論文 / 著書情報
Article / Book Information

論題(和文)	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
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子通信学会技術研究報告, Vol. SP86-77, , pp. 49-56
Citation(English)	, Vol. SP86-77, , pp. 49-56
発行日 / Pub. date	1986, 12
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。 Copyright (c) 1986 Institute of Electronics, Information and Communication Engineers.

SP86-77

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

古井貞熙（NTT通研）

1986年12月19日

社団法人 電子通信学会

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

A VQ-Based Preprocessor Using Spectral Dynamic Features for Large Vocabulary Word Recognition

古井 貞熙
Sadaoki Furui

NTT電気通信研究所
NTT Electrical Communications Laboratories

あらまし　不特定話者の大語彙単語音声認識における演算処理量を大幅に低下させることを目的として、ベクトル量子化の原理にもとづく候補単語の予備選択法について検討した。各単語の学習音声を短時間毎にケプストラム、およびケプストラムと対数エネルギーの線形回帰係数によって分析し、そのベクトル系列のクラスタ化にもとづく符号帳を各単語毎に求める。この各要素には全単語の共通符号帳の要素を用いる。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歪量にもとづいて、候補単語の予備選択を行ない、選択された候補について、共通符号帳とマルチテンプレートを用いたＳＰＬＩＴ法によって単語決定を行なう。男性20名の100都市名単語音声の認識実験により、この方法の高い有効性が確認された。

Abstract　This paper proposes a new method of a VQ (Vector Quantization)-based preprocessor for reducing the amount of computation in large vocabulary isolated word recognition. A speech wave is analyzed by time functions of both cepstrum coefficients and their short-time regression coefficients, and a universal VQ codebook for these time functions is constructed based on a multiple speaker-multiple word database. Next, a separate codebook is designed as a subset of the universal codebook for each word in the vocabulary. These word-based codebooks are used as a front-end preprocessor to eliminate word candidates whose distance scores are large. A dynamic time-warping processor based on a word-dictionary in which each word is represented as a time-sequence of the universal codebook elements (SPLIT method) then resolves the choice among the remaining word candidates. Effectiveness of this method has been ascertained by recognition experiments using a database consisting of words from a vocabulary of 100 Japanese city names uttered by 20 male speakers.

1．まえがき

近い将来の音声認識のニーズとして、不特定話者の数１０００単語程度の大語彙の単語音声認識への要求が高まってくると予想される。不特定話者の音声を認識する方法としては、これまでにマルチテンプレートによる方法[1]、HMMによる方法[2]、識別関数による方法[3]など種々の方法が提案されている。このうちマルチテンプレート法には、話者による変動の確率的モデルが不明確であっても、距離関係だけにもとづいて代表値が設定できるので、多数話者のデータベースさえあればシステムの構成が比較的容易にできるという長所がある。しかもHMMや識別関数法と異なって、単語内のスペクトル変化の時間情報がそのまま保たれているというメリットもある。ただしこの方法では、１つの単語に対して複数の標準パターンを用意するので、標準パターンの記憶容量と認識のための演算処理量が大きくなる問題がある。この問題に対処する方法として、記憶容量と距離計算量が語彙数にほとんど依存しないＳＰＬＩＴ法が有効であるが、この場合でもＤＴＷ（時間軸正規化マッチング）の演算処理量は語彙数に従って増加し、語彙数が非常に大きくなると距離計算量に比べて無視できなくなる。

このための対策としては、ＤＴＷを行なう候補単語を予備選択によって絞る方法が考えられ、最近単語別の符号帳を用いたベクトル量子化による方法が提案されている[4]。これは、各単語について、話者による変動を包含したスペクトル包絡の符号帳を用意しておき、入力音声をベクトル量子化したときの歪が比較的小さい符号帳に対応した単語を候補として選択する方法である。ただし、単純にスペクトル系列をベクトル量子化したのでは、時間情報が全く失われてしまうため、時間的に不適切な対応づけが起り、語彙数が大きくなると高い性能が得られない。このため、ベクトル量子化による符号帳の各要素に、単語音声区間内での出現位置に関する情報を確率的に各符号に持たせることが試みられ、有効性が示されている[5]。ただしこの方法は、音声区間の終端が検出されないと処理が開始できず、さらに音声区間の検出精度の影響を受けやすいという問題がある。また、単語別に符号帳を用意する方法には、距離計算量が語彙数に比例して増大する問

題がある。

ここでは、スペクトル情報にその時点での動的特徴を組み合わせ、さらに全単語に共通の符号帳を設定することにより、予備選択の大幅な性能向上を図る試みについて報告する。単語音声認識および話者認識における動的特徴の有効性についてはすでに文献[6][7][8]で確認されている。動的特徴を含むベクトル量子化による話者認識方法の有効性については、すでに文献[9]で確認されている。

2．音声分析とデータベース

音声波を遮断周波数4kHzの低域フィルタに通したのち、8kHz、11bitで標本化および量子化を行なう。32ms長のハミング窓を8msずつずらしながら音声波に乗じて波形を切り出し、LPC分析により1～10次の（LPC）ケプストラムと対数エネルギーを抽出する。次に、隣接する複数（7～15）フレームのパラメータ系列をフレーム周期毎に取り出し、動的情報として、各ケプストラム係数と対数エネルギーの時系列の線形回帰係数を抽出する。音声波は8ms毎に、ケプストラム係数とケプストラム係数及び対数エネルギーの回帰係数のセットで表現されることになる（対数エネルギーの絶対値は話者の変化や発声のたびごとに変動しやすいので用いない）。

実験には、男性20名が発声した日本語100都市名単語音声（計2000サンプル）をテスト用音声として用いた。これとは別の時期に、男性話者の声質の変動を代表していると見なされる4名の話者（上記20名とは異なる話者）が発声した100都市名単語音声を学習用音声として用いた。

3．単語毎に符号帳を作成する方法

3．1．方法

第1段階として、各単語毎に別々に符号帳を作成する方法により、特徴量の有効性などの検討を行なった。上述の男性4名の学習用音声を用いて、図1に示すように、各単語毎に4名の全フレームのパラメータセット（ケプストラム、およびケプストラムと対数エネルギーの回帰係数）をクラスタ化し、各単語のパラメータセットを代表する符号帳を作成する。クラスタ化の方法には、文献[10]の方法を用いた。各フレームのパラメータセットが異なる種類のパラメータから構成されているので、サンプル間の距離計算においては、ケプストラム、ケプストラムの回帰係数、対数エネルギーの回帰係数のそれぞれに対して適切な重みを与えた。この重みは、文献[7]の実験で得られている最適値を用いた。各単語の符号帳の大きさは、8、16、32、64の4種類に変えて実験を行なった。

未知入力音声に対しては、各単語の符号帳でベクトル量子化し、入力音声の全フレームにおける平均歪量が比較的小さい単語候補を選択して、次のDTW処理へ送る。単語候補の選択方法として、次の2種類の方法を検討した。

(1) 全認識対象語彙のうち、歪量が比較的小さい一定数の候補を選択する。［方法A］

(2) 歪量に対してしきい値を設け、しきい値よりも小さい歪量を有する候補を選択する。しきい値（θ）に関しては、次の2つの方法を試みた。

$$\begin{cases} \theta = \theta_0 & （定数）\quad ［方法B-1］ \\ \theta = \theta_0 + \min_n (dn) & ［方法B-2］ \end{cases}$$

ここでdnは、n番目の単語の符号帳による歪量である。

最後のDTW部では、候補単語についてSPLIT法によりマルチテンプレート（4名の学習サンプルをそれぞれ符号帳の要素の系列で表わしたもの）を用いてマッチングを行ない、最も距離の小さいテンプレートに対応する単語を認識結果として出力する。

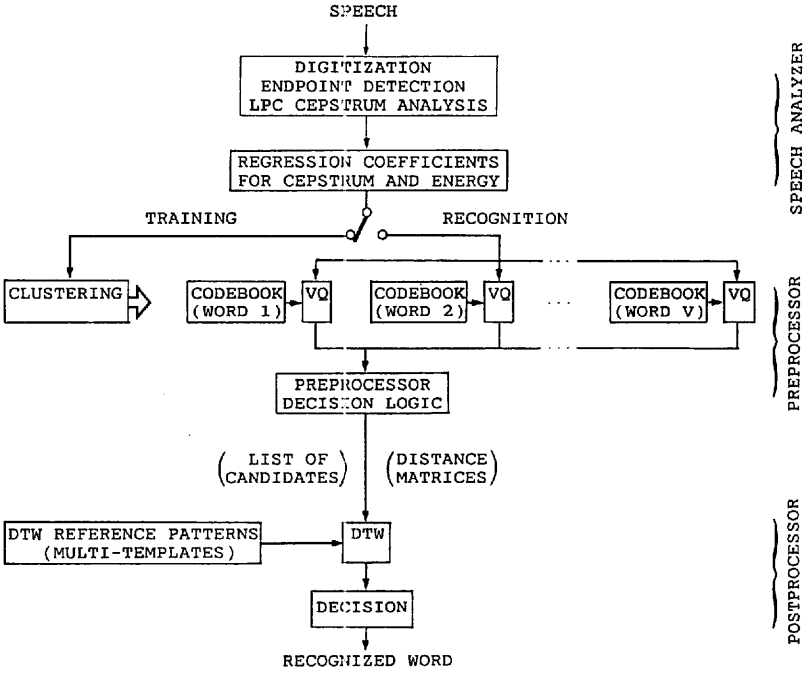


Fig. 1 - Block diagram of isolated word recognizer incorporating a VQ preprocessor based on word-based codebooks and a SPLIT post-processor.

3．2．特徴量の有効性

まずパラメータとして、ケプストラム（C）、ケプストラムの回帰係数（C'）、および対数エネルギーの回帰係数（E'）のすべてを組み合わせ、各単語の符号帳サイズを64に固定した場合について実験を行なった（'は回帰係数を表わす）。回帰係数を抽出するパラメータ系列の長さを7～15フレーム（56～120ms）に変えたときの予備選択誤り、即ち正しい単語が上位の候補に含まれない割合を図2に示す。この図から、歪量が小さい方から第2～10位程度まで選択する場合は、13フレーム以下ではフレーム数による誤り率の差はほとんどないが、第1位のみに着目すると、11フレームのときに最も小さい誤り率を示すことがわかる。

上記の実験結果から、回帰係数を抽出する長さは仮に11フレームとし、予備選択に用いる特徴パラメータを変えたときの予備選択の性能の比較を行なった。図3は実験結果を示したもので[11]、（C）のみ、（C'）のみ、（C，E'）の組み合わせ、（C'，E'）の組み合わせ、およびこれらのすべてを組み合わせた場合（C，C'，E'）において、量子化歪の小さい方から一定数の候補を選択するとき（方法A）の選択誤りの割合を示す。ケプストラムよりもケプストラムの回帰係数を用いる方が、誤り率が1／3あるいはそれ以下に小さく、ケプストラムに、ケプストラムと対数エネルギーのそれぞれの回帰係数を組み合わせることにより、誤り率が大きく低下することがわかる。（C，C'，E'）を用いて第10位まで選択するときに、正しい単語が選択される割合は99．8％である。

図4は方法B－2の場合のθ_0と、予備選択誤り率および候補単語として選択される単語数の全単語数に対する比率の関係を示す。この場合は3種類の特徴パラメータの組み合わせの場合の結果を示したが、（C）、（C，E'）、（C，C'，E'）となるに従って、誤り率が大きく低下し候補単語数を小さく絞ることができることがわかる。θ_0=0．2としたときに、正しい単語が候補に含まれる割合は99．8％、選択される単語数は100単語中平均4．5単語である。

図5は回帰係数を抽出するパラメータ系列の長さが7フレームと11フレームの場合について、方法Aを用い、各単語の符号帳サイズを8、16、32、64と変えたときの誤り率の変化を示したものである。これからわかるように、符号帳サイズが32と64の場合の差は小さく、2位～10位程度までとったときの誤り率は、回帰係数を求める長さが7フレームでも11フレームでもほとんど差がない。

一方文献[7]に示されているように、DTWによる認識では、回

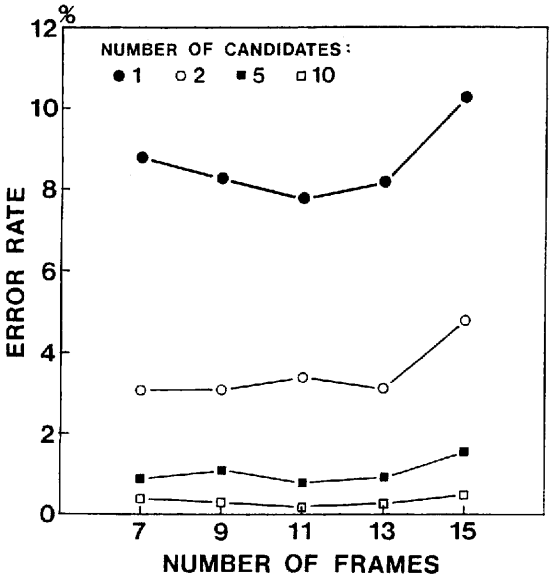


Fig. 2 - Relation between number of frames used for regression analysis and preprocessor error rate.

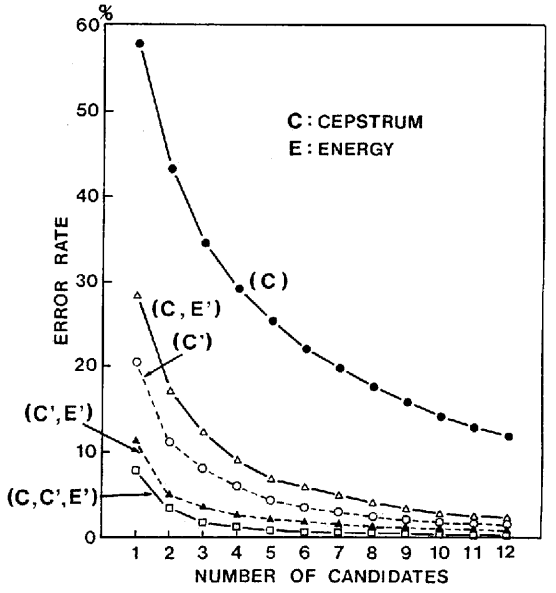


Fig. 3 - Comparison of the preprocessor performances for five feature parameter conditions (Method A: Preselection by distance rank order).

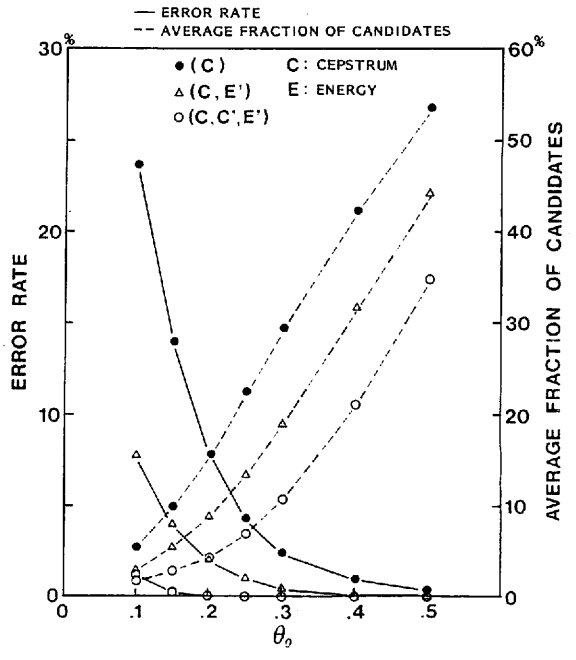
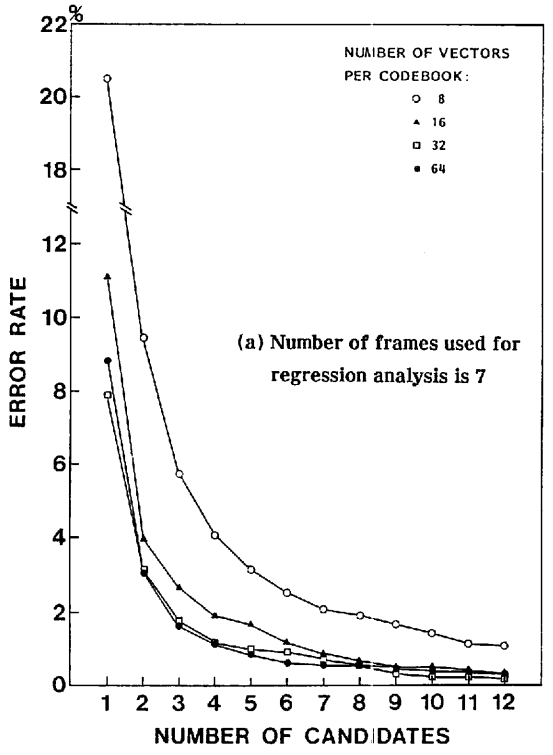


Fig. 4 - Comparison of the preprocessor performances for three feature parameter conditions (Method B-2: Preselection by biased distance threshold).



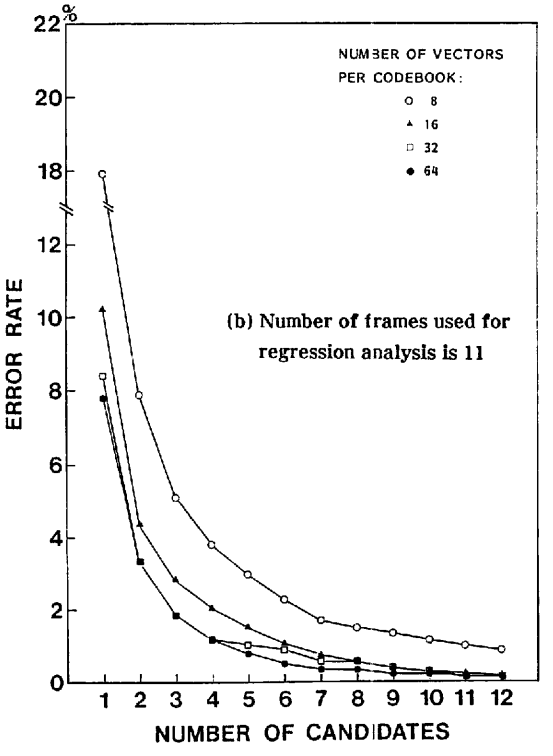


Fig. 5 - Comparison of the preprocessor performances for four codebook-size conditions (Method A: Preselection by distance rank order).

帰係数を求めるフレーム数が7フレームのときに最も高い認識性能を示す。以上の結果から、以後の検討ではいずれもフレーム数を7に定めて実験を行なった。

図6は方法B－2を用いて、符号帳のサイズを変えたときの結果を示したものである。この場合も、符号帳サイズが32と64では結果の差は小さい。以上の結果から、以後の実験では符号帳サイズは32に固定した。

図7は方法B－1を用いたときの、θ_0と予備選択誤り率および選択される単語数の割合の変化を示したものである（符号帳サイズは32）。図6の縦軸のスケールを比較するとわかるように、方法B－1では選択効率が極めて悪い。例えば$\theta_0=1.2$の場合、選択される単語は100単語中平均9.9単語、この中に正しい単語が含まれる割合は92.4%にすぎない。このため以下の実験では、方法B－1は用いず、方法Aと方法B－2のみ検討した。

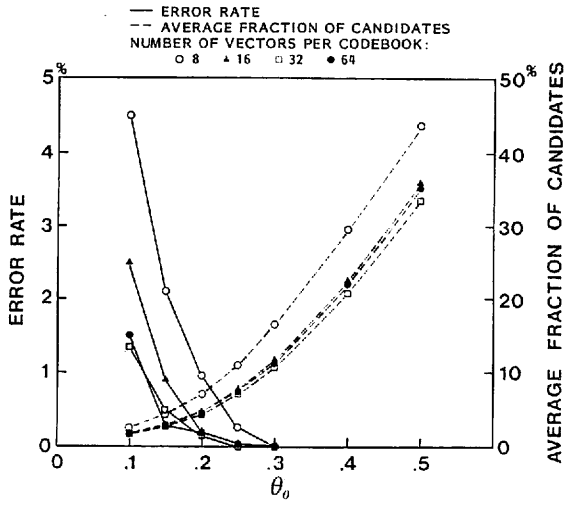


Fig. 6 - Comparison of the preprocessor performances for four codebook-size conditions (Method B-2: Preselection by biased distance threshold; Number of frames used for regression analysis is 7).

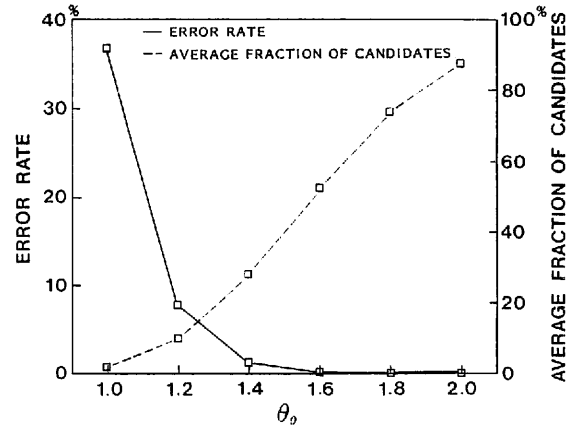


Fig. 7 - Preprocessor performance for Method B-1: Preselection by unbiased distance threshold. Number of vectors per codebook is 32.

4．単語間に共通の符号帳を用いる方法

4．1．方法

上の実験では、対象語彙毎に独立に符号帳を作成して予備選択を行なったが、この方法では対象語彙毎に入力音声の各フレームと符号帳の各要素ベクトルとの距離を計算しなければならないので、語彙数が大きくなるとそれに比例して計算量が増加する問題があり、この方法では、後段のDTWまで含めた計算量の削減は極めてわずかにしかならない。このため、ここでは全対象語彙に共通の符号帳（共通符号帳）を設け、入力音声に対してはこの共通符号帳の各要素ベクトルとの距離を計算する方法について検討した。各単語を表現する符号帳は、この共通符号帳の部分集合とすれば、各単語の符号帳によるベクトル量子化の歪量は、ＳＰＬＩＴ法の場合と同様に、共通符号帳の各要素との距離を距離行列の形で蓄えておいて、これを参照することによって簡単に求めることができる。この方法のブロック図を図8に示す。

前節までの方法では、図9(a)に示すように入力音声の各フレームとすべての単語の符号帳要素との距離計算をする必要があったが、本節の方法では、同図(b)に示すように共通符号帳の要素との距離計算をすればよいので、Ｌを各単語の符号帳サイズ、Ｍを共通符号帳のサイズとすると、距離計算量の比は

Ｍ／（Ｌ×１００）

となる。Ｌ＝３２、Ｍ＝１０２４のとき、比は１０２４／３２００≃１／３　であるが、語彙数が大きくなると、この比は語彙数に反比例して小さくなる。

共通符号帳は、一たん各単語毎に独立に作成した符号帳の要素ベクトルを全部集めて、再度クラスタ化することによって作成した。各単語の符号帳は、最初に各単語別に作成した符号帳の各要素に最も近い共通符号帳の要素を選んで作成した。共通符号帳のサイズは、５１２と１０２４の２種類の場合について検討した。

4．2．実験結果

図１０は方法Ａ、図１１は方法Ｂ－２の場合の予備選択の結果を示したものである[11]。比較のために単語別に符号帳を作成する場合の結果をあわせて示した。これらの結果からわかるように、共通符号帳のサイズを１０２４程度にすれば、単語別に符号帳を作成する場合

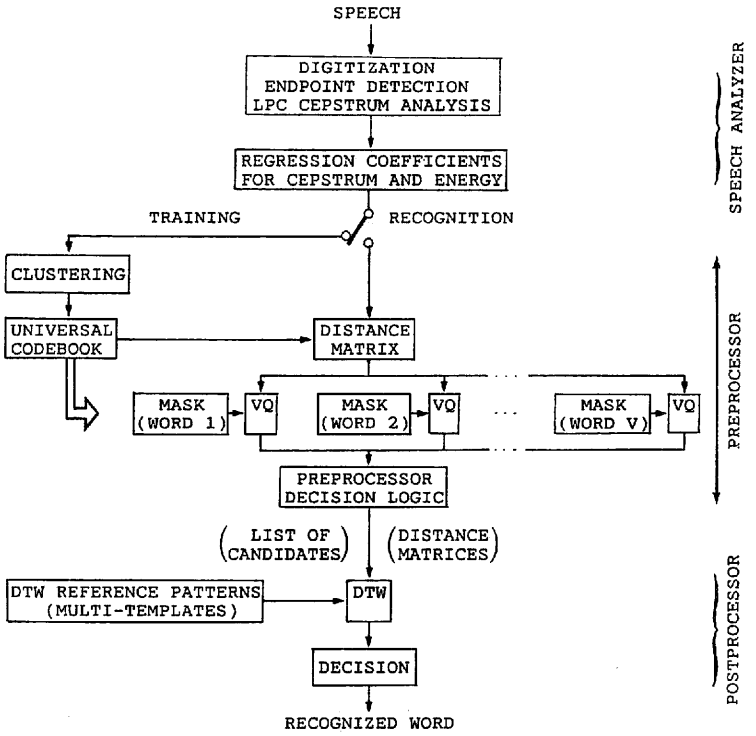


Fig. 8 - Block diagram of isolated word recognizer incorporating a VQ preprocessor based on a universal codebook and a SPLIT postprocessor.

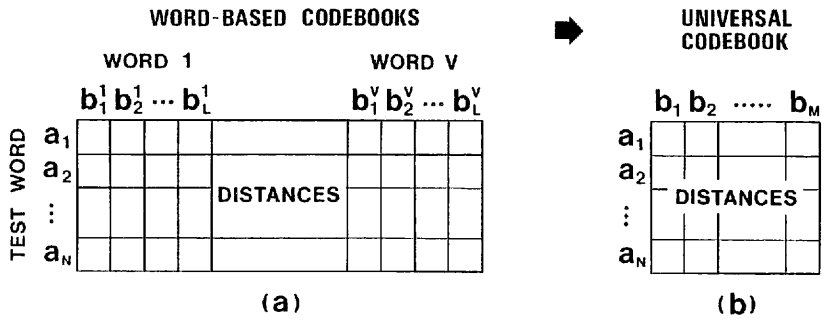


Fig. 9 - Distance computation in the word-based VQ preprocessor and the universal VQ preprocessor.

とほとんど変わらない予備選択性能を得ることができる。符号帳サイズを１０２４とし、方法Ａを用いて第１０位まで選択するときに正しい単語が予備選択される割合は９９．６％、方法Ｂ－２を用いて、θ_0＝０．２としたときに選択される単語数は平均６．８単語で、正しい単語が含まれる割合は９９．９％である。また２０００サンプル中４９６サンプルは、予備選択で候補単語が正しい１単語に絞られた。

５．ＤＴＷとの組み合わせによる認識

予備選択された候補単語に関してＤＴＷを行なって最終的に得られる認識性能を、予備選択を行なわずに全候補単語とＤＴＷを行う場合と比較した。ＤＴＷは、１０２４の要素からなる共通符号帳を用いるＳＰＬＩＴ法によって行ない、各単語のＤＴＷ用標準パターンは４名の話者の学習サンプルを共通符号帳の要素の系列で表わしたものを用いた。このため、入力音声と共通符号帳の距離行列をそのままＤＴＷにも用いることができる。ＤＰ演算には、Staggered Array法[8]を用いた。種々の実験条件の場合の認識結果をまとめて表１に示す[11]。共通符号帳を用い、方法Ｂ－２でθ_0＝０．２としたとき、ＤＴＷは平均して６．８％の候補単語とのみ行なえばよく、最終的に予備選択を行なわないときと同じ２．０％の単語認識誤り率を得ることができる。このときに必要な計算量は、予備選択を行なわない場合の約１／１０である。

６．ベクトル量子化方法に関する検討

これまでの方法では、ケプストラム係数とケプストラム係数及び対数エネルギーの回帰係数（Ｃ，Ｃ'，Ｅ'）をセットとしてクラスタ化を行なった。一方文献[9]では、ケプストラム係数（Ｃ）とその回帰係数（Ｃ'）を別々にクラスタ化して、それぞれの符号帳でベクトル量子化したときの歪量の重み付き加算により、話者認識の判定を行なっている。ここでは両者の比較を行なった。簡単のため、単語毎に符号帳を作成する場合について実験を行なった。

図１２(a)は、（Ｃ，Ｅ'）をベクトル量子化したときの歪量ｄ（Ｃ，Ｅ'）と（Ｃ'）をベクトル量子化したときの歪量ｄ（Ｃ'）について、

$$d_1=\alpha d(C, E')+d(C')$$

を計算して、これによって第１位のみを選択したときの誤り率を示したものである。同様に図１２(b)は、

$$d_2=\beta d(C', E')+d(C)$$

によって、第１位のみを選択したときの誤り率を示したものである。いずれの歪量も平均値が等しくなるように規格化してある。これらの結果からわかるようにd_1による誤り率は、αを最適値に選べば

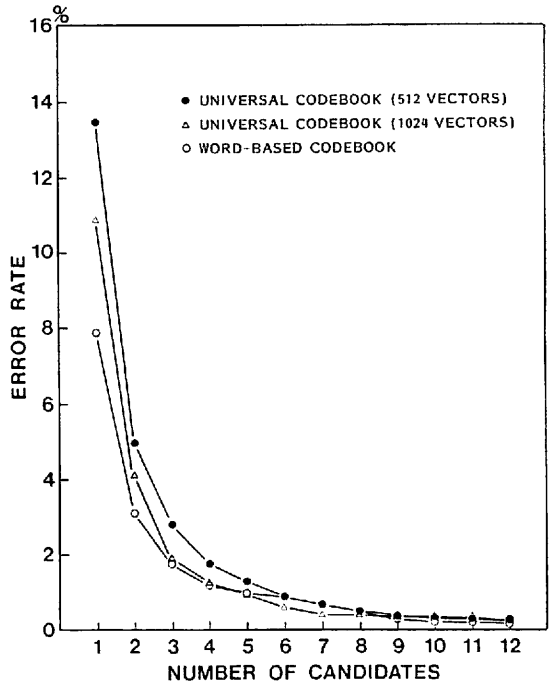


Fig. 10 - Comparison of the preprocessor performances for three codebook conditions (Method A: Preselection by distance rank order).

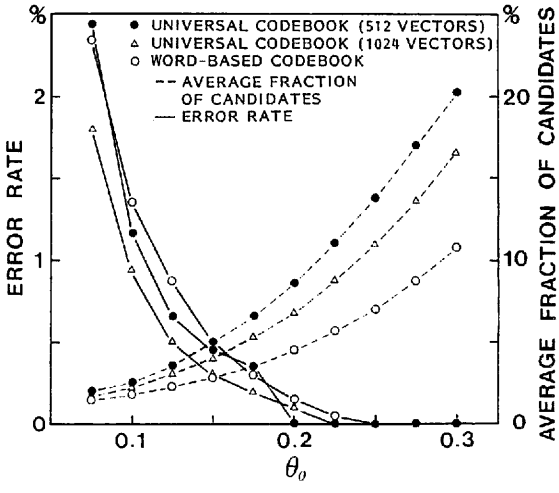


Fig. 11 - Comparison of the preprocessor performances for three codebook conditions (Method B-2: Preselection by biased distance threshold).

		OVERALL ERROR RATE (%) PRESELECTION ALONE	OVERALL ERROR RATE (%) PRESELECTION & DTW	NUMBER OF CANDIDATES FOR DTW
DTW ALONE		-	2.0	100
WORD-BASED CODEBOOK	PRESELECTION BY			
	TOP	7.9	-	-
	TOP TEN	0.2	2.1	10
	$<\theta$ (θ_0=.2)	0.2	2.1	4.5
UNIVERSAL CODEBOOK	PRESELECTION BY			
	TOP	10.9	-	-
	TOP TEN	0.4	2.3	10
	$<\theta$ (θ_0=.2)	0.1	2.0	6.8

Table. 1 - Experimental results for the word recognizer incorporating a VQ preprocessor and a DTW-based postprocessor.

d(C, E')又はd(C')単独の場合よりも小さくなるが、その値は(C, C', E')を一括してベクトル量子化したときの誤り率(7.8%)の2倍程度に大きい。またd2では、d(C', E')のみのときが最も誤り率が小さく、d(C)を組み合わせる効果は全く見られない。以上の結果から、(C, C', E')はセットとして一括してベクトル量子化することが必要なこと、言い換えるとセットとして音韻性情報を担っていることがわかる。

7. むすび

不特定話者の大語彙単語音声認識のための、候補単語予備選択方法について検討した。各単語をケプストラム、およびケプストラムと対数エネルギーの線形回帰係数からなるベクトル系列のクラスタ化にもとづく符号帳(各単語の符号帳サイズは32ベクトル)で表現し、この各要素は全単語の共通符号帳(サイズは1024)の要素で代用する。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歪量にもとづいて、候補単語の予備選択を行ない、選択された候補について、共通符号帳とマルチテンプレートを用いたSPLIT法によって最終の単語決定をおこなう。

男性20名の100都市名単語音声の認識実験の結果、この方法により、平均6.8語に候補を絞ることができることがわかった。このとき、予備選択を行なわない方法に比べて、演算処理量を約1/10に低下させることができ、しかも認識率は低下しない(平均98.0%)ことが確認された。本方法で用いている動的特徴は、文献[5]で用いている時間情報に比べて、比較的局所的な時間情報を表現している。ここで用いた予備選択方法は、単語音声の定常部で時間軸が伸縮されている限り全く歪量の増大をもたらさないという意味で、音声の伸縮特性とよく対応している。

今後さらに大語彙を対象にした評価実験を行なっていきたい。

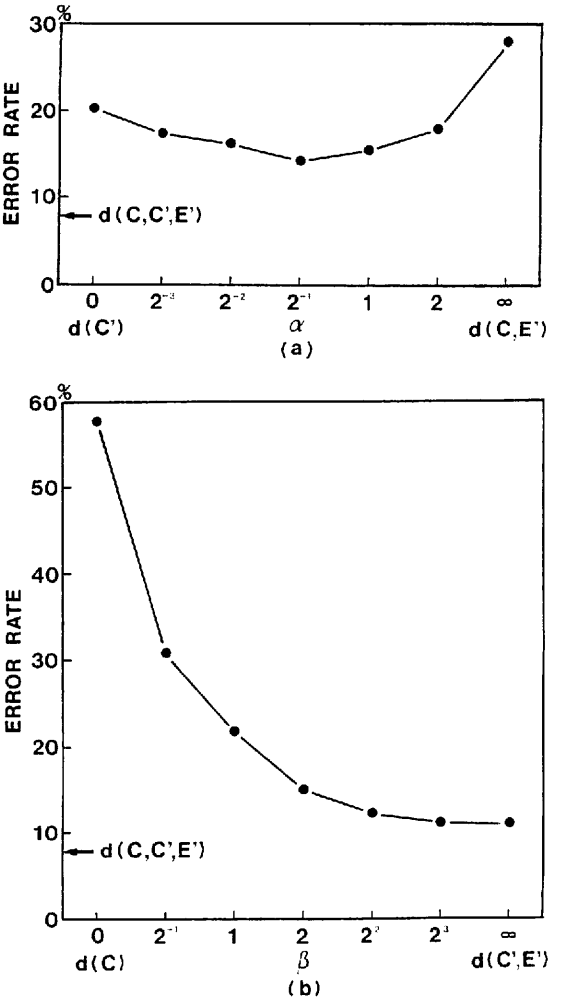


Fig. 12 - Preprocessor performance versus the factor or for combining two distortions associated with separate VQ codebooks.

文献

[1] Rabiner,L.R. Levinson,S.E. Rosenberg,A.E. and Wilpon,J.G.: "Speaker-independent recognition of isolated words using clustering techniques", IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-27,4,pp.336-349 (1979)

[2] Rabiner,L.R. Levinson,S.E. and Sondhi,M.M.:"On the application of vector quantization and hidden Marcov models to speaker-independent, isolated word recognition", Bell Syst. Tech.J.,62,4,pp.1075-1105 (1983)

[3] 千葉、亘理、渡辺:"音声認識システム"、大型プロジェクト・パターン情報処理システム・研究開発成果発表会論文集、pp.157-165 (昭55)

[4] Pan,K-C. Soong,F.K. and Rabiner,L.R.:"A vector-quantization-based preprocessor for speaker-independent isolated word recognition", IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-33,3,pp.546-560 (1985)

[5] Bergh,A.F. Soong,F.K. and Rabiner,L.R.:"Incorporation of temporal structure into a vector-quantization-based preprocessor for speaker-independent, isolated-word recognition", AT&T Tech.J.,64,5,pp.1047-1063 (1985)

[6] Furui,S.:"Cepstral analysis technique for automatic speaker verification", IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-29,2,pp.254-272 (1981)

[7] 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 (1986)

[8] Furui,S.:"Speaker-independent isolated word recognition based on emphasized spectral dynamics", Proc.IEEE Int.Conf. Acoust.,Speech,Signal Processing,Tokyo,37.10,pp.1991-1994 (1986)

[9] Soong,F.K. and Rosenberg,A.E.:"On the use of instantaneous and transitional spectral information in speaker recognition", Proc.IEEE Int.Conf.Acoust.,Speech,Signal Processing,Tokyo,17.5,pp.877-880 (1986)

[10] Linde,Y. Buzo,A. and Gray,R.:"An algorithm for vector quantizer design", IEEE Trans.Commun.,COM-28,1,pp.84-95 (1980)

[11] 古井貞熙:"大語彙単語音声認識のためのスペクトル動特性を用いた予備選択法"、音学秋季講論、1-3-17 (1986)