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.

アクセントを用いた同音異義語の不特定話者音声認識

堀田波星夫† 村上仁一† 池原悟†

† 〒 680-8552 鳥取県鳥取市湖山町南 4-101, 鳥取大学工学部知能情報工学科, 池原研究室
E-mail: †{hotta,murakami,ikehara}@ike.tottori-u.ac.jp

あらまし 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, 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. 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. 同音異義語の認識はアクセント情報と韻律的情報を含む特徴パラメータとして FBANK を用いて行う. 実験の結果, アクセントと前後音素環境情報を用いたモデルと特徴パラメータに MFCC を用いることで, 89%の精度が得られた.

キーワード 不特定話者, アクセント, FBANK, 木に基づく状態共有, 同音異義語認識

Speaker Independent Speech Recognition using Accent

Haseo HOTTA†, Jin'ichi MURAKAMI†, and Satoru IKEHARA†

† Tottori University Department of Information and Knowledge Engineering, Ikehara Laboratory, 4-11, Koyamacho-minami, Tottori-shi, Tottori-ken, Japan
E-mail: †{hotta,murakami,ikehara}@ike.tottori-u.ac.jp

Abstract In past work, word speech recognition, a phoneme feature of the speech has been chiefly used. However, the word such as "Hashi(Chopsticks)" and "Hashi(Bridge)" that can be distinguished by the difference of the accent exists in Japanese. In a past research, the speech recognition of the homonym is not so researched in Japanese. Then, the speech recognition accuracy of the homonym in speaker independnt speech is investigated in this research. A future parameter FBANK including the prosodic information and accent information are used to recognize the homonym. As a result of the experiment, the accuracy of 89% was obtained by using MFCC and the model that used phoneme environmental information before and behind the accent.

Key words speaker independent speech recognition, accent, FBANK, tree-based clustering, Japanese homonym recognition

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 [6]. 過去の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. 高橋らの研究では, 音声の音韻と韻律を別々に認識している. しかし, 音声から韻律情報のみを抽出するのは困難である.

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. 音韻と韻律を分離せずに同時に認識するために, 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の情報が音素ラベルに付与しラベル分類を行った. また, 音声認識に一般的に用いられている特徴パラメータである MFCC は音韻情報しか含んでいないため, 同音異義語の認識精度が低いと予想した. そこで, 韻律的情報を含む特徴パラメータとして FBANK を用いて MFCC と比較し評価した. 実験の結果, 特定話者においてアクセント情報と 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. 調査は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することで行う. 同音異義語を認識するために, 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と前後音素環境情報を音素ラベルに付与するモデルを提案する. アクセント情報と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 [3] と状態共有型 HMM [4] を用いて HMM の信頼性のあ

るパラメータを推定する．特徴パラメータとしては FBANK を用い, 認識精度を MFCC と比較し評価する．

2. アクセント情報に対する特徴パラメータ

本研究では, 韻律的情報が含まれている FBANK を特徴パラメータとして使用する．また, 一般に用いられている MFCC も使用する．本研究で用いる特徴パラメータの実験条件を表 1 に示す．表 1 は HTK で一般的なパラメータである．なお, 特定話者の同音異義語認識において, FBANK を用いると精度が高いことが知られており, MFCC を用いると精度が 90%程度であることが知られている [7].

表 1　特徴パラメータの実験条件

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)

3. アクセント triphone モデル

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を加えたモデル (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．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．なお,

表 2　音素ラベルの分類例
単語:秋 (ā|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

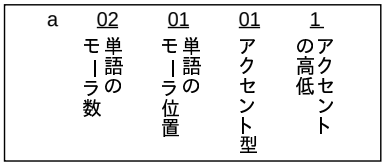

図 1　ラベル表記

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．

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 2 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．また, 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.

4. 木に基づく状態共有

本研究では, アクセント情報を音素ラベルに付与してラベル分類を行うので, 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, 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．

音の決定木に基づく状態共有型 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．状態あたりの学習データが増えることで信頼性のある HMM のパラメータの推定が可能となる．

音の決定木の各ノードには質問が付属する．本研究では, モーラ数, モーラ位置, アクセント型, アクセントの高低を考慮した質問を作成し, アクセント情報を用いたモデルの状態共有を行う．作成した質問の例を以下に示す.

- モーラ数は 3 または 4 であるか?
- モーラ位置は 1 であるか?
- アクセント型は 1,2 または 3 であるか?
- アクセントは高いか?

また, 前後音素環境を考慮した質問を作成し, 前後音素環境情報を用いたモデルの状態共有を行う．作成した質問の例を以下に示す.

- 前音素環境は鼻音であるか?
- 後音素環境は撥音であるか?

5. 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．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．半連続型 HMM の作成手順を図 2 に示す．半連続型 HMM の基本モデルは以下の手順で作成する.

（1） 連続型 HMM の初期モデルの作成
（2） 学習の実行
（3） 半連続型 HMM の作成
（4） 連結学習の実行

半連続型 HMM の基本モデル以外のモデルは以下の手順で作成する.

（1） 初期モデルの作成
（2） 連結学習の実行

木に基づく状態共有型 HMM の作成手順を図 3 に示す．状態共有型 HMM の基本モデルと triphone モデルは以下の手順で作成する.

（1） 初期モデルの作成

（2） 学習の実行

アクセントモデルとアクセント triphone モデルは以下の手順で作成する.

（1） 初期モデルの作成
（2） 学習の実行
（3） 木に基づく状態共有の実行
（4） 連結学習の実行
（5） 混合分布数の増加
（6） 連結学習の実行

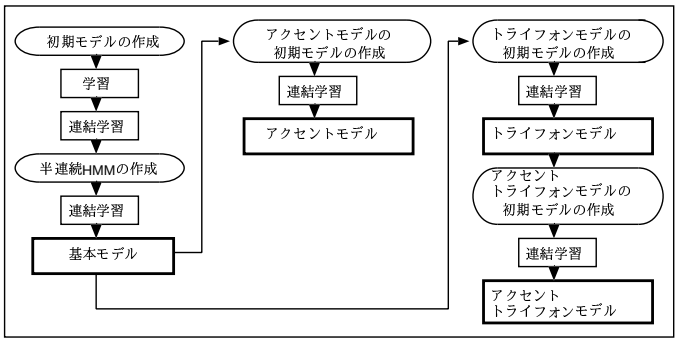

図 2 半連続型 HMM の作成手順

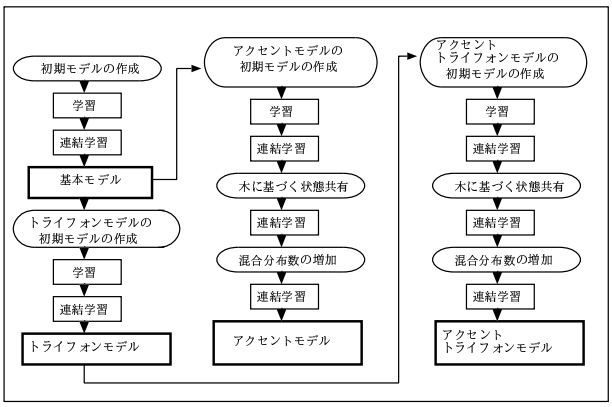

図 3 状態共有型 HMM の作成手順

6. 評価実験

6.1 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, 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 3 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.

表 3 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)		

6.2 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 4 に示す. ただし, 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.

表 4 実験条件

半連続型 HMM	MFCC 1024 Δ MFCC 1024
Diagonal 混合分布数	対数パワー, Δ対数パワー 64
半連続型 HMM	MFCC 128 Δ MFCC 128
Full 混合分布数	対数パワー, Δ対数パワー 16
状態共有型 HMM	MFCC 4 Δ MFCC 4
Diagonal 混合分布数	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	

7. 実験結果

7.1 同音異義語認識精度

半連続型 HMM で行った同音異義語認識の認識結果を表 5 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 6 に示す. 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.

表 5 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
MFCC Diagonal	男性話者	27%(18/66)	24%(16/66)
	女性話者	20%(13/66)	15%(10/66)
	平均	23%(31/132)	20%(26/132)
FBANK Diagonal	男性話者	29%(19/66)	29%(19/66)
	女性話者	20%(13/66)	14%(9/66)
	平均	24%(32/132)	21%(28/132
MFCC Full	男性話者	23%(15/66)	8%(5/66)
	女性話者	14%(9/66)	14%(9/66)
	平均	18%(24/132)	11%(14/132)
FBANK Full	男性話者	30%(20/66)	26%(17/66)
	女性話者	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） 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.

（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） MFCC は FBANK より同音異義語の認識精度が高い.

表 6 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
MFCC Diagonal	男性話者	44%(29/66)	40.9%(27/66)
	女性話者	24%(16/66)	27%(18/66)
	平均	34%(45/132)	34%(45/132)
FBANK Diagonal	男性話者	38%(25/66)	47%(31/66)
	女性話者	29%(19/66)	35%(23/66)
	平均	33%(44/132)	41%(54/132)

7.2 単語音声認識精度

半連続型 HMM の基本モデル, アクセントモデル, アクセント triphone モデルの単語音声認識の認識結果を表 7 に示す. また, 状態共有型の単語音声認識の認識結果を表 8 に示す. 表中の括弧内の分母は話者の評価データ数である. そして, 括弧内の分子の数字は誤認識した単語数を示す. なお, アクセントを用いたモデルにおいて同音異義語に誤認識している認識結果は正解として集計している.

表 7 半連続型 HMM における単語音声認識の誤り率

	基本モデル	アクセントモデル	アクセント triphone モデル	triphone モデル
MFCC, Diagonal				
男性話者	12.14% (954/7860)	8.76% (688/7860)	5.33% (419/7860)	5.36% (421/7860)
女性話者	11.72% (921/7860)	8.33% (655/7860)	6.15% (483/7860)	5.97% (469/7860)
平均	11.81% (1875/15720)	8.54% (1343/15720)	5.74% (902/15720)	5.66% (890/15720)
FBANK, Diagona				
男性話者	14.21% (1117/7860)	11.54% (907/7860)	11.18% (879/7860)	9.89% (777/7860)
女性話者	14.25% (1120/7860)	10.84% (852/7860)	10.00% (786/7860)	8.87% (697/7860)
平均	14.23% (2237/15720)	11.19% (1759/15720)	10.59% (1665/15720)	9.38% (1474/15720)
MFCC, Full				
男性話者	12.99% (1021/7860)	8.82% (693/7860)	5.38% (423/7860)	5.76% (453/7860)
女性話者	12.14% (954/7860)	7.75% (609/7860)	5.32% (418/7860)	5.53% (435/7860)
平均	12.56% (1975/15720)	8.28% (1302/15720)	5.35% (841/15720)	5.65% (888/15720)
FBANK, Full				
男性話者	14.25% (1102/7860)	10.76% (846/7860)	7.19% (565/7860)	6.59% (518/7860)
女性話者	10.34% (813/7860)	6.42% (505/7860)	5.32% (418/7860)	5.50% (432/7860)
平均	12.18% (1915/15720)	8.59% (1351/15720)	6.25% (983/15720)	6.04% (950/15720)

実験より以下の結果を得た.

（1） 半連続型 HMM の FULL の MFCC で認識精度が最も高く, 平均 94.65%が得られた.

（2） どの条件でも, 認識精度はアクセント triphone モデルまたは triphone モデル, アクセントモデル, 基本モデルの順で高い.

（3） MFCC は FBANK より同音異義語の認識精度が高い.

表 8 状態共有型 HMM における単語音声認識の誤り率

	基本モデル	アクセントモデル	アクセント triphone モデル	triphone モデル
MFCC, Diagonal				
男性話者	22.88% (1798/7860)	16.32% (1283/7860)	7.88% (619/7860)	10.05% (790/7860)
女性話者	23.79% (1870/7860)	12.79% (1005/7860)	9.38% (737/7860)	10.47% (823/7860)
平均	23.33% (3668/15720)	14.55% (2288/15720)	8.63% (1356/15720)	10.26% (1613/15720)
FBANK, Diagonal				
男性話者	44.96% (3534/7860)	16.44% (1292/7860)	13.49% (1060/7860)	12.95% (1018/7860)
女性話者	43.98% (3457/7860)	15.52% (1220/7860)	15.61% (1227/7860)	15.29% (1202/7860)
平均	44.47% (6991/15720)	15.98% (2512/15720)	14.55% (2287/15720)	14.12% (2220/15720)

8. 考　　察

8.1 同音異義語の誤認識

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 9 に示す. なお, 表 9, 12 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.

表 9 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)

8.2 単語の誤認識

半連続型 HMM での単語を同音異義語に誤認識した割合を表 10 に示す. また, 状態共有型 HMM での割合を表 11 に示す. なお, 表中の括弧内の分母は 11 組の同音異義語の誤認識数を, 分子は単語を同音異義語に誤認識した数を示す.

アクセント triphone モデルにおいて, 単語を同音異義語以外の単語に誤認識した割合は平均 39%(74/188) であり, アクセントモデルの 51%(101/200) と比べて低い. アクセントモデルの前後音素環境を考慮することによって, 単語音声認識精度が向上したためだと考えられる. 単語を同音異義語ではない別の単語に誤認識した例を表 12 に示す.

表 10　半連続型 HMM での単語を同音異義語に誤認識した割合

	アクセントモデル	アクセント triphone モデル
MFCC,Diagonal	61%(19/31)	84%(22/26)
FBANK,Diagonal	66%(21/32)	71%(20/28)
MFCC,Full	50%(12/24)	79%(11/14)
FBANK,Full	54%(13/24)	81%(17/21)
平均	59%(65/111)	79%(70/89)

表 11　状態共有型 HMM での単語を同音異義語に誤認識した割合

	アクセントモデル	アクセント triphone モデル
MFCC,Diagonal	31%(14/45)	49%(22/45)
FBANK,Diagonal	45%(20/44)	41%(22/54)
平均	38%(34/89)	44%(44/99)

表 12　単語を同音異義語ではない単語とした誤認識例

認識結果	正解
徳 (01)	置く (01)
堪える (010)	返る (100)

8.3　同音異義語のみの認識実験

平均の同音異義語認識精度が最も高い MFCC, Full のアクセント triphone モデルにおいて, 11 組の同音異義語のみの認識を行った. 実験結果を表 13 に示す. 認識精度は表 5 の結果とほとんど変わらない. 単語認識精度と同音異義語認識精度が高いためにそれほど差がでないと考えられる.

表 13　MFCC, Full, アクセント triphone モデルでの同音異義語のみの誤り率

男性話者	9%(6/66)
女性話者	11%(7/66)
平均	10%(13/132)

8.4　特定話者認識

[7] での特定話者における同音異義語の認識精度を表 14 に示す. なお, 特定話者実験は, 半連続型 HMM とアクセントモデルを用いて行われており, 本研究の実験条件とほぼ同一である. 特定話者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く, 不特定話者の同音異義語認識の傾向と同様である.

また, 単語音声認識精度を表 15 に示す. 最も単語認識精度が高いのは FBANK の Full の結果である. ただし, 女性話者で最も単語認識精度が高いのは MFCC, Full の結果であり, 不特定話者の傾向と異なる. 特定話者の同音異義語と単語音声認識精度は不特定話者より高い.

8.5　FBANK と MFCC

不特定話者のほとんどの実験結果において, FBANK の同音異義語と単語認識精度は 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と単語認識精度は MFCC と比べ高い. FBANK は MFCC より話者とモデルの依存度が高い結果となっている.

本研究では, 評価する話者以外の多数の話者のデータを学習データとして, 不特定話者のモデルを作成している. しかし, FBANK は話者を特徴付ける韻律情報を含んでおり, 多数の韻律情報が学習によって平滑化されることは認識に有効でないと考える. 本研究で用いた手法以外の不特定話者モデルの作成手法には, 多数の特定話者モデルを作成しておき評価データに最適なモデルを選択する話者選択手法や, 作成しておいたモデルを評価データに適応させる話者適応手法がある. 話者選択手法や話者適合手法を用いれば, 女性話者と特定話者において効果が見られた FBANK の韻律情報を有効に利用でき, 同音異義語の認識精度を改善できると考えられる.

表 14　特定話者, 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
MFCC Diagonal	男性話者	15%(10/66)
	女性話者	6%(4/66)
	平均	11%(14/132)
FBANK Diagonal	男性話者	11%(7/66)
	女性話者	8%(5/66)
	平均	9%(12/132)
MFCC Full	男性話者	8%(5/66)
	女性話者	8%(5/66)
	平均	8%(10/132)
FBANK Full	男性話者	5%(3/66)
	女性話者	2%(1/66)
	平均	3%(4/132)

表 15　特定話者, 半連続型 HMM における単語音声認識の誤り率

	基本モデル	アクセントモデル
MFCC, Diagonal		
男性話者	7.37% (579/7860)	4.20% (330/7860)
女性話者	7.07% (556/7860)	4.39% (345/7860)
平均	7.22% (1135/15720)	4.29% (675/15720)
FBANK, Diagonal		
男性話者	10.98% (863/7860)	7.15% (562/7860)
女性話者	9.97% (784/7860)	7.25% (570/7860)
平均	10.48% (1647/15720)	7.20% (1132/15720)
MFCC, Full		
男性話者	5.45% (428/7860)	3.45% (271/7860)
女性話者	5.01% (394/7860)	3.32% (261/7860)
平均	5.23% (822/15720)	3.38% (532/15720)
FBANK, Full		
男性話者	5.57% (438/7860)	3.03% (238/7860)
女性話者	5.34% (420/7860)	3.52% (277/7860)
平均	5.48% (858/15720)	3.29% (515/15720)

8.6　話者による認識精度の偏り

認識精度には男女差がある. 実験より以下の結果を得た.

（1） 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.

（2） 最も高い認識精度の半連続型 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, 男性の同音異義語認識精度は女性より高い.

（3） 最も高い男性の同音異義語認識精度は, 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.

（4） 最も高い女性の同音異義語認識精度は, 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%である.

ただし, 話者毎の認識結果には大きな偏りがある. 実験より以下の結果を得た.

(a) 不特定話者で最も高い認識精度は, 半連続型 HMM を用いた MFCC, Full のアクセント triphone モデルの結果である. 最も低い話者別の同音異義語認識精度は 73%(16/22) であり, 最も高い認識精度は 95%(21/22) である.

(b) a の実験条件で最も低い話者別の単語認識精度は 92.48%(2423/2620) であり, 最も高い認識精度は 96.64%(2532/2620) である.

(c) 特定話者で最も高い認識精度は, 半連続型 HMM を用いた FBANK, Full のアクセントモデルの話者別の結果である. 最も低い話者別の同音異義語認識精度は 91%(20/22) であり, 最も高い認識精度は 100%(22/22) である.

(d) c の実験条件で最も低い単語認識精度は 94.77%(2483/2620) であり, 最も高い認識精度は 96.22%(2521/2620) である.

不特定話者の話者毎の認識精度の差は特定話者に比べて大きい. 不特定話者の認識精度を改善することで, 不特定話者における話者毎の認識精度の差を小さくできると考えられる.

8.7 木に基づく状態共有

MFCC のアクセント triphone モデルにおいて共有された状態の一部を表 16 に示す. アクセント triphone モデルのアクセント情報を付与した母音・撥音・促音音素の約 20000 の状態数のうちアクセント情報の状態の質問が用いられて分類された状態の数は約 1500 である. すなわち, アクセント情報に関する質問はほとんど用いられておらず, アクセント triphone モデルの状態共有結果と triphone モデルの状態共有結果は類似している. また, モーラ数とモーラ位置の質問によって分類された状態の数はそれぞれ約 1000, アクセント型の質問によって分類された状態の数は約 600, アクセントの高低の質問によって分類された状態の数は約 150 である. 調査結果より, モーラ情報の質問がアクセント型やアクセントの高低情報より多く用いられている.

本研究では, 質問に対する評価を行っていない. 状態空間をより最適に分割する質問を用いることで状態共有型 HMM の認識精度が向上すると考えている. 本研究では最適な状態共有型 HMM の状態数も調査していない. 状態共有型 HMM において, 状態数が多すぎれば状態あたりの学習データ数が減少し信頼性のあるパラメータが推定できない. また, 状態数が少なすぎれば状態空間を十分に表現できない. モデルや特徴ベクトル等の実験条件に対して最適な HMM の状態数を求めることで, 認識精度を向上させることができると考えている.

表 16 共有された状態の例

共有状態名	共有された音素 HMM の状態
ST_e_4_1	N0402001-e0403001+k,N0402001-e0403001+ts, a0301011-e0302010+s,a0401011-e0402010+q0403010,...
ST_e_4_2	a0201000-e0202001+pau,a0201011-e0202010+pau, a0302001-e0303001+pau,a0302021-e0303020+pau,...
ST_e_4_3	N0402001-e0403001+N0404001,a0301011-e0302010+z, a0401030-e0402031+g,a0503031-e0504030+z, ...

9. 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. 調査は評価データ中に含まれるアクセントの異なる同音異義語に注目して行った. 同音異義語を音声認識するために, アクセント情報を用いたモデルを提案し単語音声認識を行った. アクセント情報と前後音素環境情報を音素に付与するラベル分類において, 音素数は膨大になり, 信頼性のある HMM のパラメータ推定は困難である. そこで, 本研究では, 半連続型 HMM と木に基づく状態共有手法を用いた状態共有型 HMM を利用して認識を行った. また, 特徴パラメータには韻律的特徴を含む FBANK と MFCC 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 モデルの MFCC において 89%の同音異義語音声認識の精度が得られた.

（2） 単語音声認識精度においてもアクセント triphone モデル, MFCC, Full, 半連続型 HMM の精度が最も高く 94.65%の精度が得られた.

（3） 韻律情報が含まれる特徴パラメータである FBANK を用いた認識精度は MFCC より低いことを確認した.

（4） 半連続型 HMM を用いた認識精度は混合分布数を同一にした状態共有型 HMM を用いた認識精度より高いことを確認した.

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認識で効果が見られた FBANK を用いることと, 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.

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.