが知られる．triphone は，前後の音韻環境を考慮しているため，この影響を小さくできたことが原因であると考えている．

また，triphone モーラモデルの結果が triphone モデルよりさらに良かった．特定話者認識では，学習データが不足して過学習を起こしたため，認識率の低下につながっていた．しかし，今回の不特定話者認識では，学習データが増加するため，過学習が起こらなかったことが原因で，認識率が向上したと考えている．この結果は，学習データが豊富であれば，単語音声認識においてモーラ情報は有効であることを示している．

6.7 音素モデル数と学習データ数

音声認識において，音響モデルの高精度化は重要である．音響モデルの高精度化に関しては，これまでに多くの研究報告がなされている [8].

HMM は統計的なモデルであるため，学習データ量とモデルの自由度との間に密接な関係がある．統計的に信頼性の高いパラメータ推定を行なうためには，学習データ量に見あったモデルの自由度を設定する必要がある．

表 17: 各モデルのモデル数と認識結果 (WER)

HMM	モデル数	誤り率	
		Diagonal	Full
基本モデル	26	5.24%	3.04%
語頭語尾モデル	38	4.87%	2.64%
モーラ語頭語尾モデル	約 140	4.10%	2.11%
モーラモデル	約 190	4.19%	2.20%
triphone モデル	約 4,600	3.73%	2.22%
モーラ tri モデル	約 6,800	4.54%	2.82%

表 17 は，本論文で用いた音素モデルのモデル数と平均誤り率を示している．この表から，ただ単にモデルを細分化すればよいというわけではないことは明らかである．例えば，モーラ情報の有効性については前述したが，triphone にモーラ情報を組み合わせた場合の認識率は，triphone モデルよりも悪くなる．これは，モデルのパラメータ数が増加したため，各パラメータの推定に使われるデータ量が細分化され，推定精度の低下 (いわゆる “過学習”) が起こったためである．

統計的に信頼度が低いパラメータは，学習セットとわずかに異なるデータに対しても誤認識を起こすようになる (頑健性の低下)．また，逆にモデルの自由度が小さく詳細化が十分でないと，満足な性能が得られない．

音声認識においては，学習データ量に見合ったモデルの構造 (または自由度) を設計する手段と，モデルの頑健性を評価する基準が重要になる．

7　まとめと今後の課題

7.1　まとめ

本論文では，ピッチパターンを単語のモーラ数およびモーラ位置で記述できると仮定し，単語のモーラ数およびモーラ位置を考慮した音素モデルを学習し，単語音声認識を行なった．また，どの程度有効であるか比較するために，従来有効な手法とされる，語頭語尾モデルや triphone モデルも学習し，単語音声認識を行なった．その結果，モーラ情報を用いることによって，Diagonal-covariance の HMM を用いた場合は，誤り率は 6 話者の平均で 1.05%減少し，Full-covariance の HMM を用いた場合は，誤り率は 6 話者の平均で 0.84%減少した．triphone モデルと比較した場合，Full-covariance では，同等の結果が得られた．また，男性話者と女性話者との結果の比較では，モーラモデルは，女性話者において，他のモデルよりも誤り率が削減され，モーラモデルは，女性話者に効果があることが認められた．認識単語の解析では，連続母音を持つ単語誤りに効果が見られ，特に，長母音を短母音に誤認識する単語に効果が見られた．

以上の結果からモーラ情報を考慮することによって認識率は向上し，本手法の有効性が示された．

7.2　今後の課題

今後の課題として，不特定話者認識におけるモーラ情報の有効性や，雑音環境下におけるモーラ情報の有効性について考察する必要があると思われる．

また，音節・半音節をサブワード単位とした場合のモーラ情報の有効性，単語のアクセント情報とモーラ情報を組み合わせた場合の有効性などについて検討していく必要があると思われる．

謝辞

最後に，三年間の長期に渡って御指導，御教授して頂きました鳥取大学工学部知能情報工学科計算機 C 研究室の池原教授と村上助教授に深くお礼申し上げます．また，論文を執筆するにあたり，助言を頂いた徳久助手にお礼を申し上げます．加えて，本稿を執筆するにあたり参考させて頂いた論文，本の著者の方々の皆様にもお礼申し上げます．

参考文献

[1] 中川 聖一，“確率モデルによる音声認識” 電子情報通信学会，(1988)

[2] 古井 貞熙，“音声情報処理” 森北出版，(1998)

[3] 中田 和男，“改訂 音声” コロナ社，(1995)

[4] 今井 聖，“音声認識” 共立出版，(1995)

[5] 鹿野 清宏，中村 哲，伊勢 史郎，“音声・音情報のディジタル信号処理” 昭晃堂，(1997)

[6] 北 研二，中村 哲，永田 昌明，“音声言語処理-コーパスに基づくアプローチ-” 森北出版，(1996)

[7] X.D. HUANG，Y. ARIKI，M.A. JACK，“HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION” Edinburgh University Press，Edinburgh，(1990)

[8] 高橋 敏，嵯峨山 茂樹，“音声認識のための音響モデルの構造” 日本音響学会講演論文集，1-6-10，pp.19-22(1999-3)

[9] 田中 信一，正井 康之，松浦 博，新田 恒雄，“単語スポッティングに適した語頭・語尾モデルの検討” 日本音響学会講演論文集，1-3-7，pp.33-34(1996-9)

[10] 水澤 紀子，村上 仁一，東田 正信，“音節波形接続による単語音声合成” 電子情報通信学会技術研究報告，SP99-2(1999-05)

[11] 前田 智広，村上 仁一，池原 悟，“モーラ情報を用いた音素ラベリング方式の検討” 電子情報通信学会技術研究報告，SP2001-53，pp.25-30(2001-8)

[12] 妹尾 貴宏，村上 仁一，前田 智広，池原 悟，“モーラ情報を用いた単語音声認識の研究” 電子情報通信学会技術研究報告，SP2001-45，pp.1-5(2001-8)

[13] 妹尾 貴宏，村上 仁一，池原 悟，“モーラ情報を用いた単語音声認識の検討” 電子情報通信学会技術研究報告，SP2002-130，pp.55-61(2002-12)

[14] 妹尾 貴宏，村上 仁一，池原 悟，“モーラ情報を用いた単語音声認識の検討” 日本音響学会 2003 年春季研究発表会発表予定，1-4-8，pp.15-16(2003-3)

[15] 妹尾 貴宏，村上 仁一，池原 悟，“モーラ情報を用いた単語音声認識” 電子情報通信学会，論文誌投稿中 (2003-2)

[16] 石田 隆浩，村上 仁一，池原 悟，“音節波形接続型音声合成の普通名詞への応用” 電子情報通信学会技術研究報告，SP2002-25，pp.7-12(2002-05)

[17] 緒方 淳，有木 康夫，“日本語話し言葉音声認識のための音節に基づく高精度な音響モデルの検討” 電子情報通信学会技術研究報告，SP2002-52，pp.49-54(2002-12)

[18] 村上 仁一，“確率的言語モデルによる自由発話認識に関する研究” 豊橋技術大学博士論文 (1996)

[19] HTK Ver2.2 reference manual，1997 Cambridge University

[20] Introducing ESPS/waves+ with EnSigTM, Entropic Research Laboratory, Inc.