論文 / 著書情報
Article / Book Information

論題(和文)	音声スペクトルの動的特徴を用いた単語音声認識
Title	
著者(和文)	古井貞熙
Author	SADAOKI FURUI
出典(和文)	日本音響学会 音声研究会資料 S84-65, , , pp. 509-516
Journal/Book name	, , , pp. 509-516
発行日 / Issue date	1984, 12

日本音響学会音声研究会資料 S84-65(December 25, 1984)

音声スペクトルの動的特徴を用いた単語音声認識

古井 貞熙
(日本電信電話公社武蔵野電気通信研究所)

内容梗概：音声スペクトルの静的特徴と動的特徴を組み合わせた単語音声認識系を提案し、不特定話者の音声認識のように、標準パタンと入力音声の話者が異なる場合に、この方法の有効性が高いことを示した。 具体的には、音声波をまずケプストラムとパワーの時系列で表現し、フレーム周期毎にそのフレームを中心とする約50msの区間の各パラメータ時系列の線形回帰係数を抽出する。 回帰係数の時系列を元のケプストラムの時系列とあわせて特徴パラメータ時系列とし、マルチテンプレートと入力音声とのStaggered Array DPマッチングによって認識を行う。 100都市名の不特定話者単語音声認識の結果、ケプストラムのみのときに6.2%であった誤認識率が、本方法を用いると2.4%に低下することがわかった。

Transactions of the Committee on Speech Research
The Acoustical Society of Japan, S84-65(December 25, 1984)

SPOKEN WORD RECOGNITION USING DYNAMIC FEATURES OF SPEECH SPECTRUM

Sadaoki FURUI
(Musashino Electrical Communication Laboratory,N.T.T.)

Abstract:This paper proposes a new isolated word recognition technique based on the combination of temporal and dynamic features of speech spectrum. This technique is highly efficient in speech recognition when the speakers of input speech and reference templates are different such as the case of speaker independent speech recognition. Spoken utterances are represented by the time sequences of cepstrum coefficients and energy, and regression coefficients for these time functions are extracted for every frame using nearly 50ms period. Time functions of regression coefficients extracted for cepstrum and energy are combined with the time functions of cepstrum coefficients, and used for Staggered Array DP matching of multiple templates and input speech. Speaker independent isolated word recognition experiments using the vocabulary of 100 Japanese city names indicate that the recognition error rate of 2.4% can be obtained by this method, whereas the error rate obtained by using only cepstrum coefficients is 6.2%.

1．まえがき

我々人間が音声を知覚するとき、そのスペクトルの静的特徴（各時点でのスペクトルパタン）だけでなく、その動的特徴（各時点でスペクトルパタンがどのように変化しつつあるかという特徴）が重要な役割を果していることはよく知られている[1]。筆者らの最近の話頭／話尾切断した単音節を用いた聴覚実験によれば、すべての単音節に共通して、スペクトルの時間変化量が極大となる時点に最も重要な音韻的特徴が含まれていることが確認されている[2,3]。

しかしこれまでに、スペクトルの動的特徴を音声認識に直接的に用いた例は少ない。その一つとして筆者らは、話者認識において、ケプストラムの各次数の時系列から10ms毎に90ms（9フレーム）の小区間をとりだし、1～2次の多項式展開係数の時系列を求めて、ケプストラム時系列と組み合わせて認識を行う方法を検討し、この有効性を確認している[4,5]。その後ケプストラムの1次の展開係数、即ち線形回帰係数を破裂時点付近でのケプストラムの時系列から求めて、破裂音の識別に用いる試みが山下らによって行われており[6]、各フレームを二つに分けて、ケプストラムとパワーの前半と後半のフレームにおける差を非定常性の情報として用いる試みも、間野らによって行われている[7]。

ここでは上述の、筆者らが以前に話者認識に用いた方法を改良して、マルチテンプレート方式による不特定話者の単語音声認識に適用した試みについて述べる。

2．単語音声認識系の構成

2．1．分析部

認識系の構成を図1に示す。入力された音声波に対して、まず標本化周波数8kHzで標本化し、短区間パワーの変化をもとに雑音区間と音声区間の分離を行う。ただし、後に端点フリーのDPマッチングを行うので、音声区間の切り出しは必ずしも厳密に行う必要はなく、むしろ音声区間の両端に雑音区間をつけて切り出しておく。次にフレーム周期8msで32msのハミング窓を乗じたのちLPC分析を行い、1～10次の線形予測係数とパワーの時系列を得る。聴覚特性との整合性を考慮して、線形予測係数はLPCケプストラムに、パワーは対数パワー（以下では簡単のため、それぞれ単にケプストラム及びパワーと呼ぶ）に変換する。

2．2．線形回帰係数の時系列の抽出

ケプストラムとパワーの8ms（フレーム周期）毎の時系列から、フレーム周期毎に7フレーム（56ms）の区間を切り出し（7フレーム分のバッファを用意しておいて、フレーム周期毎に古い1フレームのデータを捨て、新しい1フレームのデータを加える）、この区間における各次数のケプストラムおよびパワーの時系列に関して、1次の多項式展開係数、即ち線形回帰係数を次のようにして計算する。$x_i(t)$をパラメータ（$x_0(t)$：パワー、$x_1(t)$～$x_{10}(t)$：ケプストラム）の時系列とするとき、時刻tを

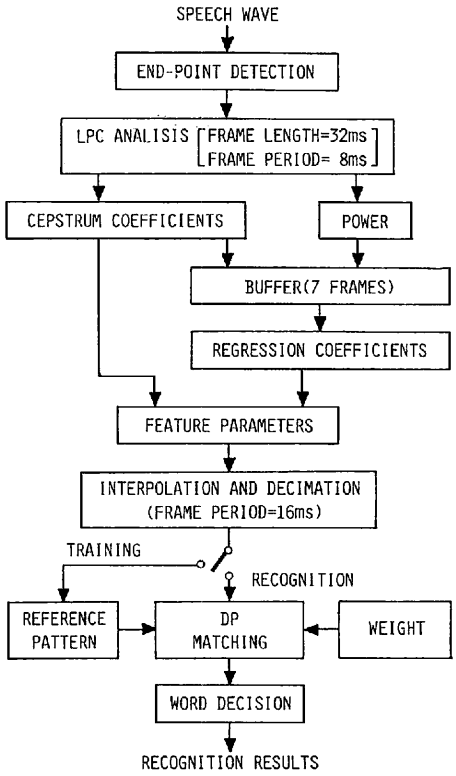

Fig. 1-Block diagram of an isolated word recognition system using dynamic features of spectrum.

時間軸の原点に取り直した時系列を$x_i(n)$とする。これに関して次のようにして線形回帰係数を計算する。

$$a_i(t)=\left(\sum_{n=-n_0}^{n_0} x_i(n)\cdot n\right)\Big/\left(\sum_{n=-n_0}^{n_0} n^2\right) \qquad (1)$$

線形回帰係数の抽出に用いるフレームの数は、後述するように、実験的な最適値として7フレームに選んだ。従って、$n_0=3$である。

予備実験では、2次の展開（回帰）係数まで求めて検討したが、単語音声認識における有効性は低いため、1次の係数のみを用いることにした。

こうして$a_i(t)$がフレーム周期毎に抽出されるので、パワーの時系列を除いたケプストラムのみの時系列$x_i(t)$を$a_i(t)$とあわせて、特徴パラメータ時系列とする。(1)式から明らかなように、音声区間の始端と終端のそれぞれn_0フレームに関しては、$a_i(t)$が定義できないので、この区間はこの段階で$x_i(t)$からも除去している。ただし、2．1節で述べたように、真の音声区間の両端に雑音区間をつけて切り出しが行われているので、真の音声区間が除去されるおそれはない。

次に認識部での処理量を減らすため、8ms毎のパラメータ時系列に対して、2フレーム毎の値を平均して平滑化し、フレームの間引きを行って16ms毎の時系列に変換する。

2．3．認識部

ケプストラムと，ケプストラム及びパワーの回帰係数の時系列を用いて、DPマッチングをベースとする認識を行う。本システムはマルチテンプレート方式の不特定話者単語音声認識を想定しているので，学習時には各単語毎にあらかじめ多数話者から選択した複数の話者の音声サンプルを標準パタンとして登録しておく。次に認識を行う場合は，各標準パタンを次々に読み出して，入力音声とのDPマッチングを行う。ここでは，DPマッチングの方法として，始端と終端の両端点フリーが実現でき，計算量が比較的少ないStaggered Array法[8]（厳密にはDP3－5の方法）を用いた。この際，標準パタンのkフレームと入力音声のlフレームとの距離D(k,l)を次のように定義する。

$$D(k,l)=\{w_1\sum_{i=1}^{10}(x_i^R(k)-x_i^I(l))^2+w_2(a_0^R(k)-a_0^I(l))^2+w_3\sum_{i=1}^{10}(a_i^R(k)-a_i^I(l))^2\}/(\sum_{j=1}^{3}w_j) \quad (2)$$

ここで Rは標準パタン、Iは入力音声を示す添字である。重み係数$\{w_j\}_{j=1}^{3}$の値は、各特徴パラメータの有効性に従って、予備実験によりあらかじめ決定しておく。

各標準パタンとのDPマッチングの結果は単語決定部に送られ、単語の決定がされる。単語の決定は，ここでは簡単のため、1単語当りの標準パタンが一つの場合も複数の場合も、入力音声から最も距離の小さい標準パタンを強制的に選ぶことによって行う。

2．4．音声データ

不特定話者単語音声認識を想定した本方式の有効性を実験的に確認するため、単語セットとしては都市名100単語を選び、次の2種類の音声データを用意した。これらの音声はいずれも計算機室（室内騒音約70dB(A)）で発声され、ダイナミックマイクロフォンで収録されている。

（1）音声データ1

男性4名が100単語を，通して2回発声した音声データである。この4名は、男性話者の音声の分布を代表する話者として、SPLIT法にもとづく不特定話者単語音声認識[9]においてマルチテンプレートを決定した際に、そのテンプレートの構成に比較的多く用いられた話者を選んだ。ただし、SPLITマルチテンプレート法で、男女50名の音声から1単語当り1テンプレートを選択した時に，これら4名の話者の音声がそれらのテンプレートに含まれる割合は，25．8％に過ぎない。

この男性4名の音声を用いた認識実験では，後の時期の音声を入力音声として，前の時期のいずれか1名の話者の音声を標準パタン（各単語について標準パタン1通り）とするすべての組み合わせ（16通り、うち同一話者4通り、異話者12通り）について実験を行った。認識サンプル数は同一話者内400サンプル、異話者間1200サンプルである。

（2）音声データ2

音声データ1の4名の話者が前の時期に各単語を1回ずつ発声した音声と、この4名と異なる男性20名が後の時期に各単語を1通り発声した音声からなる音声データである。前者の音声をすべて，即ち各単語について4名の音声をすべてマルチテンプレートの標準パタンとして登録し、後者の男性20名の音声を入力音声とする認識実験を行った。認識サンプル数は2000サンプルである。

3．微分対数スペクトル包絡とパワー変化

ケプストラムは対数スペクトルをフーリエ逆変換したものであるから，低次のケプストラムの回帰係数をフーリエ変換すると、対数スペクトル包絡の単位時間当りの変化量、即ち微分対数スペクトル包絡を得ることができる。図2は、2人の話者が発声した

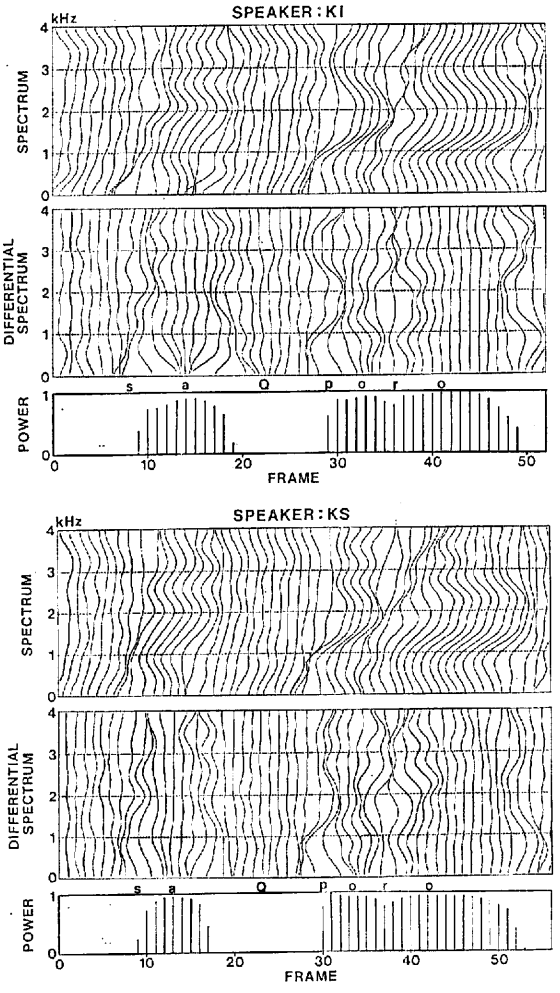

Fig. 2-Time sequences of spectrum envelope, differential spectrum envelope and power for a word /saQporo/ uttered by two male speakers.

単語「札幌」のパワー、対数スペクトル包絡、及び微分対数スペクトル包絡の時系列を示したものである。微分対数スペクトル包絡では、対数スペクトル包絡の動的特性が直接的（explicit）に表現されており、もとの対数スペクトル包絡では必ずしも明確に示されていない／r／の特徴等が明確に見られ、しかもこれらの動的特徴が、2人の話者で共通していることがわかる。

パワーの回帰係数はこの図には示されていないが、パワーの時系列から容易に推察されるように、パワーの回帰係数は、異なる話者に共通な単語のマクロ的特徴として有効であり、従来のパワーの時系列そのものを用いる方法[10,11]と違って、音声区間全体におけるパワーの最大値、最小値等で正規化しなくても、発声レベルの変動の影響を受けないため、発声区間の終了を待たずに、時間遅れなしで特徴抽出ができるという長所がある。厳密には、8msを単位とする3フレーム分の遅れが必要であるが、これは無視できる遅れであると言える。

なお、浅川ら[12]によって、パワーのフレーム間差分値からパワーが増加中か減少中かの変化傾向を抽出し、この傾向が一致しないときにはスペクトル距離に一定のペナルティーを乗ずるという方法も試みられているが、この方法では変化の符号のみが用いられ、その大きさは考慮されていない。

4．ケプストラムの回帰係数の有効性

4．1．回帰係数のみによる認識（音声データ1）

ケプストラムの回帰係数の抽出に用いる小区間の長さを、3、5、7、9、11フレームの5通りに変えて、回帰係数時系列が単独に有する認識能力を調べた。これは(1)式において、 $w_1=w_2=0$、$w_3=1$ とした場合に相当し、ケプストラムの回帰係数時系列のDPマッチングによって認識を行った。

音声データ1を用いた認識実験の結果（誤認識率）を、標準パタンと入力音声が同一話者の場合と異話者の場合に分けて、図3に示す。この結果から、回帰係数の抽出に用いるフレーム数は、3～5フレーム（24～40ms）では短かすぎ、9フレーム（72ms）でほぼ最適となることがわかる。比較のために$w_1=1$、$w_2=w_3=0$、即ちケプストラム時系列のDPマッチングによる場合の認識実験を行ったところ、結果は図の右の欄外に示すようになり、異話者間の認識で、回帰係数を求めるフレーム数が7フレームのときは、回帰係数とケプストラムによる認識結果はほぼ等しく、9フレームにした場合は、ケプストラムよりもむしろ回帰係数を用いる方が誤認識率がやや小さくなる。

4．2．ケプストラムと回帰係数の組み合わせ（音声データ1）

$w_1=1$、$w_2=0$ として、w_3の値、即ちケプストラム回帰係数の距離に対する重みを種々に変えたときの、ケプストラムと回帰係数の組み合わせによる認識の結果を図4に示す。この図には、回帰係数の抽出に用いるフレーム数を7と9とした2通りの場合の結果を示してある。これから次のことがわかる。

（1）回帰係数の抽出に用いるフレーム数は、7と9の場合ではとんど認識結果に差がない。

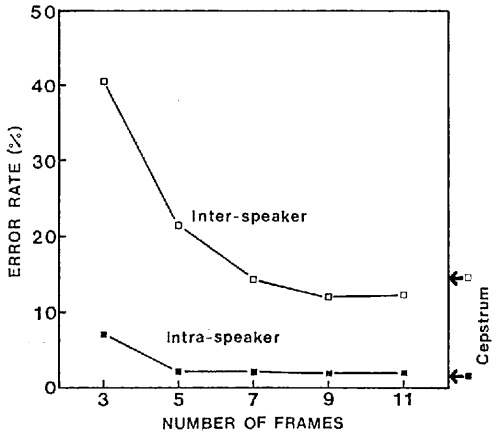

Fig. 3-Relation between number of cepstrum frames for regression analysis and recognition error rates obtained by using Database 1. Right outside is shown the error rates for the recognition by cepstrum coefficients.

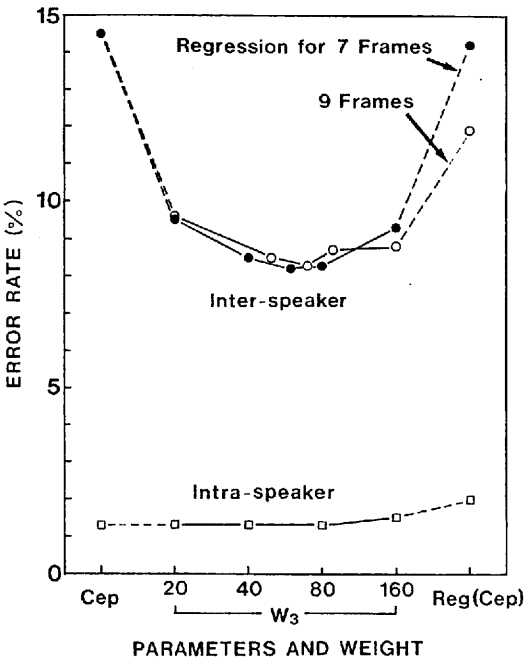

Fig. 4-Error rates for various conditions of feature parameters. Left: Cepstrum (Cep). Right: Regression coef. for cepstrum (Reg(Cep)). Middle: Combination by weight w_3. Reg(Cep) are extracted from 7 or 9 frames, and dimensions of Cep and Reg(Cep) are both fixed to 10.

（2）異話者間の認識の場合、7フレームで回帰係数を抽出し、重みw_3を適切に設定すれば、ケプストラムのみの場合及び回帰係数のみの場合に比べて、誤認識率を14．2～14．5％から8．2％へ、約3／5に減らすことができる。

（３）重みw_3の変動に対する誤認識率の変動は小さく、安定している。

（４）同一話者内の認識では、ケプストラムのみでも誤認識率が極めて小さいので、回帰係数を組み合わせる効果は無い。

図5は、異話者間の認識において、ケプストラムは１０次まで用い、回帰係数もケプストラムと同じく１０次まで用いた場合（ｐ＝１０）と、比較的マクロなスペクトル特徴の動特性に絞って7次までで打ち切った場合（ｐ＝７）の結果を比較して示したものである。これから、回帰係数もケプストラムと同じく１０次まで用いることが必要であることがわかる。

以上の結果から、回帰係数も１０次まで用いることとし、回帰係数の抽出に用いるフレーム数は、７フレームとすることにした。

なお、上述の実験において、異話者間の認識における１．０％、同一話者内の認識における０．５％の誤りは、もともと標準パタンと入力音声の音声区間長が大きく異なっていて、ＤＰマッチングができないための誤りである。

5．パワーの回帰係数の有効性（音声データ１）

前節までの結果にもとづき、ケプストラムに対するその回帰係数の距離の重みを最適（$w_1=1$、$w_3=60$）に設定して、パワーの回帰係数の距離に対する重みw_2を変えたときの誤認識率の変化を図6に示す。図の左端は、$w_2=0$即ちケプストラムとその回帰係数のみの場合の認識結果、右端はw_2を最適にしたときのケプストラムとパワーの回帰係数による認識の場合（$w_1=1$、$w_2=10$、$w_3=0$）の結果である。ケプストラムのみの場合の誤認識率は１４．５％であるので、ケプストラムに最適な重みでパワーの回帰係数を組み合わせることによって、誤認識率が３．８％低下するが、ケプストラムとその回帰係数を組み合わせたときの方が、効果が大きい（６．３％の低下）。後者の組み合わせに、さらにパワーの回帰係数を組み合わせることによって誤認識率が０．４％低下する。

6．マルチテンプレートによる不特定話者単語音声認識（音声データ２）

6．1．種々の組み合わせによる誤認識率の改善

音声データ１を用いた４章及び５章の結果から、ケプストラム、ケプストラムの回帰係数、及びパワーの回帰係数を組み合わせるときの重みの最適値が明らかとなったので（$w_1=1$、$w_2=10$、$w_3=60$）、この条件における音声データ２を用いた認識実験を行った。すでに述べたように、標準パタンとしては男性４名の１００単語音声をすべて蓄え、これ以外の男性話者２０名の音声の認識実験を行った。認識に用いるパラメータの組み合わせを種々に変えたときの誤認識率の変化を図7に示す。なおパワーの回帰係数のみによる認識では、誤認識率が極めて大きくなるので、この実験は行っていない。

音声データ１を用いた場合の結果と同様に、ケプストラムを用いるよりもむしろその回帰係数を用いる方が誤認識率がやや小

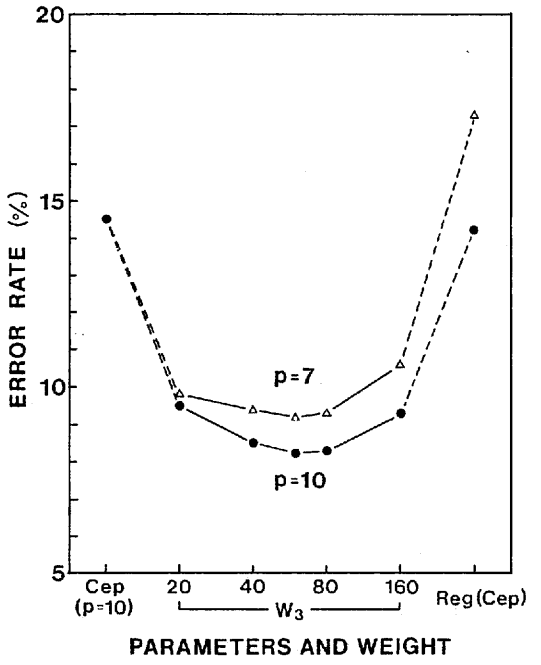

Fig. 5-Effects of the dimension of regression coef. for cepstrum(p) on the word recognition error rates under the inter-speaker matching condition.

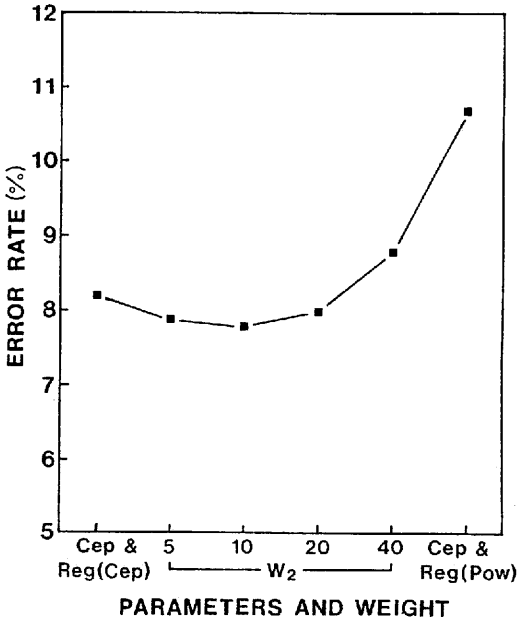

Fig. 6-Error rates for various conditions of feature parameters. Left:Cepstrum(Cep) and their regression coef.(Reg(Cep)). Right:Cep and regression coef. for power(Reg(Pow)). Middle: Combination of Cep, Reg(Cep) and Reg(Pow) by weight w_2(w_1=1, w_3=60).

さく、ケプストラムにその回帰係数を組み合わせることにより、誤認識率が6．2％から3．1％へ1／2に低下する。ケプストラムにパワーの回帰係数を組み合わせたときの誤認識率は3．8％で、音声データ1の場合と同様に、パワーの回帰係数よりもケプストラムの回帰係数の方が効果が大きい。ケプストラムとケプストラム及びパワーの回帰係数をすべて組み合わせたときの誤認識率は2．4％で、ケプストラムのみの場合の誤認識率の約2／5である。

6．2．認識率の改善の分析

ケプストラムのみの場合、ケプストラムとその回帰係数を組み合わせた場合、及びそれにさらにパワーの回帰係数を組み合わせた場合の、話者による誤認識数の分布を図8に示す。ケプストラムのみの場合は、100個の音声サンプル中14、17、18サンプルを誤認識する話者があるが、ケプストラムにその動的特徴である回帰係数を組み合わせると、このような誤りの極端に多い話者がなくなり（最大で8サンプル）、これにパワーの動的特徴を組み合わせると、全般に誤認識数の分布がさらに小さい方へ寄る（最大でも6サンプル）ことがわかる。音声認識において、全体の話者の中ではわずかな割合ではあるが、極端に誤りの多い話者を生ずる現象は一般に”Sheep and Goats現象”と呼ばれており、音声認識システムを実現する上で重要な問題の一つである。スペクトルの静的特徴と動的特徴を組み合わせることによって、平均的な誤認識率が低下するだけでなく、極端に誤りの多い話者がなくなることは、極めて望ましいことであると言えよう。

次に誤認識を生じた音声サンプルについて、その単語の音素表記と誤認識結果の音素表記を用いて、どのような誤りを生じたかを、音素を単位とするLevenshtein Distance（LD、一方の記号系列から他方の記号系列へ変換するときに必要な置換（r）、挿入（i）、脱落（d）の数の和を表わしたもの）により分析した。

$$LD(A \rightarrow B) = min(r + i + d) \qquad (3)$$

ここではLDをさらに選ばれた標準パタンの音素数で割った値を求め、この値を全誤認識サンプルについて平均した値を求めた。その結果、ケプストラムのみを用いたときの平均値は0．57であるのに対し、ケプストラムとその回帰係数を組み合わせたときの平均値は、0．47に低下することがわかった。さらに音素でなくモーラを単位とし、置換（r）を除いた挿入、脱落のみによる評価を行ったところ、モーラ数で正規化し全誤認識サンプルで平均した値は、ケプストラムのみのとき0．55であったのに対し、その回帰係数を組み合わせると、0．45に低下することがわかった。これらの結果は、スペクトルの静的特徴に動的特徴を組み合わせると、DPマッチングの際に子音と母音を対応づけたり、音素が脱落したり挿入されたりする極端な誤りが減少することを示している。

静的特徴であるケプストラムのみの場合と、動的特徴であるケプストラムとパワーの回帰係数を組み合わせた場合に生ずる誤認識のうち、2回以上生じた単語の組み合わせは表1に示す通り

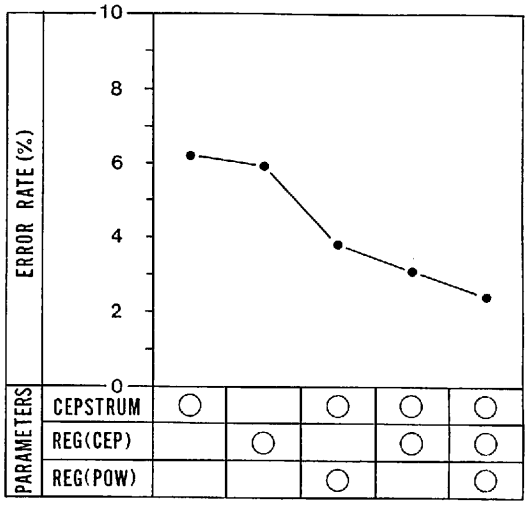

Fig. 7-Recognition results for five conditions of the feature parameter combinations, using Database 2. ○ indicates the parameter set used.

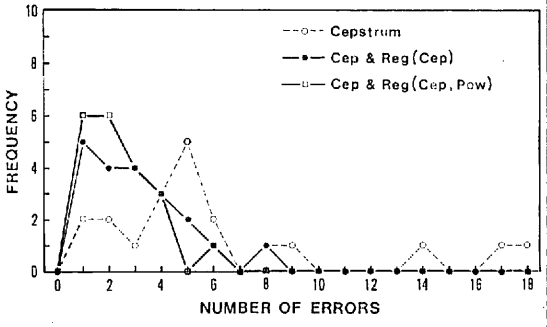

Fig. 8-Distribution of the number of errors in 100 test utterances for 20 speakers.

である。表の上部に静的特徴と動的特徴を組み合わせたときの誤りを多い順にまとめて示しており、ここに示したものが全話者の誤認識48サンプル中の2／3を占めている。この結果から、静的特徴のみを用いたときの誤認識の組み合わせには、聴覚的には理解できないような誤りもかなり含まれているが、動的特徴の組み合わせによってこれらが大幅に減り、聴覚的に比較的類似した組み合わせだけが残ることがわかる。

7．むすび

人間の聴覚特性の知見にもとづいて、音声スペクトルの静的特徴と動的特徴を組み合わせた単語音声認識系を提案し、不特定話者の音声認識の場合のように、標準パタンと入力音声の話者が異なる場合にこの方法の有効性が高いことを示した。具体的には、音声波をまずフレーム周期8msでLPC分析し、ケプストラムとパワーの時系列で表現する。次に56ms（7フレーム）の区間を8msずつずらしながら、各区間での1～10次の各ケプストラムとパワーの時間パタンの線形回帰係数を抽出し、回帰係数

の時系列を元のケプストラムの時系列とあわせて、特徴パラメータ時系列とする。演算量削減のため、フレームの補間と間引きを行って、フレーム周期を2倍の16msに変換し、標準パタンと入力音声とのStaggered Array法によるDPマッチングによって認識を行う。このとき、標準パタンと入力音声のDP平面上での各点における距離の計算において、ケプストラム、ケプストラムおよびパワーの回帰係数のそれぞれについて計算された値を、適切な重みを乗じて加算し全体としての距離とする。

この方法を用いた100都市名単語の認識実験の結果、男性4名の全単語の特徴パラメータ時系列を標準パタンとして、これらの話者と異なる男性20名の音声を認識する場合、ケプストラムのみを用いたときの誤認識率が6．2％であったのに対し、ケプストラムの回帰係数を組み合わせたとき3．1％、さらにパワーの回帰係数を組み合わせたとき2．4％に低下することが確認された。回帰係数で表現される動的特徴を静的特徴に組み合わせて用いると、誤りの極端に多い話者がなくなり、聴覚的類似性の少ない極端な単語間の誤りが減少することがわかった。

ケプストラムにこの回帰係数を組み合わせると、パラメータの数が分析次数の2倍となるが、ケプストラムのみで分析次数を20次程度に上げても、認識率は10次の場合よりほとんど向上しない。パラメータの総数を20より小さく抑えたときに、ケプストラムと回帰係数の最適な割合がどのようになるかは、まだ未検討である。

なお、ケプストラムの代わりにWLR尺度（スペクトルのピークのずれに重みをおいた距離尺度）[13]を用いて、これにケプストラムの回帰係数による距離を組み合わせる場合についても補足的実験を行っており、このときは、WLRのみの場合の誤認識率がケプストラムの場合よりも小さいので、ケプストラムの回帰係数を組み合わせる効果はやや小さいが、誤認識率が3．3％から2．6％へ約4／5に低下することを確認している。これにパワーの回帰係数を組み合わせれば、さらに誤認識率が低下すると期待される。ただし、WLR尺度とケプストラムの回帰係数の組み合わせは、パラメータの数が分析次数の3倍程度に大きくなる問題がある。パラメータの次数を20程度にする場合は、WLR尺度を用いるよりも、ケプストラムとその回帰係数を用いた方が誤認識率が小さい。

上で述べたように、ここでは回帰係数を安定に抽出する目的から、分析のフレーム周期は8msとし、認識時のフレーム周期は16msとしているが、分析のフレーム周期も16msとしたときに結果に差を生ずるか否かは、今後の検討課題の一つである。その他の今後の課題としては、さらに多数話者の音声を用いた不特定話者単語音声認識の実験による本方法の評価がある。このためには本方法をSPLIT法[9,14～16]に適用して、スペクトルの静的特徴と動的特徴をセットでベクトル量子化し、演算処理量と記憶容量の削減をはかるのが望ましい。またここで提案した方法は、連続音声中からあらかじめ決めておいた単語だけを自動的にみつけだすワードスポッティング等にも有効であると推定される。動的特徴には伝送系の影響を受けにくい長所もある[5]。図2に示した微分対数スペクトル包絡は、スペクトル包絡の動特性を視覚的にとらえる方法として、有効な手段となり得ると予想される。

Table 1-Comparison of the frequency of major confusion word pairs under the two feature parameter conditions, [ ]:Devocalized, N#:Frequency;
{N1:Temporal features
N2:Combination of temporal & dynamic features
N3:N2-N1

WORD PAIRS	FREQUENCY		
	N1	N2	N3
KAWASA[KI] - AMAGASA[KI]	6	10	+4
NARA - NAHA	5	4	-1
ICHINOMIYA - NISHINOMIYA	8	3	-5
NARA - URAWA	2	3	+1
[TO]KUSHIMA - [FU]KUSHIMA	2	3	+1
WAKAYAMA - OKAYAMA	3	2	-1
[FU]KUYAMA - TOYAMA	2	2	0
KOH[CHI] - OH[TSU]	2	2	0
[FU]CHUU - [HA]CHIOJI	0	2	+2
TAKASA[KI] - NAGASA[KI]	9	0	-9
MATSUDO - SASEHO	5	0	-5
YAMAGATA - HIRAKATA	3	0	-3
KANAZAWA - NAGANO	2	0	-2
[FU]KUI - FUJI	2	0	-2
I[CHI]KAWA - FUJISAWA	2	0	-2
I[CHI]KAWA - I[CHI]HARA	2	0	-2
FUNABA[SHI] - MAEBA[SHI]	2	0	-2
YOKOHAMA - [TO]KOROZAWA	2	0	-2
HAMAMA[TSU] - TAKAMA[TSU]	2	0	-2
TOYOHA[SHI] - TAKASA[KI]	2	0	-2
AKA[SHI] - TAKA[TSU]KI	2	0	-2
KURA[SHI]KI - TAKA[TSU]KI	2	0	-2
SHIMONOSE[KI] - NAHA	2	0	-2
OTHERS	55	17	
TOTAL	124	48	

本方法において、ケプストラムの回帰係数の抽出に用いる音声区間の長さの最適値は、英語の文章音声を用いた話者認識の場合の最適値の1／2程度に短い。これが、用いている特徴の違い（韻律的特徴と音韻的特徴）によるものか、言語の違いによるものかは今のところ不明であるが、興味ある点である。

[謝辞]種々の適切なアドバイスを頂いた、情報通信基礎研究部第四研究室・寛室長、好田調査役、Staggered Array DPのプログラムを提供して頂いた鹿野調査役、有益な討論をして頂いた研究室の皆様に心から感謝の意を表する。

[文　献]

[1]境、中山:"聴覚と音響心理"、日本音響学会編、コロナ社（昭53）

[2]古井 貞熙:"音節知覚におけるスペクトルの動的特徴の役割りについて"、音響学会秋季講論、1-1-12（昭59-10）

[3]古井 貞熙:"音声知覚におけるスペクトル変化情報の役割り"、音響学会聴覚研資、予定（昭60-01）

[4]Furui,S and Rosenberg,A.E.:"Experimental Studies in a New Automatic Speaker Verification System Using Telephone Speech",Proc.IEEE Int.Conf.Acoust.,Speech,Signal Processing,Denver,CO,pp.1060-1062 (1980)

[5]Furui,S.:"Cepstral Analysis Technique for Automatic Speaker Verification",IEEE Trans.Acoust.,Speech,Signal Processing,ASSP-29,2,pp.254-272 (1981)

[6]山下、柳田、溝口、角所:"極周波数の変動傾向に基づいた有声破裂音の識別"、音響学会音声研資、S83-77 (昭58)

[7]間野、白井:"音声過渡部のラベリング"、音響学会音声研資、S84-55 (昭59)

[8]鹿野、相川:"Staggered Array ＤＰマッチング"、音響学会音声研資、S82-15 (昭57)

[9]菅村、相川、鹿野、好田:"SPLIT、マルチテンプレート法による不特定話者単語音声認識、音響学会音声研資、S82-64 (昭57)

[10]相川、鹿野、古井:"パワー情報で重みづけた距離による単語音声認識"、音響学会音声研資、S81-59 (昭56)

[11]Rabiner,L.R.,Sondhi,M.M. and Levinson,S.E.:"A Vector Quantizer Combining Energy and LPC Parameters and Its Application to Isolated Word Recognition",Bell Labs. Tech.J.,63,5,pp.721-735 (1984)

[12]浅川、畑岡、小松、市川:"不特定話者連続数字認識方式の検討"、音響学会音声研資、S83-53 (昭58)

[13]杉山、鹿野:"ピークに重みをおいたＬＰＣスペクトルマッチング尺度"、信学論、J64-A、5、pp.409-416 (昭56)

[14]菅村、古井:"疑音韻標準パタンによる大語い単語音声認識"、信学論、J65-D、8、pp.1041-1043 (昭57)

[15]Sugamura,N.,Shikano,K. and Furui,S.:"Isolated Word Recognition Using Phoneme-Like Templates",Proc.IEEE Int.Conf.Acoust.,Speech,Signal Processing,Boston,MA, pp.723-726 (1983)

[16]菅村、鹿野、好田:"SPLIT、単語マルチテンプレート法による不特定話者単語音声認識"、信学論、J67-D、10、pp.1210-1217 (昭59)