日本語の同音異義語の認識*

○村上 仁一, 堀田 波星夫, 池原 悟 （鳥取大学）

1 はじめに

日本語では,「箸」,「橋」のような音韻的には同一だがアクセントの違いによって弁別できる単語が存在する．しかし，従来の単語音声認識においては主に音声の音韻的特徴が用いられており，日本語における同音異義語の音声認識の研究はあまり行われていない [8]．過去の韻律的特徴を用いた研究としては高橋ら [7] の研究がある．この研究では，韻律情報を認識率を向上させるために用いている．また中国語では，アクセントの違いにより単語の意味が異なる．このため音声認識において韻律を含めて認識する研究が多く行われている．これらの研究の多くは，音声の音韻情報として MFCC を，韻律情報としてピッチ周波数を抽出して，２つの情報を組み合わせて認識している [1][2]．しかし，母音はピッチがあるが子音はピッチがないため，音声から信頼性のあるピッチ周波数を抽出するのは困難である．特に不特定話者では，倍ピッチや半ピッチが出力されやすい．

本研究では，同音異義語を認識するために，ピッチ周波数の抽出をおこなわずに，ピッチが音響パラメータに影響を与えることを利用する．具体的には，単語のアクセント型と各モーラ位置でのアクセントの高低情報を音素ラベルに付与した HMM を作成する．このモデルを用いて同音異義語の音声認識を行い，認識精度を調査する．調査は評価データ中の同音異義語の対を抽出して行う．また，アクセント情報や前後音素環境情報を音素に付与すると音素数が膨大になるので，本研究では半連続型 HMM[5] を用いる．音響パラメータとしては MFCC と FBANK を用いる．

2 アクセントモデル･アクセント triphone モデル

本研究では音素ラベルに単語のアクセント型と各モーラ位置のアクセントの高低の情報を加えたモデル (以下，アクセントモデル)[9] を用いる．このアクセントモデルは母音，撥音，促音の音素ラベルにモーラ数およびモーラ位置およびアクセント位置の情報を付与する．具体的には母音，撥音，促音の音素の後に 7 桁の数字を加える．始めの 2 桁の数字は単語のモーラ数を表す．次の 2 桁の数字はモーラ位置を表す．次の 2 桁の数字は単語のアクセントの型を表す．最後の数字 1 桁はそのモーラ位置でのアクセントの高低を示し，0 か 1 である．0 は低，1 は高であることを表す．

また，本研究ではアクセントモデルに前後の音素環境情報を加えたモデル (以下，アクセント triphone モデル) を提案し，評価する．また，通常の音素ラベルを用いて学習した音素 HMM を基本モデル，通常の音素ラベルにおいて前後音素環境を考慮したモデルを triphone モデルとする．

アクセントモデルとアクセント triphone モデルと triphone モデルのラベルの分類例を表 1 に示す．表中のラベル表記で，+ の後の音素は後音素環境を，− の前の音素は前音素環境を表現する．図 1 に例を示す．

Table 1 音素ラベルの分類例

単語:秋 (a k i)

基本モデル	a	k	i
アクセントモデル	a0201011	k	i0202010
triphone モデル	a+k	a−k+i	k−i
アクセント triphone モデル	a0201011 +k	a0201011− k +i0202010	k− i0202010

3 評価実験

3.1 同音異議語の認識

本研究では同音異義語を認識するために，音素ラベルに単語のアクセント型と各モーラ位置のアクセントの高低の情報を加えたモデル (以下，アクセントモデル) を提案する．そしてアクセントモデルを利用して，同音異義語の認識精度を調査する．従来の音声

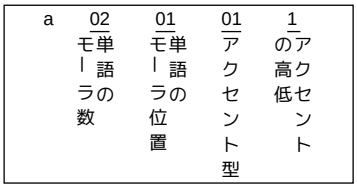

Fig. 1 アクセントを付与したラベル表記

*Speaker Independent Homonyms Speech Recognition using Accent Model. by Jin'ichi Murakami, Haseo HOTTA, Satoru Ikehara (Faculty of Engineering, Tottori University)

認識で用いられている音響パラメータの MFCC は音韻情報である．しかし，韻律情報は音韻情報に影響を与える．そのため，同音異義語の認識が可能になる．また，FBANK は，音韻情報と韻律情報を含む．そこで本研究では，MFCC と FBANK で認識実験を行う．

3.2 学習データと評価データ

データベースには ATR 単語発話データベース Aset の 5240 単語/話者の男女各 10 話者を用いる．データは男女別に，実験対象話者以外の 9 話者分の奇数番を学習データに用いる．評価データには，実験対象話者の偶数番の同音異義語を用いる．単語のアクセントは NHK 日本語発音アクセント辞典 [3] を利用する．ATR の Aset データベース中には同音異議語が 31 組 62 単語ある．しかし，表記と異なるアクセントの音声があるため，人手による聴取結果と一致する音声のみ使用する．その結果，11 組 22 単語の同音異義語を使用する．実験で用いられる同音異義語を表 2 に示す．

Table 2 評価データ（同音異義語の対）

1.	居る (01)	射る (10)
2.	代える (011)	返る (100)
3.	欠ける (011)	駆ける (010)
4.	機嫌 (011)	起源 (100)
5.	公開 (0111)	航海 (1000)
6.	置く (01)	億 (10)
7.	指名 (011)	氏名 (100)
8.	度 (01)	足袋 (10)
9.	徳 (01)	解く (10)
10.	付ける (010)	漬ける (011)
11.	因る (01)	夜 (10)

括弧内の数字の 0 はアクセントの低，1 は高を意味する．

3.3 分析条件

評価実験は，男性話者 3 名と女性話者 3 名で行う．実験には単語音声認識ツールの HTK [4] を使用する．本研究では，韻律的情報が含まれている FBANK と，一般に用いられている MFCC を音響パラメータとして使用する．HMM の共分散行列は，Diagonal covariance で行う．MFCC と FBANK は，共に同じ混合ガウス分布数を利用する．本研究で用いる音響パラメータは HTK の default 値を利用する．また HMM の実験条件を表 3 にまとめる．

Table 3 HMM のパラメータ

音響モデル	3 ループ 4 状態 半連続分布型
stream 数	3
混合ガウス分布数 (Diagonal)	MFCC 1024 + ΔMFCC 1024 + 対数パワー 64 + Δ 対数パワー 64
混合ガウス分布数 (Full)	MFCC 128 + ΔMFCC 128 対数パワー 16 + Δ 対数パワー 16

3.4 アクセントモデルとアクセント triphone モデルの作成手順

HMM は初期モデルが重要であるため，アクセントモデルと triphone モデルの初期モデルは基本モデルから作成する．また，アクセント triphone モデルの初期モデルは triphone モデルから作成する．アクセントモデルとアクセント triphone モデルの作成手順を図 2 に示す．

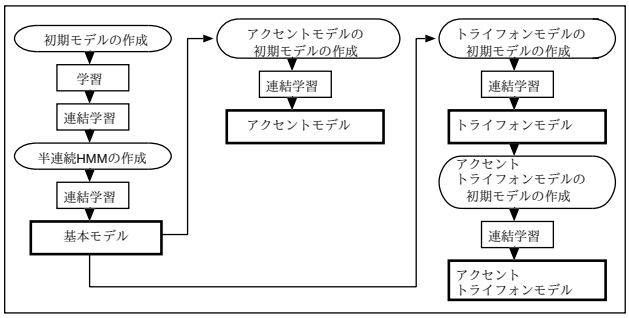

Fig. 2 アクセントおよびアクセント triphone モデルの作成手順

4 同音異義語の認識実験結果

実験結果を表 4，5，6，7 に示す．表中の mau，mmy，mnm は男性話者であり，faf，fms，ftk は女性話者である．表 4 は，音響パラメータに MFCC，共分散行列に Diagonal Covariance を用いた同音異義語の誤り率，表 5 は，音響パラメータに FBANK，共分散行列に Diagonal Covariance を用いた同音異義語の誤り率，表 6 は，音響パラメータに MFCC，共分散行列に Full Covariance を用いた同音異義語の誤り率，表 7 は，音響パラメータに FBANK，共分散行列に Full Covariance を用いた同音異義語の誤り率，の結果である．

Table 4　MFCC，Diagonal を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	27%(6/22)	18%(4/22)
mmy	18%(4/22)	27%(6/22)
mnm	36%(8/22)	27%(6/22)
faf	23%(5/22)	18%(4/22)
fms	9%(2/22)	0%(0/22)
ftk	6%(6/22)	27%(6/22)
男性平均	27%(18/66)	24%(16/66)
女性平均	20%(13/66)	15%(10/66)
平均	23%(31/132)	20%(26/132)

Table 5　FBANK，Diagonal を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	23%(5/22)	27%(6/22)
mmy	23%(5/22)	27%(6/22)
mnm	41%(9/22)	32%(7/22)
faf	23%(5/22)	23%(5/22)
fms	5%(1/22)	0%(0/22)
ftk	32%(7/22)	18%(4/22)
男性平均	29%(19/66)	29%(19/66)
女性平均	20%(13/66)	14%(9/66)
平均	24%(32/132)	21%(28/132)

Table 6　MFCC，Full を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	14%(3/22)	5%(1/22)
mmy	23%(5/22)	5%(1/22)
mnm	32%(7/22)	14%(3/22)
faf	5%(1/22)	5%(1/22)
fms	9%(2/22)	9%(2/22)
ftk	27%(6/22)	27%(6/22)
男性平均	23%(15/66)	8%(5/66)
女性平均	14%(9/66)	14%(9/66)
平均	18%(24/132)	11%(14/132)

Table 7　FBANK，Full を用いた同音異義語の誤り率

話者	アクセントモデル	アクセント triphone モデル
mau	18%(4/22)	14%(3/22)
mmy	27%(6/22)	32%(7/22)
mnm	45%(10/22)	32%(7/22)
faf	0%(0/22)	9%(2/22)
fms	5%(1/22)	0%(0/22)
ftk	14%(3/22)	9%(2/22)
男性平均	30%(20/66)	26%(17/66)
女性平均	6%(4/66)	6%(4/66)
平均	18%(24/132)	16%(21/132)

実験より以下の結果を得た．

1. 認識率が最大になる実験条件

   同音異義語の平均認識率は，HMM にアクセント triphone モデル，音響パラメータに MFCC，共分散行列に Full Covariance を用いた実験において，最も高い値，平均 89%が得られた（表 6）．しかし，男性話者の平均と女性話者の平均では，音響パラメータによって結果が異なる．男性話者の平均では，HMM にアクセント triphone モデル，音響パラメータに MFCC，共分散行列に Full Covariance を用いた実験で，認識精度 92%が得られた（表 6）．女性話者の平均では，HMM にアクセント triphone モデル，音響パラメータに FBANK，共分散行列に Full Covariance を用いた実験で，認識精度 94%が得られた（表 7）．

2. 男性と女性の比較

   男性と女性を比較すると，女性話者のほうが認識率は高い．しかし男性話者においては，MFCC は FBANK より認識率が高い．しかし逆に女性話者においては，FBANK は MFCC より認識率が高い．

3. MFCC と FBANK の比較

   平均の認識率をみると，MFCC は FBANK より同音異義語の認識精度がわずかに高いが差は小さい．また，男性話者は MFCC が有効であるのに対し，女性話者では FBANK が有効である．

4. 話者別の比較

   どの実験条件においても，話者によって認識率が大きく異なる．例えば，HMM にアクセント triphone モデル，音響パラメータに FBANK，共分散行列に Full Covariance を用いた実験では，fms の認識率は 100%であったのに対し，faf の認識率は 91%であった（表 7）．

5. アクセントモデルとアクセント triphone モデルの比較

   多くの場合，アクセント triphone モデルの方がアクセントモデルより同音異義語の認識率は高い．しかし，認識率が大きく改善されるのは，音響パラメータが MFCC で共分散行列が Full Covariance の男性話者のときである．（誤り率が 23%から 8% に改善された（表 6）．その他の実験では，大きな差はない．

5 考察

5.1 同音異義語の誤認識の分析

全ての実験条件において，同音異義語の誤認識としては，モーラ数 2 の高低のアクセントの同音異義語と，モーラ数 3 の低高高のアクセントの同音異義語が誤認識する例が多かった．

5.2 FBANK と MFCC

不特定話者において，同音異義語の平均認識精度をみると，多くの実験では,MFCC は FBANK より高い．しかし，男性話者の平均値では MFCC は FBANK より高いのに対し，女性話者の平均値では，FBANK が MFCC より高い．FBANK は，音韻情報と韻律情報を含む．一方 MFCC は，音韻情報を表現している．しかし，韻律情報は，音韻情報に影響を与える．そのため MFCC でも同音異義語の認識が可能になる．しかし，同音異義語の認識は，FBANK のほうが MFCC より妥当性があると考えていた．不特定話者における女性話者では，この予想が正しかった．しかし，不特定話者の男性話者においては FBANK より MFCC が有効であった．

5.3 男性話者と女性話者

同音異議語の認識実験の結果から，女性話者は男性話者より認識率が高い．この原因として，女性話者は男性話者と比較して，単語発話中のピッチ周波数の変化が大きいためと考えている．

ところで一般に女性話者は男性話者と比較してピッチが高い．そのため，韻律情報と音韻情報の分離が困難である．そのため，一般の単語音声認識では，女性の認識は男性の認識と比較して認識率が低下する．同音異議語では，逆の結果になっている．

6 おわりに

本研究では，従来の音声認識においてあまり行われてこなかった不特定話者における同音異義語の音声認識精度を調査した．同音異義語を音声認識するために，ピッチ周波数を抽出するのではなく，アクセントを音素ラベルに付与したモデルを提案した．そして，同一発話内容でアクセントが異なる同音異義語対に対して，音声認識率を調査した．音響パラメータには MFCC と FBANK を利用して認識を行った．不特定話者における同音異義語音声認識の実験結果より以下を確認した．

1. アクセント triphone モデルにおいて音響パラメータに MFCC を利用し，共分散行列に Full Covariance のとき，同音異義語対 11 組 22 単語において 89% の同音異義語音声認識の精度が得られた．
2. 平均値で見たとき，MFCC の認識精度は FBANK より低い．しかし男性話者では MFCC のほうが有効であったのに対し，女性話者では FBANK のほうが有効であった．
3. 話者において認識精度に大きな差がある．

今後，不特定話者における同音異義語の認識精度を高める手法として，特定話者と女性話者で効果が見られた FBANK を用いること，そのために話者適応手法や話者選択手法を用いることが考えられる．

参考文献

[1] Yi-hao Kao, Lin-shan Lee. Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language. InterSpeech 2006 , pp. 1814-1817, 2006.

[2] Dau-Cheng Lyu, Min-Siong Liang, Yuang-Chin Chiang, Chun-Nan Hsu, Ren-Yuan Lyu, “Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling”, Eurospeech 2003, 1861-1864, 2003.

[3] NHK 日本語発話アクセント辞典新版. NHK 出版, 1998. ISBN4-14-011112-7.

[4] *HTK Ver3.2 reference manual*. Cambridge University, 2002.

[5] X.D.Huang, Y.Ariki, and M.A.Jack. Hidden markov models for speech recognition.

[6] S.J. Young, J.J. Odell, and P.C. Woodland. Tree-based state tying for high accuracy acoustic modelling. *Proc. ICASSP*, pp. 307–312, 1994.

[7] 高橋, 松永, 嵯峨山. ピッチパタン情報を用いた単語音声認識. 日本音響学会講演論文集, No. 1-3-20, pp. 39–40, 1990.

[8] 村上, 荒木, 池原. 音声におけるポーズ長およびアクセント位置の情報量の考察. 日本音響学会講演論文集, No. 3-3-11, pp. 89–90, 1988.

[9] 堀田, 村上, 池原. モーラ情報およびアクセント位置をもちいた単語音声認識. 日本音響学会講演論文集, No. 3-Q-4, pp. 151–152, 2004.