不特定話者における同音異義語音声認識*

○堀田波星夫, 村上仁一, 池原悟 (鳥取大)

1 はじめに

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きた. しかし, 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. 過去の研究において, 韻律的特徴を用いた研究としては, 高橋ら [5] の研究があるが, 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 [6].

そのため以前の研究において, 特定話者における同音異義語の音声認識を行った. そして, アクセント情報を利用し韻律的情報を含む特徴パラメータとして FBANK を用いることで同音異義語の認識精度が高いことを確認した [7].

そこで本研究では, 不特定話者における同音異義語の音声認識精度を調査する. 具体的には, アクセント情報と FBANK を利用して単語音声認識を行い, 評価データ中の同音異義語の認識結果に注目して評価する.

2 木に基づく状態共有

本研究では, アクセント情報を音素に付与してラベル分類を行い音韻とアクセントを同時に認識する. そのため音素数が多くなり, HMM の信頼性のあるパラメータの推定が困難である. そこで本研究では半連続型 HMM[3] と音の決定木に基づく状態共有手法 [4] を用いた状態共有型 HMM を利用する.

状態共有型 HMM では, 音響的特徴が類似した音素 HMM の状態集合に対して, 決定木に基づくクラスタリングを行い状態共有する. そして, 状態あたりの学習データを増やして信頼性のある HMM のパラメータを推定する.

3 評価実験

本研究では音素 HMM 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加えたモデル (以下, アクセントモデル)[7] を用いる. 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. なお, 研究において単語のアクセントは NHK 日本語発音アクセント辞典 [1] を利用する. 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.

3.1 ラベル分類

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. なお, 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. そして, アクセントを用いるモデルは母音, 撥音, 促音音素の後ろ 7 桁の数字でアクセントとモーラ情報を表現する. 7 桁の数字の意味を図 1 に示す.

3.2 音素 HMM の作成

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. そして, 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. 半連続型 HMM の実験手順を図 2 に示す. また, 木に基づく状態共有型 HMM の実験手順を図 3 に示す.

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

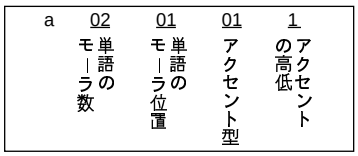

Fig. 1 ラベル表記

3.3 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. そして男女別に, 実験対象話者以外の 9 話者分の奇数番を学習データに, 実験対象話者の偶数番を評価データに用いる. 同音異義語認識実験は, 2620 単語の音声認識を行い, 評価データ中のアクセントの異なる 11 組の同音異義語の認識結果で評価する. 実験で用いられる同音異義語を表 2 に示す. なお, データベース中には表記と異なるアクセントの音声があるので, 評価データ中の同音異義語のアクセントは人手による聴取結果と一致することを確認した.

3.4 実験条件

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 [2] を使用する. その他の実験条件を表 3 に示す. なお, 状態共有型 HMM において, 半連続型 HMM の Diagonal の実験条件と同一にするために, 混合分布数は 2112(1024+1024+64) とする.

4 実験結果

半連続型 HMM で行った同音異義語認識の認識結果を表 4 に示す. また, 状態共有型 HMM で行った同音異義語認識の認識結果を表 5 に示す.

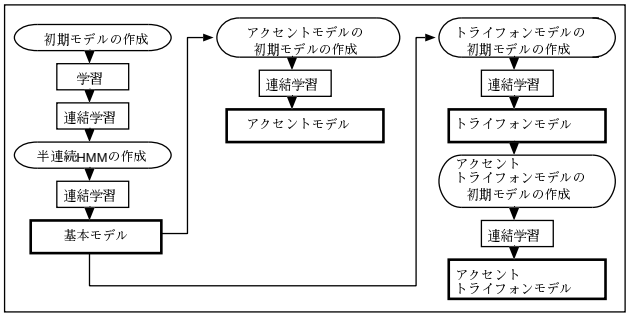

Fig. 2 半連続型音素 HMM の作成手順

*Speaker Independent Speech Recognition using Accent. by HOTTA Haseo, MURAKAMI Jin'ichi and IKEHARA Satoru(Tottori University)

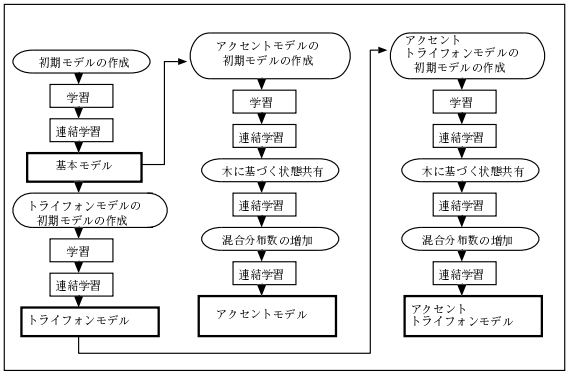

Fig. 3　状態共有型音素 HMM の作成手順

Table 2　評価データ中の同音異義語の対
括弧内の数字の 0 はアクセントの低,1 は高を意味する

居る (01)	射る (10)	指名 (011)	氏名 (100)
代える (011)	返る (100)	度 (01)	足袋 (10)
欠ける (011)	駆ける (010)	徳 (01)	解く (10)
起源 (100)	機嫌 (011)	付ける (010)	漬ける (011)
公開 (0111)	航海 (1000)	因る (01)	夜 (10)
置く (01)	億 (10)		

表中の括弧内の分母は評価データ中の全ての同音異義語数を, 分子は誤認識した同音異義語数を示す.

実験より以下の結果を得た.

1. MFCC は FBANK より同音異義語の認識精度が高い.
2. 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精度が高い.
3. 状態共有型 HMM の認識率は半連続型 HMM と比べ低い.
4. 最も同音異義語を認識できた実験では平均 89%の精度が得られた.

5　考察

5.1　同音異義語認識

多くの実験結果において, 男性話者の同音異義語認識精度は女性話者より低い. しかし, 認識精度が最も高い半連続型 HMM を用いた MFCC, Full のアクセント triphone モデルの実験においてのみ, 男性の認識精度は 92%(61/66), 女性は 86%(57/66) となる. そして, 最も高い男性の認識精度は半連続型 HMM を用いた MFCC, Full のアクセント triphone モデルでの 92%である. また, 最も高い女性の認識精度は, 半連続型 HMM を用いた FBANK, Full のアクセントモデルとアクセント triphone モデルでの 94%(62/66) である.

Table 3　実験条件

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms
音響モデル	3 ループ 4 状態
stream 数	3
MFCC 特徴ベクトル	12 次 MFCC+Δ 12 次 MFCC +対数パワー+Δ 対数パワー (計 26 次)
FBANK 特徴ベクトル	24 次 FBANK+Δ 24 次 FBANK +対数パワー+Δ 対数パワー (計 50 次)
半連続型 HMM Diagonal 混合分布数	MFCC 1024 Δ MFCC 1024 対数パワー, Δ 対数パワー 64
半連続型 HMM Full 混合分布数	MFCC 128 Δ MFCC 128 対数パワー, Δ 対数パワー 16
状態共有型 HMM Diagonal 混合分布数	MFCC 4 Δ MFCC 4 対数パワー, Δ 対数パワー 2
FBANK の混合分布数は MFCC と同様なので省略	

Table 4　半連続型 HMM における同音異義語の誤り率

	アクセントモデル	アクセント triphone モデル
Diagonal, MFCC	23%(31/132)	20%(26/132)
Diagonal, FBANK	24%(32/132)	21%(28/132)
Full, MFCC	18%(24/132)	11%(14/132)
Full, FBANK	18%(24/132)	16%(21/132)

Table 5　状態共有型 HMM における同音異義語の誤り率

	アクセントモデル	アクセント triphone モデル
Diagonal, MFCC	34%(45/132)	34%(45/132)
Diagonal, FBANK	33%(44/132)	41%(54/132)

同音異義語の実験では, アクセントが低で始まる 3 モーラ以上の単語を誤認識する結果が多い. 誤認識の例を表 6 に示す. なお, 表 6 の括弧内の数字の 0 はアクセントの低, 1 は高を意味する.

ただし, 認識結果には話者によって大きな偏りがある.

Table 6　同音異義語の誤認識例

認識結果	正解
航海 (1000)	公開 (0111)
付ける (100)	漬ける (011)

5.2　FBANK と MFCC

不特定話者のほとんどの実験結果において, FBANK の同音異義語認識精度は MFCC と比べて低い. しかし, 特定話者において FBANK を用いた同音異義語認識精度は MFCC と比べ高い [7]. そのため, 話者選択や話者適合手法を用いると FBANK の認識率を改善できると考えている.

6　おわりに

本研究では, 不特定話者においてアクセントを用いることで同音異義語の音声認識実験を行った. その結果, 前後環境も考慮したアクセント triphone モデルの MFCC において 89%の精度が得られた. 今後, 認識精度を高める手法として FBANK を用いることが考えられる.

参考文献

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.