モーラ情報を用いた単語音声認識*

○妹尾貴宏，村上仁一，池原悟 (鳥取大・工)

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む．単語音声認識では，入力パラメータにフォルマント情報を使用し，ピッチ情報は，通常，使用されていない．

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．この依存感関係を利用することにより音声合成 [1] や音素ラベリングの研究 [2] 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．このことから，単語音声認識においてもこの依存関係を利用し，単語のモーラ数およびモーラ位置情報を考慮して音素 HMM を作成した場合に，HMM の精度は向上すると推定できる．この結果，単語音声認識の認識率は向上すると考えられる．

2　モーラ情報とピッチ情報

特定話者の単語の発話において，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].

[1] の報告では，ピッチ周波数の分散は，ピッチ周波数の変動に比べて，非常に小さいことから，ピッチ周波数は単語に関係なく単語のモーラ位置で決定できることが示されている．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同様の傾向を示す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やすい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音素 HMM を分類して学習を行なうことで，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なった場合，単語音声認識の精度が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する．

3　評価実験

本手法の有効性を調べるために，モーラ情報を用いないモデル (以下，本論文では基本モデルと呼ぶ．) とモーラ情報を用いたモデル (以下，本論文ではモーラモデルと呼ぶ．) をそれぞれ学習し，単語音声認識実験を行なう．また，単語の語頭と語尾を考慮したモデル (以下，本論文では語頭語尾モデルと呼ぶ．) と，従来有効な手法として知られる triphone モデルも比較の対象として学習し，単語音声認識を行なう．

3.1　ラベルファイルの音素ラベルの分類

学習および評価を行なうデータには，波形ファイルとラベルファイルを含むデータベースを使用する．データベース中の全ラベルファイルの音素ラベルを単語のモーラ数とモーラ位置情報を用いて分類する．音素ラベルは母音と撥音のみ分類する．促音や子音はピッチの影響が小さいため，本研究では分類しない．なお，語頭語尾情報を用いた音素ラベルも同様に母音と撥音を分類する．triphone はすべての音素ラベルを分類する．

モーラ情報を考慮した音素ラベルの分類方法は，音素ラベルの前後に単語のモーラ数とモーラ位置情報を付加することによって分類する．語頭語尾情報を考慮した音素ラベルの分類方法は，単語の語頭および語尾にある音素ラベルを a および z を付加することによって分類する．triphone は前後の音素を-と+でつなぐことによって分類する．具体的な分類例を表 1 に示す.

表 1: 音素ラベルの分類例

基本音素	a	ts	u	s	a
モーラ情報	3a1	ts	3u2	s	3a3
語頭語尾情報	aa	ts	u	s	a
Triphone	a+ts	a-ts+u	ts-u+s	u-s+a	s-a

モーラ情報を考慮した音素ラベルは音声ラベルが”atsusa”の場合，単語のモーラ数は 3 なので，母音の前方に 3 をつけ，後方に各々のモーラ位置をつけて分類する．1 番目と 5 番目の音素 a は，分類後は 3a1 と 3a3 という音素に置き換えられ，モーラ位置が異なるため，異なった音素として扱う.

3.2　音素 HMM の作成

本論文では，基本モデル，モーラモデル，語頭語尾モデル，triphone モデルの 4 種類を使用する．HMM は初期モデルが重要であるため，モーラモデル，語頭語尾モデル，triphone モデルの初期モデルは，基本モデルから作成する．また，学習データ不足による音素 HMM の信頼性の低下を防ぐために，半連続型 HMM を使用する [5]．音素 HMM の具体的な作成手順を図 1 に示す.

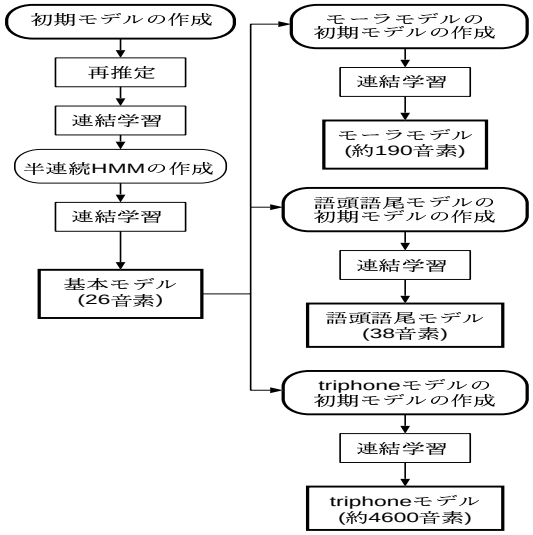

図 1: 音素 HMM の作成手順

作成された基本モデル，モーラモデル，語頭語尾モデル，triphone モデルに対して，それぞれ単語音声認識を行ない，認識率について比較を行なう.

3.3　実験条件

データベースには ATR の単語発話データベース Aset の 5240 単語を使用し，奇数番を学習データに，偶数番を評価データに使用する．使用する音声データは全て，人手によって音素境界位置が付与されている.

評価実験は，男性話者 3 名と女性話者 3 名で行なう．実験に使用する単語音声認識ツールは HTK[6] を使用する.

HMM の共分散行列には，Diagonal-covariance(以下，Diagonal) と Full-covariance(以下，Full) の 2 種類を使用する．その他の実験条件を表 2 に示す.

4　実験結果

Diagonal の HMM を用いた実験の結果を表 3 に，Full の HMM を用いた実験の結果を表 4 に示す．実験の結果から，

*Isolated-Word Speech Recognition Using Mora Position and Mora Length
By Takahiro Seo (Tottori Univ.)

表 2: 実験条件

標本周波数	16kHz		
分析窓	Hamming 窓		
分析窓長	20ms		
フレーム周期	5ms		
音響モデル	4 状態 3 ループ 半連続型		
混合分布数	256mixture		
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		
学習 DB 音素数	2,620 単語 約 15,500	評価 DB 音素数	2,620 単語 約 15,500

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向上した．その割合は，6 話者の平均で，Diagonal の場合 1.05%で，Full の場合 0.84%であった．また，triphone モデルの結果と比較すると，モーラモデルの認識率は，Diagonal では低くなるが，Full では同等の結果であった．

表 3: 単語音声認識結果 (Diagonal-covarience)

話者	認識率			
	基本	モーラ	語頭語尾	triphone
mau	96.26%	96.91%	96.30%	98.13%
mmy	93.59%	95.23%	94.39%	96.26%
mnm	94.27%	95.65%	94.58%	96.68%
faf	95.00%	95.92%	94.73%	94.62%
fms	94.47%	95.38%	96.22%	95.45%
ftk	94.96%	95.80%	95.11%	96.45%
平均	94.76%	95.81%	95.22%	96.27%

表 4: 単語音声認識結果 (Full-covarience)

話者	認識率			
	基本	モーラ	語頭語尾	triphone
mau	96.22%	98.09%	97.63%	98.51%
mmy	96.98%	97.40%	97.21%	97.75%
mnm	96.76%	97.29%	96.68%	97.52%
faf	97.75%	98.40%	97.71%	97.48%
fms	96.87%	97.90%	97.40%	97.71%
ftk	97.18%	97.71%	97.52%	97.71%
平均	96.96%	97.80%	97.36%	97.78%

5 考察

5.1 認識単語の解析

基本モデルとモーラモデルの認識単語を比較すると，モーラ情報を用いることによって効果の見られた単語の多くは，連続母音をもつ単語であった．特に長母音を短母音に誤認識する単語によく効果が見られた．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ている．効果の見られた単語の例を表 5 に示す．

表 5: 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

5.2 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 6 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 6 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．

表 6: 男性話者と女性話者の平均認識率

話者	HMM	基本	モーラ	語頭語尾	triphone
男性	Diagonal	94.71%	95.93%	95.09%	97.02%
	Full	96.65%	97.60%	97.18%	97.98%
女性	Diagonal	94.81%	95.70%	95.36%	95.51%
	Full	97.26%	98.00%	97.54%	97.63%

5.3 各モーラ語の認識結果

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，2 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 7 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 7: 各モーラ語の認識率

モーラ数	基本	モーラ	語頭語尾	triphone
1 モーラ語	94.2%	93.6%	48.1%	88.5%
2 モーラ語	95.41%	96.30%	95.33%	94.15%
3 モーラ語	96.63%	97.01%	97.86%	97.89%
4 モーラ語	97.55%	99.00%	98.56%	99.01%
5 モーラ語	100%	99.5%	100%	100%
6 モーラ語	100%	97.8%	100%	100%
7 モーラ語	100%	100%	100%	100%

表 7 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 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モデルを学習し，単語音声認識を行なった．その結果，Diagonal-covariance の HMM を用いた場合は，認識率は 6 話者の平均で 1.05%向上し，Full-covariance の HMM を用いた場合は，認識率は 6 話者の平均で 0.84%向上した．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，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．

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．

今後の課題として，不特定話者認識におい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．

参考文献

[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