# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | 階層的スペクトル動特性を表現した符号帳による音声認識 |
| Title(English) | |
| 著者(和文) | 古井貞熙 |
| Authors(English) | SADAOKI FURUI |
| 出典(和文) | 電子情報通信学会論文誌, Vol. J74-A, No. 5, pp. 750-757 |
| Citation(English) | , Vol. J74-A, No. 5, pp. 750-757 |
| 発行日 / Pub. date | 1991, 5 |
| URL | http://search.ieice.org/ |
| 権利情報 / Copyright | 本著作物の著作権は電子情報通信学会に帰属します。<br>Copyright (c) 1991 Institute of Electronics, Information and Communication Engineers. |

論　文

# 階層的スペクトル動特性を表現した符号帳による音声認識

正　員　古井　貞熙†

# Speech Recognition Using VQ-Codebooks Representing Hierarchical Spectral Dynamics

Sadaoki FURUI†, *Member*

**あらまし**　音声の各時点ごとのスペクトル特徴を，ケプストラムとパワーの静的特徴および階層的な動的特徴の組合せによって表現し，各単語あるいは各音素に含まれるこれらの特徴のクラスタ化によって作成した符号帳により，単語あるいは音素の特徴を表現する音声認識法を提案する．階層的な動的特徴は，各時点を中心とする長短の異なる長さの時間窓におけるパラメータ時系列の回帰係数として抽出する．こうして作成した各単語および各音素に特有の符号帳によって，入力音声をベクトル量子化し，音声区間における平均量子化ひずみの大きさによって，単語の候補を予備選択したり，単語や音素を認識したりすることができる．不特定話者の日本語100単語音声の認識実験の結果，本方法によって極めて効率良く候補単語を予備選択でき，高い精度で単語の認識ができることが確認された．また大語い単語中の/b//d//g/の音素の認識実験の結果，従来のHMM，ニューラルネット等による方法と同程度の高い音素認識性能が得られた．本論文では更に，異なる種類の特徴を単一符号帳で表現する方法と，別々の複数符号帳で表現する方法の比較についても検討を行った．

## 1. まえがき

音声知覚において重要な役割を果たしているスペクトルの動的特徴[1]を，音声から適切に抽出し，音声認識に用いる方法を開拓することは，音声認識における重要な課題の一つである．このような方法の一つである，いわゆる$\Delta$（デルタ）ケプストラム法（ケプストラムや対数パワーの線形回帰係数の時系列を用いる方法）[2]は，音声認識において極めて有効であることが種々の実験で実証され，現在広く種々の音声認識系で用いられている[2),(3]．更に，大語い単語音声認識において，ケプストラム，$\Delta$ケプストラムおよび$\Delta$パワーを組み合わせたベクトルを要素とする符号帳を単語ごとに作成し，これらの符号帳でベクトル量子化（VQ）したときのひずみを尺度とすることにより，候補単語が効率良く予備選択できることも，既に確認されている[4),(5]．

これらの方法で，回帰係数を計算するフレーム数（時間長）は，その設定値が認識性能に大きな影響を与えることはないので，通常，比較的良い結果をもたらす範囲の値（40～100 ms）に適当に決めることが多い．しかし，音声に含まれる重要な動的特徴は，実際には音素長，音節長などに対応して，種々の時間的単位の中に含まれていると考えられる．人間の聴覚においても，種々の時間分解能が存在することが示唆されている[6]．ここでは，音声の各時点ごとに，その時点を中心とする長短の異なる時間長について求めた$\Delta$ケプストラムおよび$\Delta$パワーを階層的に組み合わせて，これらのパラメータを要素とするベクトルの符号帳を，単語あるいは音素別に作成し，これによるVQひずみに基づく音声認識を試みた結果について述べる[7]．

異種の特徴ベクトルを組み合わせて用いる場合，すべての特徴ベクトルを一括したベクトルについて符号帳（単一符号帳）を作成する方法と，符号帳サイズをあまり大きくせずにVQひずみを小さく保つために，特徴ベクトルの種別ごとに符号帳（複数符号帳）を作成し，個々の符号帳によるひずみを加え合わせて全体のひずみを計算する方法が用いられる．特に学習サンプルの数が少ない場合に，後者の方法を用いることが多いが，この方法は異なる特徴ベクトル間の独立性を前提にしているので，これらの間の相関が無視されるという問題がある．ここでは，単一符号帳と複数符号帳の効果の比較についても検討する。

2. では，不特定話者の単語音声認識において，VQひ

† NTTヒューマンインタフェース研究所，武蔵野市
NTT Human Interface Laboratories, Musashino-shi, 180 Japan

ずみに基づいて候補単語を効率良く予備選択する方法とその効果について，3. では，VQ ひずみに基づいて，特定話者の大語い単語中の破裂子音/b//d//g/の認識をする実験とその結果について述べる．最後に 4. で結論と今後の課題について述べる．

## 2. 単語候補の予備選択

### 2.1 方　　法

候補単語の予備選択機能を含む単語音声認識系全体の構成を図 1 に示す．主な実験条件は次のとおりである．

（a） 認識対象：日本語 100 都市名

（b） 学習サンプル：男性 4 名（400 サンプル）

（c） テストサンプル：上と異なる男性20名(2,000 サンプル)

（d） 標本化周波数：8 kHz

（e） フレーム周期：8 ms，フレーム長：32 ms(ハミング窓)

（f） 分析法：10 次（LPC-）ケプストラム，10 次 $\Delta$ケプストラム，$\Delta$パワー

学習用音声の話者 4 名は，30 名の男性話者のスペクトルの分析に基づき，その分布を代表する話者を選んだものである．文献(4)，(5)では，$\Delta$ケプストラムと$\Delta$パワーを求めるフレーム数を 7 フレームとしたが，今回は図 2 に示すように，階層的動特性として，7 フレーム（56 ms）とその 3 倍の 21 フレーム（168 ms）の 2 種類を組み合わせて用いる効果について検討した．各時点での音声特徴は，ケプストラムと合わせて，$10(c)+10(\Delta c_7)+10(\Delta c_{21})+1(\Delta p_7)+1(\Delta p_{21})=32$ 次元のベクトルで表現されることになる．ここで $c$ はケプストラム，$p$ は対数パワー，引数は回帰係数を計算するフレーム数を示す．

各単語を特徴づける符号帳サイズはそれぞれ 64 とし，全単語の符号帳を総合した共通符号帳のサイズは，256, 512, 1,024 の 3 種類に変えた．具体的手順として

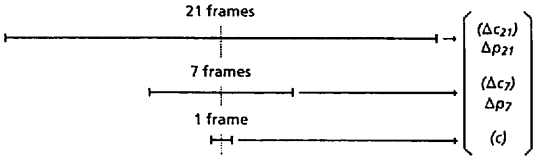


図 2　階層的スペクトル動特性を表現する特徴ベクトルの構成

Fig. 2　Structure of a feature vector representing hierarchical spectral dynamics.

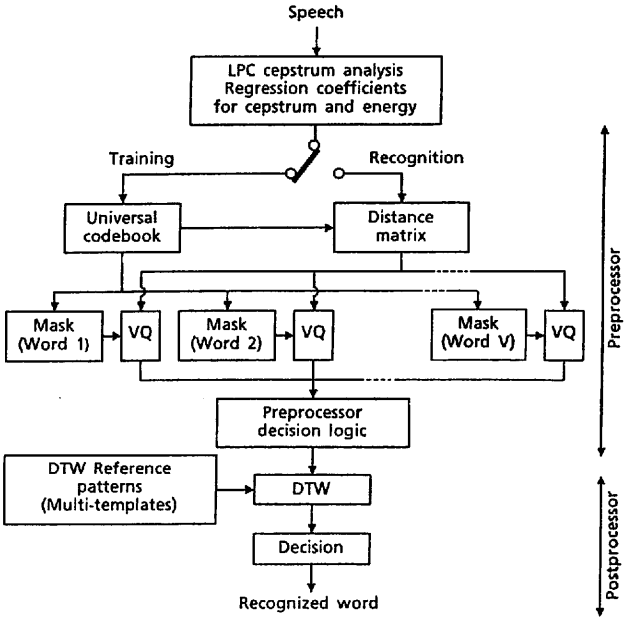


図 1　予備選択と SPLIT 法の組合せによる単語認識系の構成

Fig. 1　Block diagram of word recognizer incorporating a VQ preprocessor and a SPLIT postprocessor.

は，まず各単語ごとに，4名の学習用音声の全フレームをLBGアルゴリズム[8]でクラスタ化して，サイズが64の符号帳を作成する．次に全単語の符号帳を組み合わせた集合(64×100=6,400要素)に再度LBGアルゴリズムを適用して，上記の3種類のサイズの共通符号帳を作成する．最後に，各単語の符号帳要素を，最も距離が近い共通符号帳要素に置き換える．従って各単語の符号帳は，共通符号帳の部分集合になっている．

入力音声は，パラメータ系列に変換された後，各フレームごとに共通符号帳の全要素との距離が計算され，その結果は距離行列に蓄えられる．次にこの距離行列を用いて，入力音声を各単語の符号帳でベクトル量子化したときの量子化ひずみが計算される．各単語の符号帳は，共通符号帳の部分集合であるので，距離行列のうち，各単語の符号帳要素に対応する部分のみをマスクで取り出して，その範囲内での最小値を全フレームで加算することにより，容易に各単語符号帳に対する量子化ひずみを計算することができる．このひずみの値が，あらかじめ定めてあるしきい値よりも小さい場合は，その単語を正解の候補として残し（予備選択)，それ以外は候補ではないとして棄却する．選択された候補単語の中から，SPLIT法によってどの単語であるかを決定する．SPLIT法では，各単語が符号帳要素の時系列（4名の学習用音声から作成した4種類のマルチテンプレート）で表現されているので，上述の距離行列の値をそのまま用いて，容易に単語の判定を行うことができる．この予備選択とSPLIT法の組合せにより，大語いの単語音声認識における演算処理量を大幅に削減することができる[4],[5]．

距離（ひずみ）値を計算するときの各特徴量に対する重みは，次のように決めた．ケプストラムおよび$\Delta$ケプストラムに関しては，それぞれ，各次数ごとの分散を全次数について平均し，全次数について一様に平均分散の逆数で重みづけた．$\Delta$パワーに対する重みは，従来の実験で経験的に決めてある定数(3.0)に分散の逆数を乗じた値に設定した．

### 2.2 認識結果

SPLIT法の部分は用いずに，量子化ひずみに基づく予備選択の部分のみを用いて，ひずみの最も小さい単語に認識したときの認識誤り率を，SPLIT法のみで認識した場合の結果と比較して表1に示す．表の中で( )は，用いた特徴ベクトルすなわち符号帳の構成を示す．量子化ひずみによる実験に関しては，表の上から，動特性を抽出するフレーム数が7フレームの場合

表1 100単語音声の認識誤り率

| 方法 | 符号帳 | 符号帳サイズ 256 | 512 | 1024 |
|---|---|---|---|---|
| VQひずみ* | 単一符号帳$(c, \Delta_7, \Delta p_7)$ | __% | 13.45% | 10.90% |
| | 単一符号帳$(c, \Delta c_{21}, \Delta p_{21})$ | — | 16.45 | 10.40 |
| | 複数符号帳 $D(c, \Delta c_7, \Delta p_7)+\alpha D(c, \Delta c_{21}, \Delta p_{21})$ | — | 9.05 | 6.65 |
| | 単一符号帳 $(c, \Delta c_7, \Delta p_7, \Delta c_{21}, \Delta p_{21})$ | — | 4.45 | 3.60 |
| SPLIT | 単一符号帳$(c)$ | 6.50 | 6.20 | 6.50 |
| | 複数符号帳$D(c)+\alpha D(\Delta c_7, \Delta p_7)$ | 2.15 | 2.35 | 2.25 |
| | 単一符号帳$(c, \Delta c_7, \Delta p_7)$ | 3.00 | 2.15 | 1.95 |

* 各単語の符号帳サイズは64

$(c, \Delta c_7, \Delta p_7)$，21フレームの場合$(c, \Delta c_{21}, \Delta p_{21})$，両者のひずみを組み合わせて重み付き加算した値を用いた場合，両者の動特性を単一のベクトル$(c, \Delta c_7, \Delta p_7, \Delta c_{21}, \Delta p_{21})$に組み合わせ，単一の符号帳にしてひずみを計算した場合，のそれぞれの結果を示した．いずれの場合も，静的な特徴であるケプストラムは，常に動特性と組み合わせて用いているので，特徴ベクトルの次元数はそれぞれ，21，21，21+21=42，32である．複数符号帳によるひずみの重み付き加算は，次のようにして行った（$\alpha$：重み)．

$$D_{SUM}=(D(c, \Delta c_7, \Delta p_7) + \alpha D(c, \Delta c_{21}, \Delta p_{21}))/(1+\alpha) \quad (1)$$

SPLIT法による実験に関しては，ケプストラムのみの場合，その符号帳と動的特徴（$\Delta$ケプストラムおよび$\Delta$パワー）の符号帳を別々（複数符号帳）に作成し両者のひずみを重み付き加算した場合，およびこれらの特徴を単一符号帳にした場合の結果を示した．

(1) 量子化ひずみによる方法の結果

表1の上半分の結果から，量子化ひずみによる認識方法において，7フレーム，21フレームそれぞれ単独の場合に比べて，両者のひずみ量を組み合わせるか，単一のベクトルに組み合わせてひずみを計算することによって，大幅に誤りが低下することがわかる．例えば，共通符号帳サイズが1,024の場合，7フレームまたは21フレーム単独符号帳による誤り率は，それぞれ10.90%と10.40%であるが，両者のひずみを加算した値による誤り率は6.65%に低下する．このことは，7フレームまたは21フレームから求めた動的特徴の認識における効果はほぼ同程度であるが，実際に抽出されている特徴には相違があり，このために組合せ効果

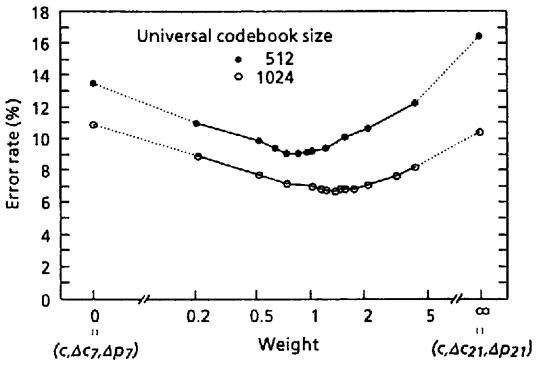


図 3　7フレームと21フレームの回帰係数を含む2種類の符号帳によるVQひずみの組合せによる単語認識結果

Fig. 3　Error rates for word recognition based on the weighted sum of the two VQ-distortions obtained using the 7-frame and 21-frame regression coefficient codebooks, respectively.

が発揮されることを示している．

この両者のひずみを重み付き加算する場合の，重み $\alpha$ に関する認識誤り率の変化を図3に示す．重みの変化に対する組合せ効果の変動は小さく，重みの値はほぼ $\alpha=1$ 程度に設定すればよいことがわかる．

更に，表1の結果から，距離値で組み合わせるよりも，階層的動特性を含む単独の符号帳にした場合の方が，誤りが前者の1/2近くに低下し，共通符号帳サイズを1,024とすれば，3.60%の誤り率が得られることがわかる．

量子化ひずみによる認識方法における共通符号帳のサイズに関しては，一貫して1,024の場合の結果が512の場合の結果を上回っている．このため追加実験として，符号帳サイズを更に大きくした場合と，その極限（無限大サイズ）として共通符号帳をもたずに各単語の学習音声から作成したものをそのまま用いる場合について実験を行った．その結果，1,024よりも大きくしても，認識性能の向上は見られず，1,024が最適であることが確認された．

（2）　SPLIT法による結果

表1の下半分の結果から，SPLIT法における共通符号帳のサイズに関しては，ケプストラムの単独符号帳，およびケプストラムと動的特徴の複数符号帳の場合は，256でほぼ十分であるが，ケプストラムと動的特徴を組み合わせた単一符号帳の場合は，512では不十分であり，1,024が必要であることがわかる．動的特徴を用いる場合の，複数符号帳と単一符号帳の結果を比較すると，符号帳サイズが256の場合は複数符号帳，512の場合はほぼ同程度，1,024の場合は単一符号帳にした方が誤りが少なくなることがわかる．符号帳サイズが1,024の場合の，単一符号帳による誤り率は1.95%，複数符号帳による誤り率は2.25%である．

Kai-Fu Leeらの HMM を用いた音声認識実験[3]では，256のサイズの符号帳を用いて，単一符号帳よりも複数符号帳の方が良いことを示しているが，この結果はここで得られた実験結果とも一致している．符号帳サイズを大きくすることができれば，単一符号帳の方が良くなると予想される．

なお表1の最下段の，動的特徴を含む単一符号帳を用いる場合について，符号帳サイズに関する補足実験を行った結果，サイズは1,024で十分であり，ベクトル量子化を行わなかった場合の誤り率は，1,024の場合の誤り率（1.95%）にほぼ等しいことが確かめられた．

（3）　量子化ひずみによる方法とSPLIT法の比較

以上の実験から，7フレーム，21フレームの2種類の動特性を組み合わせた符号帳を用いれば，DPやHMMを用いなくても，不特定話者の100単語音声の認識において，3.6%の認識誤り率が得られることがわかった．この値は，動的特徴（$\Delta$ケプストラムおよび$\Delta$パワー）を用いないSPLIT法の場合の約1/2である．$\Delta$ケプストラムおよび$\Delta$パワーで表現される動特性は，音声の比較的ミクロな時間特性であり，DPやHMMで表現される比較的マクロな時間特性とは異なるが，上の実験結果は，ミクロな時間特性の集積でも単語の特徴をとらえることができることを示している．

なお，SPLIT法の場合に7フレームと21フレームの動特性を組み合わせる効果に関する実験も行ったが，この場合はほとんど効果が認められなかったので，この結果は省略した．より複雑なタスクで実験を行えば効果が得られると思われる．

### 2.3　予備選択の結果

次に，第1位のみでなく，距離値にしきい値を設けて，しきい値より小さい値をもつ複数の候補を予備選択する場合について検討した．しきい値（$\theta$）は，式（2）に示すように，各入力に対して全候補との距離（$D_n$）の内の最小値を求め，その値に一定のバイアス（$\theta_0$）を加えた値に設定した．

$$\theta=\theta_0+\mathrm{Min}\{D_n\} \qquad (2)$$

7フレーム，21フレームの特徴を別々の符号帳にし

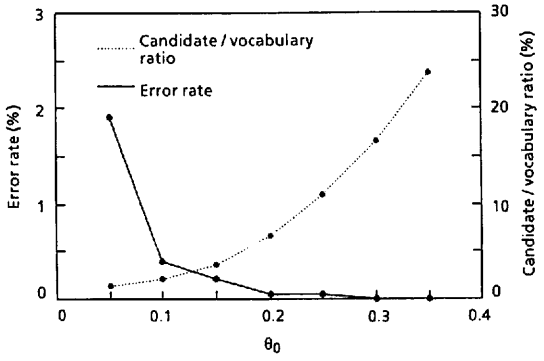


図 4 単語候補の予備選択の結果（二つの符号帳による距離を組み合わせる場合）

Fig. 4 Preprocessor performance for the multiple codebook method as a function of the threshold bias $\theta_0$.

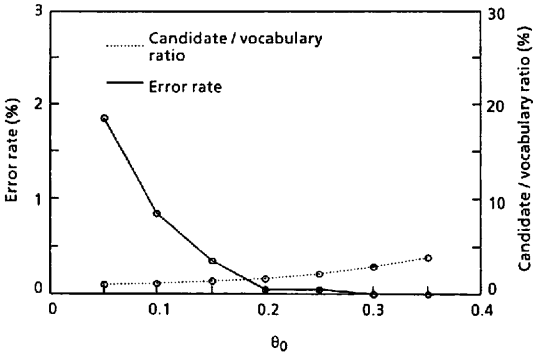


図 5 単語候補の予備選択の結果（二つの特徴を組み合わせた単一符号帳の場合）

Fig. 5 Preprocessor performance for the single codebook method incorporating hierarchical dynamic features.

て，最適な重みで両者の距離を加算した場合の結果を図4に，単一符号帳にした場合の結果を図5に示す．いずれの場合も，しきい値バイアス（$\theta_0$）を変えたときに，正しい候補が脱落する誤り率と，選択される候補数の全候補に対する割合を示した．単一符号帳にした方が，候補数を大幅に減らせることがわかる．脱落誤りが0.05%（$\theta_0=0.2$）のとき，選択される候補数の全候補に対する割合は，複数符号帳のとき6.61%，単一符号帳のとき1.67%である．

更に，選択される候補数が，各入力音声に対して1単語のみの場合を調べたところ，図6に示す結果が得られた．$\theta_0=0.2$の条件では，複数符号帳の場合に全入力音声の28.0%，単一符号帳の場合に全入力音声の71.8%に対しては1単語のみが選択される．これらの

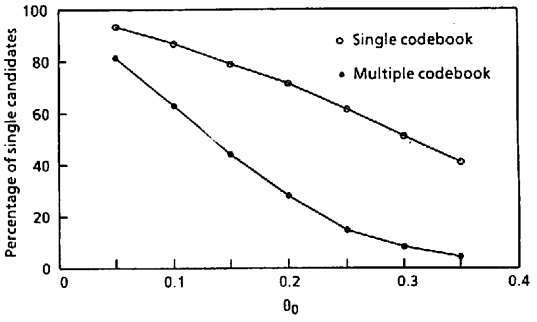


図 6 予備選択によって1入力音声に対し単一候補が選択される割合

Fig. 6 Probability of obtaining a single candidate for each input utterance as a function of the threshold bias $\theta_0$.

場合は，以後の認識処理（DP演算）が不要となるので，この数を差し引くと，単一符号帳の場合に認識処理が必要な候補数は，全候補数の0.95%になる．

## 3. 音素認識

### 3.1 方　法

ATRで作成された5,240単語データベース（発声者MAU）に含まれる/b//d//g/音声を用いて，動特性を含む符号帳による認識実験を試みた．話頭・話尾切断音節を用いた音声知覚に関する研究[1]において明らかにしたように，一般的に人間が音節および音素を知覚するときに最も重要な情報は，子音部から母音部へのスペクトル変化の最も大きい区間にある．この知見に基づき，上記データベースから，/b/, /d/, および/g/とそれらに後続する母音の境界を中心に一定長の音声区間を切り出し，これを用いて/b//d//g/の認識実験を行った．

上記データベースには/b/, /d/, /g/がそれぞれ445, 382, 512個含まれるので，それぞれ奇数番目に出現したものを学習サンプルに，偶数番目に出現したものをテストサンプルに用いた．標本化周波数，分析次数などの音声分析条件は単語の場合と同様であるが，音素認識には単語認識よりも時間的にやや細かい情報が必要と考えられるので，フレーム長は25 ms，フレーム周期は7.5 msとした．$\Delta$ケプストラム，$\Delta$パワーを求めるフレーム数は，3フレームと11フレームの間で変化させて実験を行った．

フレーム数を変えたときの，$\Delta$ケプストラムと$\Delta$パワーの/b//d//g/に関する平均分散を図7に示す．$\Delta$ケプストラムについては，1次から10次までの係数に

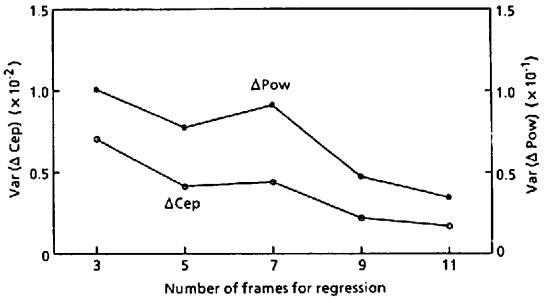


図 7 $\Delta$ ケプストラムと $\Delta$ パワーの/b, d, g/に関する平均分散

Fig. 7 Mean variances of $\Delta$ cepstra and $\Delta$ power, calculated for various numbers of frames.

表 2 音素/b/, /d/, /g/の認識誤り率

| | | 符号帳サイズ | | | |
|---|---|---|---|---|---|
| | | 32 | 64 | 128 | 256 |
| 特徴量 | ($c$) | 7.32% | 3.89% | 3.74% | 2.99% |
| | ($c$, $\Delta c_5$, $\Delta p_5$) | — | 3.44 | 1.49 | 1.35 |

関して平均した値（距離計算のときの重みの逆数に相当）を示した．

### 3.2 実験結果

認識および学習に用いる音声区間を，子音・母音境界を中心に 26 フレーム (195 ms) とし，回帰係数を求めるフレーム数を 5 フレームとして，各音素を特徴づける符号帳のサイズを 32 から 256 の間で変えたときの，種々の符号帳条件における認識誤り率を表 2 に示す．なお本実験では，各音素に対応する 3 種類の符号帳を融合した共通符号帳は作成せず，各符号帳を独立に取り扱った．

表 2 の結果より，符号帳サイズは，ケプストラムのみのとき (c) は 64 程度，回帰係数を組み合わせるとき ($c$, $\Delta c_5$, $\Delta p_5$) は 128 程度が必要であり，それ以後もサイズを増すに従ってやや誤り率が低下する傾向があることがわかる．ケプストラムのみの場合に比べて，回帰係数を組み合わせることにより，認識誤りの大きな低下が見られ，符号帳サイズを 256 とすれば 1.35%の誤り率が得られる．この値は，ケプストラムのみの場合の 1/2 以下である．以上の結果に基づき，以下の実験はいずれも，ケプストラムに回帰係数を組み合わせ，符号帳サイズを 256 として行った．

回帰係数を求めるフレーム数を変えたときの，誤り率の変化を図 8 に示す．フレーム数としては，5 ないし 7 (37.5〜52.5 ms) 程度が最適であることがわかる．回帰係数を求めるフレーム数を 5 フレームに固定し，認識に用いるフレーム数（音声区間の両端のそれぞれ 2 フレームは回帰係数が算出できないので認識には用いないため，最大のフレーム数は，26−2×2=22 となる）を変えたときの，誤り率の変化を図 9 に示す．認識に用いるフレーム数は 8 フレーム (60 ms) あればほぼ十分であり，それ以上の範囲では長さによる誤り率の変化は小さい．細かく見ると，12 フレーム (90 ms) のときに誤り率が最小となり，このときの値は 1.20% である．

次に，回帰係数を求めるフレーム数が 5 フレームの場合と，11 フレームの場合の特徴量を組み合わせる効果を調べた．単語のときと同様に，別々の符号帳 ($c$, $\Delta c_5$, $\Delta p_5$)($c$, $\Delta c_{11}$, $\Delta p_{11}$) として距離（ひずみ）値で組み合わせる場合と，単一符号帳 ($c$, $\Delta c_5$, $\Delta p_5$, $\Delta c_{11}$, $\Delta p_{11}$)

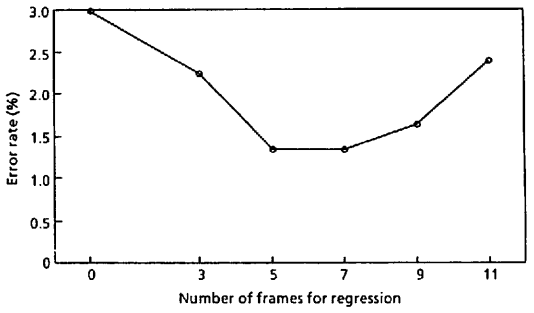


図 8 回帰係数を求めるフレーム数を変えたときの/b, d, g/認識誤り率の変化

Fig. 8 /b/,/d/,/g/ recognition error rates as a function of the number of frames for regression analysis.

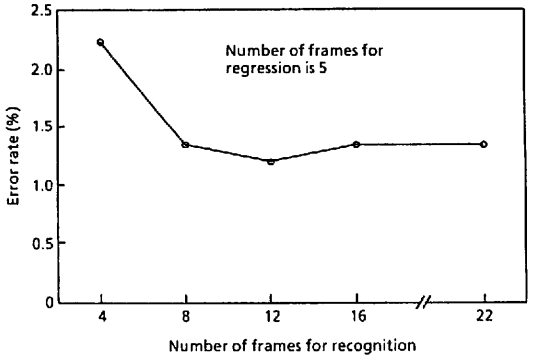


図 9 認識に用いるフレーム数を変えたときの/b, d, g/認識誤り率の変化（回帰係数を求めるフレーム数は 5 フレーム）

Fig. 9 /b/,/d/,/g/ recognition error rates as a function of the speech length.

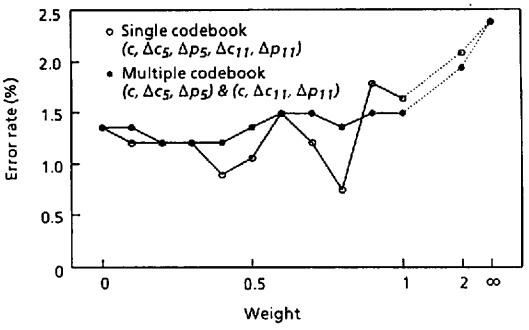


図 10 二つの符号帳による距離を組み合わせる場合と二つの特徴を組み合わせた単一符号帳の場合の/b, d, g/の認識結果

Fig. 10 Comparison of /b/, /d/, /g/ recognition error rates using the multiple codebook and single codebook methods as a function of the weighting factor combining two distortions.

とする場合の2種類の実験を行った．11フレームの場合の特徴量に対する重みを変えたときの，誤り率の変化を図10に示す．いずれの場合も2種類の特徴量を組み合わせる効果が見られるが，単一符号帳の場合の方がその効果がやや大きい．重みが0.4のときに両符号帳に共通して誤り率はほぼ最小となり，このときの単一符号帳による誤り率は0.90%（認識率99.10%）である．この値は，同じ音声データに関して，HMMやNeural Netを用いた方法によって得られている値[9],[10]とほぼ同等である．

### 3.3 音素スポッティング

これまでの実験では，人手による音素ラベリングに従って切り出した音声区間を未知音声として用いていたが，より一般的な状況での，スペクトル動特性を表現した符号帳による音声認識法の有効性を確かめるため，/b/, /d/, および/g/のスポッティング実験を行った．

本実験では，まず各音素について，すべての学習サンプルの，子音と後続母音の境界を中心とする12フレームの音声区間を用いて，ケプストラムと単一階層の$\Delta$ケプストラムおよび$\Delta$パワーからなる符号帳($c$, $\Delta c_5$, $\Delta p_5$)を作成した．未知音声の区間を，子音と後続母音の境界を中心とする300 msの区間に拡大し，その区間における音素のスポッティングを行った．なお，語頭などで300 msの区間の一部が音声区間をはずれる場合は，全区間が音声区間となるように適宜その区間をずらして実験を行った．

その未知音声区間について，18フレーム(135 ms)の窓を1フレームずつずらしながら，その窓内の特徴ベクトルを各音素の符号帳でベクトル量子化した．その窓内で平均化した量子化ひずみが最小となる窓の位置および符号帳を検出し，その最小のひずみを生ずる符号帳の音素が発声されているものと認識した．その認識結果が実際に発声された音素と同じであれば正解とし，検出位置に関しては評価は行わなかった．

認識実験の結果，ラベリングに従って切り出した音声を用いた実験結果とほぼ同程度の，1.05%の平均誤り率が得られた．この結果は，ここで提案した方法が安定に動作することを示している．

## 4. むすび

音声中の各時点を中心とする長短の異なる長さの区間から抽出した，階層的なスペクトル動特性を用いる新しい音声認識方法を提案し，認識実験により評価を行った．その結果，単語や音素別にこのような特徴ベクトルからなる符号帳を作ることにより，単語や音素の特徴が表現・抽出できることがわかった．入力音声をこれらの符号帳でベクトル量子化したときの量子化ひずみを用いることによって，高い精度で認識を行ったり，認識精度を下げずに効率良く候補を予備選択することができる．

単一符号帳と複数符号帳の比較実験の結果，符号帳の大きさが小さい場合は複数符号帳を用いた方がよいが，符号帳を十分大きくできる場合には，単一符号帳を用いた方が，高い認識性能を得ることができることが示された．

このようなスペクトルの過渡特性に着目した認識方法は，発声速度の変化にも強いことが赤木の実験[11]でも示唆されており，このような符号帳を各音素について用意し，音声区間を並列的にスキャンすることによって，音素スポッティングをベースとした連続音声認識を行うことも可能であると思われる．階層的スペクトル動特性の表現は，一種のセグメント特徴の表現法と考えることもできる．今後は，動的特徴を含むセグメント符号帳によるVQひずみを入力とする認識方法についても，検討してみたい．

**謝辞** 音素認識実験に用いた音声データに関して便宜を図って頂いたATR研究所の各位，ならびに実験を手伝って下さった当所松井知子研究員に感謝する．

### 文 献

(1) Furui S.: "On the role of spectral transition for speech perception", J. Acoust. Soc. Amer., 80, 4, pp. 1016-1025 (Oct. 1986).

(2) Furui S.: "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust., Speech, Signal Processing, **ASSP-34**, 1, pp. 52-59 (Feb. 1986).
(3) Lee K.-F., Hon H.-W and Reddy R.: "An overview of the SPHINX speech recognition system", IEEE Trans. Acoust., Speech, Signal Processing, **ASSP-38**, 1, pp. 35-45 (Jan. 1990).
(4) 古井貞熙："大語彙単語音声認識におけるスペクトル動特性を用いた単語予備選択", 音声研資, SP86-77 (1986-12).
(5) Furui S.: "A VQ-based preprocessor using cepstral dynamic features for speaker-independent large vocabulary word recognition", IEEE Trans. Acoust., Speech, Signal Processing. **ASSP-36**, 7. pp. 980-987 (July 1988).
(6) 寺西立年："聴覚の時間的側面", 難波精一郎編, 聴覚ハンドブック, 第7章, pp. 276-319, ナカニシヤ出版(1984).
(7) Furui S.: "On the use of hierarchical spectral dynamics in speech recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Albuquerque, S15a. 10 (April 1990).
(8) Linde Y., Buzo A. and Gray R. M.: "An algorithm for vector quantizer design", IEEE Trans. Commun., COM-28, 1. pp. 84-95 (1980).
(9) Waibel A., Hanazawa T., Hinton G., Shikano K. and Lang K. J.: "Phoneme recognition using time-delay neural networks", IEEE Trans. Acoust., Speech, Signal Processing, **ASSP-37**. 3, pp. 328-339 (March 1989).
(10) 平田好充, 橋本泰秀, 中川聖一："混合連続出力分布 HMM を用いた有声破裂音の識別", 1989 信学秋季全大, A-18.
(11) Akagi M. and Tohkura Y.: "On the application of spectrum target prediction model to speech recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, New York, S3. 11 (April 1988).

(平成2年8月29日受付, 12月3日再受付)

**古井　貞熙**

昭43東大・工・計数卒, 昭45同大大学院修士課程了. 同年NTT電気通信研究所入社. 以後, 同研究所において, 音声認識, 話者認識, 音声知覚などの基礎研究に従事. 昭53～54ベル研究所客員研究員. 現在NTTヒューマンインタフェース研究所音声情報研究部長. 工博. 昭50米沢賞, 昭63論文賞, 平2著述賞, 昭60, 62音響学会論文賞, 平1科学技術庁長官賞, IEEE ASSP Society Senior Award各受賞. 著書「ディジタル音声処理」「Digital Speech Processing, Synthesis, and Recognition」「音声情報工学」(分担),「音の科学」(分担)など. IEEE, 日本音響学会各会員.