社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

# モーラ情報を用いた単語音声認識の検討

妹尾　貴宏†　　村上　仁一†　　池原　　悟†

† 鳥取大学工学部　〒 680–0945 鳥取県鳥取市湖山町南 4–101
E-mail: †{seo,murakami,ikehara}@ike.tottori-u.ac.jp

**あらまし**　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，語頭語尾を考慮したモデルや Triphone モデルの認識結果よりも高い，という結果が得られた．その結果，モーラ情報の有効性が認められた．

**キーワード**　単語音声認識，モーラ数，モーラ位置，語頭語尾，Triphone

# Isolated-Word Speech Recognition using Mora Position and Mora Length

Takahiro SEO†, Jin'ichi MURAKAMI†, and Satoru IKEHARA†

† Faculty of Engineering, Tottori university　Koyama-cho 4–101, Tottori, 680–0945 Japan
E-mail: †{seo,murakami,ikehara}@ike.tottori-u.ac.jp

**Abstract**　Recently, in word speech recognition, cepstram is normally used as parameters. And, pitch information is not normally used because it is difficult to extract pitch frequency steady. However, in recent reserches, it is known that there is a relation between pitch frequency and mora information. Then in this paper, we performed word speech recognition using mora information in stead of pitch and examined its validity. In the results, by using mora information, error rate of word speech recognition could decrease for many speeker. And comparing the results of word speech recognition using mora infomation with using begining and ending of a word or using triphone , the recognition rate using mora infomation was higher . And the efficiency of this way was confirmed.

**Key words**　Word speech recognition,Mora length,Mora position,Begining and ending of a word,Triphone

## 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 [1] や音素ラベリング [2] 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する．

## 2. モーラ情報とピッチ情報

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 [1]．図 1 は [1] から引用したもので，単一話者のナレータが発声した 5 モーラ語の地名 (固有名詞)2，800 件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+ [7] を使用している．

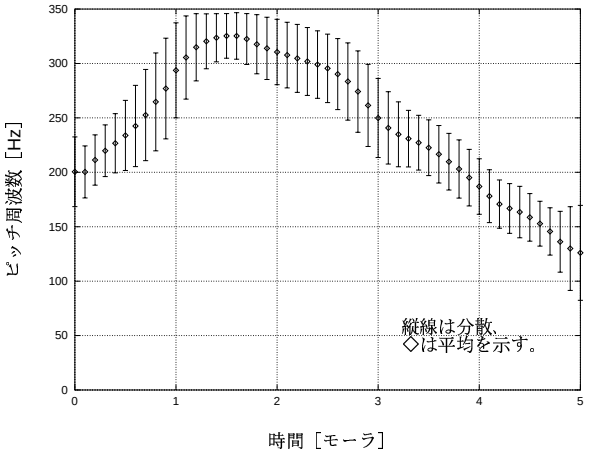


図 1　5 モーラ語 2，800 件のピッチ周波数平均値と分散
Fig. 1　Relation between mora information and pitch frequency

図 1 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，4，6 モーラ語も同様の傾向を示し，分散も 5 モーラ語と同程度であったと報告されている．

また，固有名詞だけでなく，普通名詞においても図 1 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 [4]．

一方，フォルマントはピッチ周波数によって影響を受けることが知られている．このことから，単語のモーラ数およびモーラ位置で分類して音素 HMM の学習を行うことで，フォルマントにおけるピッチ周波数の影響を分離できると考えられる．その結果，音素 HMM の精度が向上し，この音素 HMM を使用して単語音声認識を行った場合，単語音声認識の精度は向上すると推定できる．なお本研究では，単語のモーラ数およびモーラ位置をモーラ情報と定義し，モーラ情報を考慮した音素 HMM をモーラモデルと呼ぶ．

## 3. 語頭語尾モデルと Triphone モデル

単語音声認識において，単語の語頭の音節と語尾の音節を考慮して音素 HMM を作成し，単語認識を行った場合に，認識率が向上することが知られている．これは，同じ音素でも，語頭，語中，語尾において特徴パラメータが大きく異なるため，語頭，語中，語尾ごとに音素 HMM を作成することによって，単語音声認識の認識率は向上すると考えられる．以下，語頭語尾を考慮した音素 HMM を語頭語尾モデルと呼ぶ．

Triphone は，前後の音素環境を考慮した HMM モデルで，現在の音声認識で最も有効な手法として知られている．これは，焦点となる音素と前後の音素間の調音結合の影響を反映できるため，音声認識の認識率が向上する．

## 4. 単語音声認識実験

本手法の有効性を調べるために，基本的なモデル，モーラモデル，語頭語尾モデル，Triphone モデルをそれぞれ学習し，単語音声認識実験を行なう．

### 4.1 ラベルファイルの作成

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分類方法を述べる．また，具体的な分類例を表 1 に示す．

#### 4.1.1 モーラ情報を用いたラベルファイルの作成

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 akasaka(赤坂) である場合について，母音と発音の分類例を示す．赤坂という単語は 4 モーラ語であるため，1 番目，3 番目，5 番目，7 番目の音素 a は 4a1，4a2，4a3，4a4 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．

#### 4.1.2 語頭語尾情報を用いたラベルファイルの作成

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 akasaka(赤坂) 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 a は az 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．

#### 4.1.3 Triphone モデルのラベルファイルの作成

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 akasaka(赤坂) 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 k なので a+k と置き換えられる．また，3 番目の音素 a は前音素は k，後音素は s なので k-a+s と置き換えられる．

### 4.2 音素 HMM の学習

#### 4.2.1 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

表 1　モーラモデル、語頭語尾モデル、Triphone モデルの分類例
Table 1　Examples of labels

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 基本音素 | a | k | a | s | a | k | a |
| モーラ情報を使用 | 4a1 | k | 4a2 | s | 4a3 | k | 4a4 |
| 語頭語尾情報を使用 | aa | k | a | s | a | k | az |
| Triphone | a+k | a-k+a | k-a+s | a-s+a | s-a+k | a-k+a | k-a |

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

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

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 [5] を使用する．

（2） 初期モデルの作成方法

HMM 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．

### 4.3　音素 HMM の作成手順

#### 4.3.1　基本的な音素 HMM の作成

学習データに，従来使用されている基本的な 26 種類の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，26 種類の半連続分布の音素 HMM を作成する．以下，この基本的な音素 HMM を基本モデルとする．

#### 4.3.2　モーラモデルの作成

4.3.1 節で作成した基本モデルうち，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．以下，モーラ情報を用いた音素 HMM をモーラモデルとする．

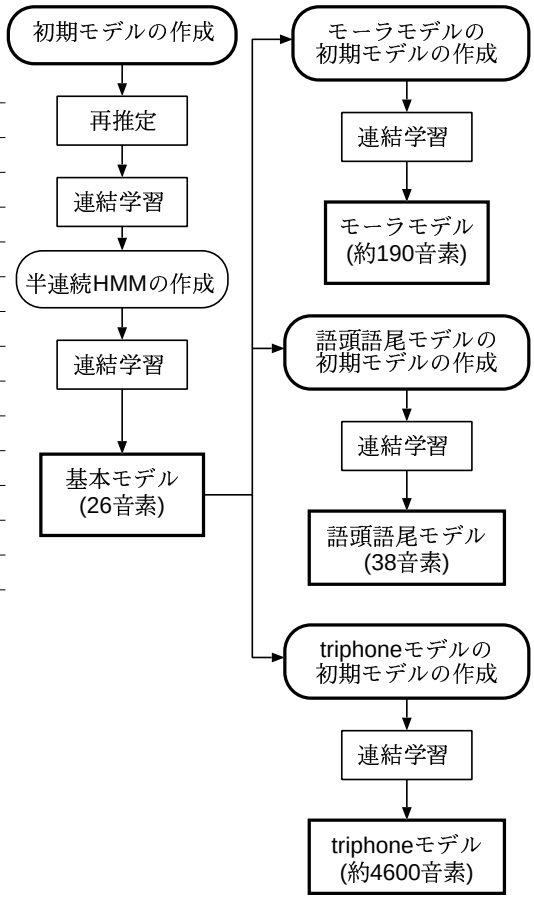


図 2　音素 HMM の作成手順
Fig. 2　The procedure creation of Pheneme HMM

#### 4.3.3　語頭語尾モデルの作成

4.3.1 節で作成した基本モデルの音素 HMM 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．以下，語頭語尾情報を用いた音素 HMM を語頭語尾モデルとする．

#### 4.3.4　Triphone モデルの作成

Triphone モデルの初期モデルは，4.3.1 節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．

基本モデル，モーラモデル，語頭語尾モデル，Triphone モデルに対し，それぞれ単語音声認識実験を行い，認識結果の比較を行なう．

### 4.4　実験条件

単語音声認識を行うツールとして，HTK [6] を使用する．音

響分析条件を表 2 に示す．また，音素 HMM 学習時および認識時の実験条件を表 3 に示す．

表 2　音響分析条件
Table 2　The conditions of the acoustic analysis

| | | | |
|---|---|---|---|
| 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) | | |
| 音響モデル | 3 ループ 4 状態 | | |
| | 半連続分布型 | | |
| 共分散行列 | Diagonal-covarience，Full-covarience | | |
| 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 | | |

表 3　実 験 条 件
Table 3　Experimental conditions

| | |
|---|---|
| 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 |

実験には，ATR の単語発話データベース Aset を使用する．このデータベースには，話者ごとに 1 モーラから 7 モーラまでの単語 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．

音素 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，Diagonal の HMM の場合，母音，撥音，無音を 10mixture とし，その他の音素を 4mixture とする．また，Full の HMM の場合は母音，撥音，無音を 10mixture とし，音素 b，g，w，z，zh を 2mixture とし，音素 p を 1mixture 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．

男性 3 話者 (mau，mmy，mnm) 女性 3 話者 (faf，fms，ftk) の計 6 話者に対し単語音声認識を行い，モーラ情報の有効性を調査する．

表 4　単語音声認識結果 (Diagonal-covarience)
Table 4　The results of the experiment of word speech recognition(Diagonal-covarience)

| 話者 | 認識単語総数 | 基本モデル | | 語頭語尾モデル | | モーラモデル | |
|---|---|---|---|---|---|---|---|
| | | 誤り数 | 誤り率 | 誤り数 | 誤り率 | 誤り数 | 誤り率 |
| mau | 2620 | 100 | 3.92% | 84 | 3.21% | 81 | 3.09% |
| mmy | 2620 | 153 | 5.84% | 144 | 5.50% | 131 | 5.00% |
| mnm | 2620 | 159 | 6.07% | 155 | 5.92% | 125 | 4.77% |
| faf | 2620 | 157 | 5.99% | 151 | 5.76% | 130 | 4.96% |
| fms | 2620 | 156 | 5.95% | 123 | 4.69% | 126 | 4.81% |
| ftk | 2620 | 132 | 5.04% | 119 | 4.54% | 102 | 3.89% |
| 平均 | | | 5.45% | | 4.94% | | 4.42% |

### 4.5 評 価 方 法

評価には (1) 式に示す誤り率を使用する．また，モーラ情報や語頭語尾情報などの情報を使用することによって改善された誤りの度合いとして，(2) 式に示す改善率を使用する．なお，誤認識した単語の総数を，単に，誤り数とする．

$$\text{誤り率} = \frac{\text{誤り数}}{\text{認識単語総数}} \times 100 \tag{1}$$

$$\text{改善率} = \frac{\text{誤り数}_{add\_info} - \text{誤り数}_{normal}}{\text{誤り数}_{normal}} \times 100 \tag{2}$$

ここで 誤り数$_{add_info}$ は，モーラ情報や語頭語尾情報を使用したときの誤り数を，誤り数$_{normal}$ は，基本モデルの誤り数を示す．

### 4.6 実 験 結 果

#### 4.6.1　モーラモデルと語頭語尾モデルの認識結果

表 4 に Diagonal の HMM を用いた単語音声認識の実験結果を，表 5 に Full の HMM を用いた単語音声認識の実験結果を示す．結果から，すべての話者において，モーラ情報や語頭語尾情報を使用することによって，基本モデルよりも，誤り率を減少させることができた．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について述べる．

（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の誤り率は，この実験で使用したデータベース 6 話者平均で，モーラ情報を使用することにより 5.45%から 4.42%に減少した．また，語頭語尾情報を使用した場合では，4.94%に減少した．改善率は，データベース 6 話者平均で，モーラ情報を使用することにより，18.9%の誤りが改善された．また，語頭語尾情報を使用した場合では，9.45%が改善された．

（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の誤り率は，この実験で使用したデータベース 6 話者平均で，モーラ情報を使用することにより 3.36%から 2.30%に減少した．語頭語尾情報を使用した場合は，3.02%に減少した．また，改善率は，データベース 6 話者平均で，モーラ情報を使用することにより，31.6%の誤りが改善された．語頭語尾情報を使用した場合は，10.2%が改善された．

表 5　単語音声認識結果 (Full-covarience)

Table 5　The results of the experiment of word speech recognition(Full-covarience)

| 話者 | 認識単語総数 | 基本モデル | | 語頭語尾モデル | | モーラモデル | |
|---|---|---|---|---|---|---|---|
| | | 誤り数 | 誤り率 | 誤り数 | 誤り率 | 誤り数 | 誤り率 |
| mau | 2620 | 98 | 3.74% | 72 | 2.75% | 55 | 2.10% |
| mmy | 2620 | 100 | 3.82% | 108 | 4.12% | 76 | 2.90% |
| mnm | 2620 | 89 | 3.40% | 89 | 3.40% | 65 | 2.48% |
| faf | 2620 | 75 | 2.86% | 63 | 2.40% | 56 | 2.14% |
| fms | 2620 | 71 | 2.71% | 54 | 2.06% | 46 | 1.76% |
| ftk | 2620 | 95 | 3.63% | 88 | 3.36% | 63 | 2.40% |
| 平均 | | | 3.36% | | 3.02% | | 2.30% |

表 6　Triphone における認識結果

Table 6　The results of word speech recognition using Triphone Model

| 話者 | 認識単語総数 | Diagonal | | Full | |
|---|---|---|---|---|---|
| | | 誤り数 | 誤り率 | 誤り数 | 誤り率 |
| mau | 2620 | 39 | 1.49% | 44 | 1.68% |
| mmy | 2620 | 89 | 3.40% | 65 | 2.48% |
| mnm | 2619 | 91 | 3.47% | 53 | 2.02% |
| faf | 2620 | 123 | 4.69% | 98 | 3.74% |
| fms | 2618 | 97 | 3.08% | 76 | 2.90% |
| ftk | 2619 | 80 | 3.05% | 69 | 2.63% |
| 平均 | | | 3.30% | | 2.57% |

#### 4.6.2　Triphone モデルの認識結果

Triphone を用いた単語音声認識実験の結果を表 6 に示す.

実験の結果から，Triphone モデルの誤り率は，データベース 6 話者平均で，Diagonal の HMM を用いた場合は 3.30%，Full の HMM を用いた場合は 2.57%であった．Triphone モデルを用いることによって，男性話者では効果がみられたが，女性話者ではあまり効果がみられなかった.

また，モーラモデルの認識率と比較した場合，データベース 6 話者平均の認識率では，Diagonal の HMM においては Triphone モデルが方が高く，Full の HMM においてはモーラモデルの方が高かった.

## 5.　考　　察

### 5.1　効果の見られた単語

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 7 に改善された具体的な単語例を示す.

### 5.2　効果の見られなかった単語

基本モデルの認識結果と比較すると，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 8 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 7　改善された単語例

Table 7　The improved examples of words

| 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 t s u u | → | こつ | k o t s 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 8　改善されなかった単語例

Table 8　The exam;les of words which have not been improved

| 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 |

### 5.3　モーラ数ごとに分類した認識結果

１モーラ語や２モーラ語は，柿や牡蛎などのように，アクセントの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モーラ語は図 1 に示されるような，ピッチ周波数とモーラ数およびモーラ位置の関係が成り立つとは限らない．そこで，３モーラ語以上の単語の誤り率について検討した．表 9 は各モーラ数における単語の誤り率を示している．なお，表の誤り率は認識実験で使用した 6 話者の平均値を示している.

表 9 から，モーラ情報や語頭語尾情報によって３モーラ語と４モーラ語の単語で誤り率の大きな減少が見られた．これは，３モーラ語以上の単語では，アクセント型によらず，ピッチパターンが安定しているためと考えられる．３モーラ以上の誤り率では，基本モデルが 4.66%，モーラモデルが 3.54%，語頭語尾モデルが 3.42%となり，語頭語尾モデルが一番よい認識率を示した.

また，語頭語尾モデルにおいて，１モーラ語の誤り率が増加し，全体の認識率の低下していた．これは，音素を語頭，語中，語尾で分類した場合に，１モーラ語の音素はどれにも属さないことが原因であると考えられる.

### 5.4　モーラ情報の有効性

モーラ情報の有効性を，基本モデル，語頭語尾モデルおよび Triphone モデルの実験結果と比較して考察する．表 10 は，基本モデル，語頭語尾モデル，モーラモデルおよび Triphone モデルのデータベース 6 話者平均の誤り率を示している.

表 9　各モーラ語の誤り率
Table 9　The error rate of each mora length of words

| モーラ数 | 単語総数 | 基本モデル | | 語頭語尾モデル | | モーラモデル | |
|---|---|---|---|---|---|---|---|
| | | 誤り数 | 誤り率 | 誤り数 | 誤り率 | 誤り数 | 誤り率 |
| 1 モーラ語 | 156 | 12 | 7.7% | 84 | 54% | 12 | 7.7% |
| 2 モーラ語 | 2460 | 204 | 8.29% | 222 | 9.02% | 196 | 8.29% |
| 3 モーラ語 | 5793 | 350 | 6.04% | 274 | 4.73% | 325 | 5.61% |
| 4 モーラ語 | 6577 | 291 | 4.42% | 195 | 2.96% | 157 | 2.39% |
| 5 モーラ語 | 584 | 0 | 0.0% | 1 | 0.17% | 5 | 0.86% |
| 6 モーラ語 | 138 | 0 | 0.0% | 0 | 0.0% | 0 | 0.0% |
| 7 モーラ語 | 12 | 0 | 0.0% | 0 | 0.0% | 0 | 0.0% |

表 10　各 HMM モデルにおける誤り率
Table 10　The error rate of the each HMM models

| HMM モデル | 誤り率 | |
|---|---|---|
| | Diagonal | Full |
| 基本モデル | 5.45% | 3.36% |
| 語頭語尾モデル | 4.94% | 3.02% |
| モーラモデル | 4.42% | 2.30% |
| Triphone モデル | 3.30% | 2.57% |

表 10 の実験結果から，今回の実験で一番よい結果を示したのは，Full の HMM を用いた場合のモーラモデルであった．これは，従来，有効であるといわれている Triphone モデルよりも，よい結果であった．このことから，モーラ情報は，単語音声認識において，Full の HMM を用いた場合，Triphone よりも有効であることが認められた．

また，モーラモデルの認識率は基本モデルと語頭語尾モデルの認識率よりも高くなり，モーラ情報の有効性が認められた．

### 5.5　モーラ語頭語尾モデル

#### 5.5.1　モーラモデルと語頭語尾モデルの中間モデルの検討

モーラ情報は，単語音声認識において，誤り率を減少させるのに有効な情報であることを示した．語頭語尾情報は，3 モーラ語以上の単語に対して，モーラ情報より有効であることを示してきた．そこで，二つの情報の利点を生かし，さらに誤り率を削減させるために，モーラモデルと語頭語尾モデルの中間的なモデルを検討する．

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，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い，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検討した．

#### 5.5.2　認識実験の結果と考察

単語音声認識実験は 4.4 節で示した条件下で行なった．実験の結果を表 11 に示す．

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 4.15%であり，改善率は 23.3%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 2.22%であり，改善率は 33.9%であった．

表 11　モーラ数と語頭語尾を考慮した認識結果
Table 11　The results of word speech recognition using mora length and begining and ending of a word

| 話者 | 単語総数 | Diagonal | | Full | |
|---|---|---|---|---|---|
| | | 誤り数 | 誤り率 | 誤り数 | 誤り率 |
| mau | 2620 | 75 | 2.86% | 53 | 2.02% |
| mmy | 2620 | 130 | 4.96% | 76 | 2.90% |
| mnm | 2620 | 124 | 4.73% | 63 | 2.40% |
| faf | 2620 | 129 | 4.92% | 50 | 1.91% |
| fms | 2620 | 113 | 4.31% | 43 | 1.64% |
| ftk | 2620 | 94 | 3.59% | 64 | 2.44% |
| 平均 | | | 4.23% | | 2.22% |

表 12　Triphone モーラモデルの認識結果
Table 12　The results of word speech recognition using Triphone Model and Mora position and length

| 話者 | 認識単語総数 | Diagonal | | Full | |
|---|---|---|---|---|---|
| | | 誤り数 | 誤り率 | 誤り数 | 誤り率 |
| mau | 2620 | 63 | 2.40% | 64 | 2.44% |
| mmy | 2620 | 118 | 4.50% | 74 | 2.82% |
| mnm | 2619 | 110 | 4.20% | 65 | 2.48% |
| faf | 2620 | 135 | 5.15% | 85 | 3.24% |
| fms | 2618 | 140 | 5.35% | 96 | 3.67% |
| ftk | 2619 | 89 | 3.40% | 89 | 3.40% |
| 平均 | | | 4.16% | | 3.01% |

この結果から，モーラ語頭語尾モデルの認識結果は，モーラモデルの認識結果と比較して，ほぼ同等の結果であると言える．

### 5.6　Triphone とモーラ情報を考慮したモデルの有効性

次に，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う．Triphone にモーラ情報を考慮したモデルは，以下の手順によって作成した．

#### 5.6.1　Triphone モーラモデルの作成手順

Triphone にモーラ情報を考慮したモデルの初期モデルは，4.3.2 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，Triphome モーラモデルとする．

#### 5.6.2　認識実験の結果と考察

単語音声認識は 4.4 節で示した条件下で行なった．実験の結果を表 12 に示す．

表 12 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．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. 結　　論

本研究では，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モデルを学習し，単語音声認識を行った．その結果，モーラ情報の認識結果は，基本的なモデルと比較した場合，Diagonal-covariance の HMM において，誤り率を 5.42%から 4.32%に減少させることができた．また Full-covariance の HMM においては，誤り率を 3.36%から 2.30%に減少させることができた．改善率はそれぞれ 20.3%と 31.6%であった．そして，モーラモデルの認識率は，Full-covariance の HMM において，従来，有効な手法として使用される Triphone モデルよりも高かった．また，語頭語尾情報を使用したモデルを認識に使用した場合は，基本モデルより，認識率は向上するが，モーラ情報を使用したモデルよりも低かった．

以上の結果から，モーラ情報は，単語音声認識において，誤りを減少させることができ有効な情報であることが認められた．また，モーラ情報を用いた単音声認識の認識率は，Full-covariance の HMM を用いた場合，Triphone モデルよりも高いことが認められ，本手法の有効性が示された．

今後の課題としては，不特定話者認識や雑音下における単語音声認識に対してモーラ情報を考慮し，その有効性を確認する必要がある．

## 文　　献


[1] 水澤，村上，東田，“音節波形接続による単語音声合成” 信学技報，SP99-2(1999-05)

[2] 前田，村上，池原 “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)

[3] 妹尾，村上，前田，池原 “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)

[4] 石田，村上，池原 “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)

[5] X.D.HUANG,Y.ARIKI,M.A.JACK “HIDDEN MARKOV MODELS FOR SPEECH RECOGNITION”

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

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