音声の認識と発生*

中田和男**

1. 一般的な音声認識の困難さ

だれの声でも，どんな内容の音声でも認識することのできる，いわゆる一般的な音声認識装置ができればその便利さはいうまでもない．しかし現在の技術からみてそれはほとんど不可能といってよいほどむずかしい．

その困難さの原因は，次の2点に集約される．

（1） 話者の変化による音声の特長の広範な変化．

（2） 音声の内容（音素環境）の変化による音声の特長の広範な変化．

男の声と女の声，おとなの声と子供の声，それは同じ言語内容を話していながら，その音声波形の物理的な特長は非常に異なっており，現在の技術ではそれが同一の音声を話したものであると認識することは非常にむずかしい．また同一人の音声であっても /keisanki/ の /a/ と /onseitaipu/ の /a/ とではその物理的な特長（たとえば周波数スペクトルの構造）は非常に異なっており，日本語として可能なすべての音素環境(音素の前後関係）のもとで常に確実に /a/ という音素を認識するということは現状ではほとんど不可能に近い．

このような困難さを解決する手段として，言語情報の利用ということが取り上げられ，種々研究されているが，まだ，決定的な解決となるまでには至っていない[1]．したがって一般的な音声認識（音素もしくは音節単位の音声認識）は，学問的な興味からは数多く研究されているけれども，実用を目的として研究されるには至っていない．

2. 音声認識への工学的アプローチ

実用を意識した音声認識の研究，いわば音声認識への工学的なアプローチともいうべき立場では，話者と語彙を限定することで，上記の一般的な音声認識において出くわす本質的な困難さを避けようと試みている．

認識の対象とすべき語彙を十数語から数十語の特定の単語，たとえば，0から9までの数字とか FORTRAN の命令語とか主要都市名とかいったものにかぎり（入力されるものは必ずその一つであると仮定し），しかも話者を特定の一個人に限定する．

このような限定条件のもとでは，認識のための基本的なアルゴリズムとして，単語を単位とするパターンマッチの手法が使えることになり，上記の本質的な困難さは一応避けられるようにみえる．しかし実際に実験してみればすぐわかるように，なお，次の2点が問題点として残っている．

（1） 話す速さの違い

アナウンサーのような特別な訓練を受けた人は別として，普通の人では，たとえ簡単な0から9までの 10 個の数字音にしても，その音声波形の時間的な構造は話すたびにかなり大幅に変わっている．一例として数字音の継続時間（時間的な長さ）をとってみても，けっして一定ではない．しかもその内部的な構造の変化は，全時間長に比例した線形な時間軸の伸縮によって一致するような簡単なものではない．しかも音声の情報は時間的な流れであり，時間を独立変数もしくはパラメータとしないような表現（記述）はできない．とすれば簡単にパターンマッチというけれど，標準のどの部分（時点）と入力のどの部分（時点）とを対応させてマッチングをとればよいかが簡単には決まってこない．

（2） 確実にして有効な特長

簡単にパターンマッチとはいうものの，音声波形そのものを直接比較したのでは情報量が多すぎて標準パターンの記憶容量と認識のための処理時間が非現実的なものとなってしまう．

たとえば0から9までの 10 個の数字音を認識するとして，一つの数字音が平均0.5秒の長さであり，その波形情報を 8 kHz サンプリング，8 ビットで記憶するとして，1語の平均情報量は，$8\times8\times10^3\times0.5=32$ k ビットとなり，必要な全記憶容量はその 10 倍で 320 k ビットとなる．

しかも音声の波形そのものは，振幅の比例的な

---

* Recognition and Generation of Speach, by Kazuo Nakata (Central Research Laboratory, Hitachi Ltd.)

** 日立製作所中央研究所

大小変化は別としても，マイクロホンや増幅器やフィルターなどの位相特性，口からマイクロホンまでの距離などによってすぐに変わってしまう．

とすればなにかもっと有効な特長(パラメータ)を音声波形から抽出して，その空間でのパターンマッチを行なうことが望ましく，その特長はなにかということになる．これがまたそう簡単にはきまらない．

しかしこの特長は少なくとも次のような性質をできるだけ備えたものであることが望ましい．

（a）　抽出が容易(装置が簡単)，確実(再現性あり)，迅速(できれば実時間)であること，

（b）　音声の言語的な内容によく対応したもので，認識に十分役だつものであること．

（c）　情報量が圧縮されたものであること．

従来，音声スペクトルの主要な成分であるフォルマント周波数とか，音声波形の零交差波とかが特長としてよく用いられきたが，前者は，上記の（a）に問題があり，後者は，（b）に問題があった．

この二つの問題に対する新しい解決の可能性(breakthrough)が最近の研究に現われてきているので，以下それを中心に解説をすすめることにする．

3.　動的計画法を利用した時間軸正規化

音声情報の時間的構造の特長は，その順序性と非線形性にある．すなわち，同一の言語内容を話した音声では，その音素は決められた順序に現われてき，その前後関係が変わるということはない．しかし，異なった時点での発声による同一音声の間の時間的構造の対応はけっして線形な時間軸の変換（平行移動と一様な伸縮）で解決される性質のものではない．

この問題に対する新しい解決法として音声認識にもちこまれたのが動的計画法(Dynamic Programming)の手法である．動的計画法はもともと多段決定過程の最適化手法として開発されたものであるが，音声情報の時間的な順序性の制限のもとでの非線形性を処理する手法として応用することができる[2]．

単語 A の音声の特長の時間的変化を $\boldsymbol{A}(t)$，B のそれを $\boldsymbol{B}(t)$ とし，それぞれの継続時間を T_A，T_B とすれば，

$$A=\boldsymbol{A}(t),\quad 0\leq t\leq T_A$$
$$B=\boldsymbol{B}(t),\quad 0\leq t\leq T_B$$

そこでこの二つの単語のパターン $\boldsymbol{A}(t)$ と $\boldsymbol{B}(t)$ を比較するに当って，順序性を保ちながらの非線形な時間軸の変換 $u(t)$ を考え，$\boldsymbol{B}(t)$ を $\boldsymbol{B}'(t)$ $(0\leq t\leq T_B')$ に変換して考える．

$u(t)$ は，（ⅰ）　t の連続，単調増加関数

（ⅱ）　$u(0)=0$，$u(T_B')=u(T_B)$

（ⅲ）　$u(t)$ の値は t の近くにある

の3条件を満たすものとする．（ⅲ）は極端な変換は適切でないという実用上の制限である．

このような $u(t)$ のなかで，たとえば次式で定義されるような整合誤差 E を最小にするようなものを選ぶ．

$$E(A,B,u)=\frac{\int\|\boldsymbol{A}(t)-\boldsymbol{B}(u(t))\|\,ds}{\int ds}$$

上式で $\|\boldsymbol{A}(t)-\boldsymbol{B}(u(t))\|$ は $\boldsymbol{A}(t)$ と $\boldsymbol{B}(u(t))$ の時点 t における距離とし，$ds=\sqrt{1+u'(t)}dt$ 積分は曲線 $\tau=u(t)$ にそって $(0,\ 0)$ から $(T_A,\ T_B)$ まで行なうものとする（図 3.1 参照）．このときこの曲線 τ が $(0,\ 0)$ と $(T_A,\ T_B)$ を結ぶ直線からあまり大きくは

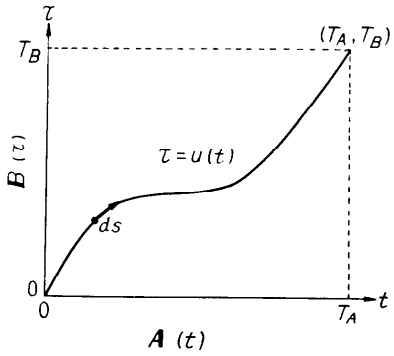

図 3.1　$\tau=u(t)$ による非線形時間軸正規化の原理

ずれないことを要求するのが上記の条件（ⅲ）である（T_A と極端に異なる T_B を与えるようなパターン $\boldsymbol{B}$ は最初から $\boldsymbol{A}$ と整合をとるべきパターンの候補から除く）．

$$D(A,B)=\min_{u(t)}[E(A,B,u)]$$

をもって音声 A と B の時間ずれを整合して比較した距離を考える．

時間的にサンプリングされたパターンについての具体的な動的計画法による計算手続きについては，迫江，千葉の論文[2]にゆずるが，この計算では収束性の問題

(くり返し計算によって最適化する)がなく計算時間が速いこと，終点の条件を開放とすることにより連続音声中の単語の区分・認識を行なうことができるなどの利点をもっている．

動的計画法の音声認識への応用については，日本における迫江，千葉らの研究と同時に，ソ連でも研究が行なわれていることが報告されている[3]．

実験結果の報告によれば[2]，NEAC 3100 (サイクルタイム 2 μsec) で，1桁数字当りの認識に約 1.5 秒を要している．この実験では音声の特長パラメータは専用の金物によって実時間で抽出されているから，この時間は標準パターン 10 個との間の動的計画手法による時間軸の整合を含めたパターン全体の整合処理に要している時間とみることができる．

確かに動的計画法による時間軸の非線形変換整合は従来の手法に比べて効果・速度の点で格段にすぐれた手法ということができるが，上記の実験例でみてもわかるようになお実時間処理には至っていない．この点完全な動的計画手法の実現ではないが，それとほぼ同じような効果をもつ簡易化手法の開発が今後の実用上の問題点としてあげられよう．

4. 音声の新しい特長パラメータ

音声の言語（音素）的な情報を最もよく表わすパラメータとして，従来音声波形の周波数スペクトル（正しくはその包絡）もしくはその主要成分としてのフォルマント周波数が取り上げられてきた[4),5)]．この考えの基本的な正しさは音声発生の音響的な理論からいって，より根源的であるか否かの論議は別として，まずまちがいない．

ただ問題は，それを理論的なものにできるだけ近い形で，確実かつ容易に抽出する手法がなかなかみつからなかったことにある．

金物のフィルターによる分析は，ピッチ周波数の影響を受けやすく，Analysis-by-Synthesis による分析手法は収束性（そのための手続きと制御パラメータの数），したがって抽出までの所要時間に問題がある[6]．

最近になって，板倉，斉藤は最尤法による音声スペクトル包絡の推定法を開発した．われわれの実験的な経験によれば，音声のスペクトル包絡もしくはそれを記述する主要成分としてのフォルマント周波数の抽出に関して，この手法は最も確実でかつ高速な手法ということができる．

この最尤スペクトル推定法において中間的なパラメータとしてαパラメータと呼ばれているものが抽出されるが，音声スペクトルのモデルとして零点をもたない有理スペクトル密度を仮定したことから，それは同時に音声波形の最小自乗誤差の線形予測係数となっている．

一方，板倉，斉藤はαパラメータによる音声の情報圧縮伝送系（分解・合成系）の改良から，音声波形の偏自己相関係数として k パラメータと呼ばれている係数を導き出し，α パラメータとの関係を求めているが，この関係式は Schmit の直交化形式となっており，α パラメータを直交化したものが k パラメータであることを示している．この間の関係を図解的に図 4.1 に示す．

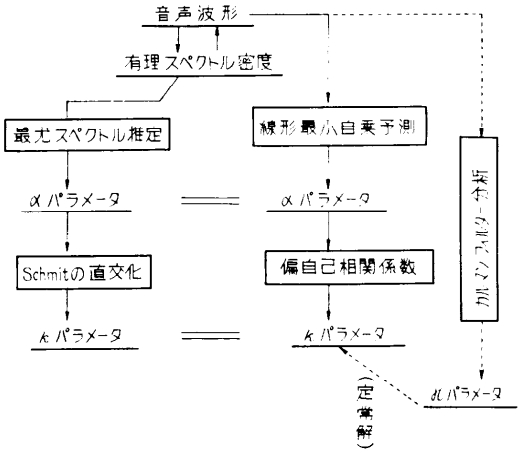

図 4.1 音声の特長パラメータ間の関係

パターン認識の基礎理論からいって，認識パラメータ(特長)の選び方の基本は，できるだけ相互に直交化したものを選ぶのが有効であることはよく知られている（K-L 系展開はその代表的な一般的手法）．

以上概観したように，音声の認識に音声のスペクトル（包絡）の情報が原理的に有効な特長であるとするならば，それの推定のデータとなるαパラメータもまた音声認識に有効な情報を内蔵していることになり，それを直交化した k パラメータは認識に最も適した形でのパラメータ（特長）ということができよう．以下板倉，斉藤の論文[7)~9)]を要約してこの間の関係を整理しながら，k パラメータ（音声の偏自己相関係数）による音声認識の問題点に触れてみよう．

5. 音声のスペクトル包絡パラメータ $\{\alpha\}$

音声のスペクトル包絡のモデルとして，零点をもた

ない p 次の有理スペクトル密度 $f_1(\lambda)$ を仮定する.

$$f_1(\lambda)=\frac{\sigma^2}{2\pi}\frac{1}{\left|\prod_{i=1}^{p}\left(1-\frac{z}{z_i}\right)\right|^2}$$

$$=\frac{\sigma^2}{2\pi}\cdot\frac{1}{\left|\sum_{i=0}^{p}\alpha_i z^i\right|^2} \qquad (1)$$

ここで λ は基準化角周波数であり，$\lambda=\pm\pi$ が周波数 $\pm W$ Hz ($2W$ は音声波形のサンプリング周波数) に相当する．σ^2 はそのエネルギーである.

$z=\exp(-j\lambda)$, z_i は次式の根である.

$$1+\alpha_1 z+\alpha_2 z^2+\alpha_3 z^3+\cdots+\alpha_p z^p=0 \qquad (2)$$

このような仮定のもとで，N 個の観測音声波形のサンプル値 $(x_1, x_2, \cdots, x_N)$ を得る対数尤度を最大にするパラメータ $\{\alpha_i\}_1^p$ は次の連立方程式を解くことによって得られる.

$$\sum_{j=0}^{p}\hat{v}_{j-i}\cdot\alpha_j=0 \quad (i=1,\cdots,p,\ \alpha_0=1) \qquad (3)$$

ここで，$\hat{v}_\tau=\frac{1}{N}\sum_{t=1}^{N-|\tau|}x_t\cdot x_{t+|\tau|}$

すなわち x_t の短時間自己相関係数.

$\{\alpha_i\}_1^p$ が求まれば式(2)の p 次方程式を解くことによって z_i が求まり，z_i からいわゆるフォルマントが次式で与えられる.

$$\left.\begin{aligned}&\text{フォルマント周波数}\\&F_i=\arg z_i/2\pi\Delta T=W\cdot\arg z_i/\pi\\&\text{フォルマント帯域幅}\\&B_i=\log|z_i|/2\pi\Delta T=W\log|z_i|\pi\end{aligned}\right\} \qquad (4)$$

問題の α パラメータは観測波形から p 単位時間の遅れまでの短時間相関係数 $\{\hat{v}_\tau\}_0^p$ によって求められる.

このときの積分時間長（窓関数の時間幅）によってどの程度の平滑化を行なうかが決定される.

ここで見方を変えて次のような問題を考える.

定常時系列 $\{x_t\}$ において，相つぐ p 個の観測値 $(x_{t-1}, x_{t-2}, \cdots, x_{t-p})$ から x_t を予測する線形予測係数を $\{a_i\}_1^p$，推定誤差を $\varepsilon_t^{(p)}$ とすれば，

$$\varepsilon_t^{(p)}=x_t-\hat{x}_t=x_t+\sum_{i=1}^{p}a_i x_{t-i}$$

$$=\sum_{i=0}^{p}a_i x_{t-i} \quad (i=0, 1, \cdots, p) \qquad (5)$$

このとき平均自乗誤差 $E^{(p)}=E[(\varepsilon_t^{(p)})^2]$ を最小にする線形予測係数 $\{a_i\}_1^p$ は，簡単な計算から，次の p 元連立1次方程式の解であることがわかる.

$$\sum_{j=0}^{p}\hat{v}_{i-j}\cdot a_j=0 \quad (i=1,\cdots,p,\ a_0=1) \qquad (6)$$

ここで $\hat{v}_{i-j}$ は定常時系列 $\{x_t\}$ の観測サンプル値間の相関係数であり，$\hat{v}_{i-j}=\hat{v}_{j-i}$ であるから，式(6)は式(3)に等しいことがわかる.

いいかえれば，最尤推定法によるスペクトル包絡パラメータ $\{\alpha_i\}_1^p$ は，定常時系列 $\{x_t\}$ の p 個のサンプル値 $\{x_{t-i}\}_1^p$ から x_t を最小自乗誤差で推定するときの線形予測係数 $\{a_i\}_1^p$ に等しいことがわかる.

このことは，音声波形 $\{x_t\}$ が式(1)のような p 次の有理スペクトル（包絡）密度をもつと仮定することと，式(5)で表わされるような p 個までの線形依存性をもつと仮定することがまったく同じことの2側面であることを理解すれば容易にうなずけることである.

6. 音声の偏自己相関係数 $\{k\}$

板倉らはさらに理論を展開させ，音声波形 x_t と x_{t-n} との間の偏自己相関係数として，その間のサンプル $(x_{t-1}, \cdots, x_{t-n+1})$ によって前向きおよび後向きに線形予測できる部分を取り除いた部分（予測誤差）の相関を定義し，これを偏自己相関係数または k パラメータと略称した.

$$k_n\triangleq\frac{E[\varepsilon_{ft}^{(n-1)}\cdot\varepsilon_{bt}^{(n-1)}]}{\sqrt{E[\{\varepsilon_{ft}^{(n-1)}\}^2]\cdot E[\{\varepsilon_{bt}^{(n-1)}\}^2]}}=\frac{w_{n-1}}{u_{n-1}} \qquad (7)$$

ここで，

$$\left.\begin{aligned}&\varepsilon_{ft}^{(n-1)}=x_t-\sum_{i=1}^{n-1}\alpha_i x_{t-i}=\sum_{i=0}^{n-1}\alpha_i x_{t-i}\\&\varepsilon_{bt}^{(n-1)}=x_{t-n}-\sum_{i=1}^{n-1}\beta_i x_{t-n-i}=\sum_{i=1}^{n}\beta_i x_{t-n-i}\\&\beta_i=\alpha_{n-i} \quad (i=1,\cdots,n)\end{aligned}\right\} \qquad (8)$$

実際に音声サンプル間の $(p+1)$ の相関係数 v_i $(i=0,\cdots,p)$ から α および k を逐次的に求めるフローを図 6.1 に示す[8].

このフローに示される関係から，k は α を Schmit の直交化式に従って直交化したものであることがわかる．事実，α パラメータはその次数 p を変えるとすべての値が変わってしまうが，k パラメータの場合には，次数を変えても低次のほうの値は変わらずに新たに拡張された次数に応じた部分をつけ加えていくだけでよい.

α にしても k にしても，その計算は逐次的ではある

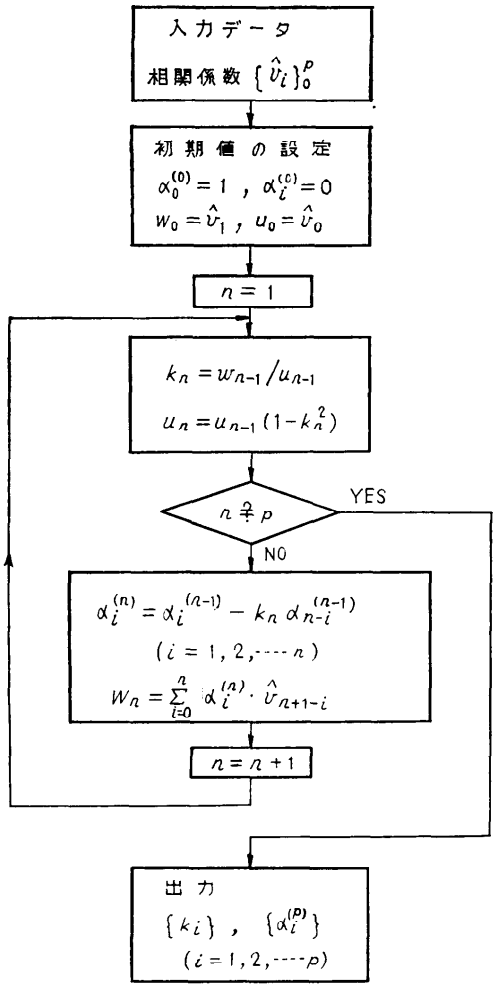

図 6.1 α および k パラメータの逐次計算フロー

が，くり返し的ではなく収束性の問題もなく，計算機によって高速に求めることができる．ことにそれらを求めるために必要な $(p+1)$ 個の相関値さえ専用ハードによって高速に求めれば，実時間で導き出すことも困難ではない．

k パラメータが音声波形の偏自己相関係数として定義され，各遅れ時間ごとに，既知のサンプル値から線形予測可能な部分を引き去った残りの部分の相関係数として順次定義されていくことは，予測系を直交化しできるだけ能率的な表現をしようとしていることであり，直交化パラメータであることの物理的なイメージもとらえやすいといえよう．

以後，偏自己相関係数として定義される $\{k_i\}_{p_1}$ のことを k パラメータと略称する．

7. k パラメータによる音声認識の問題点

以上述べたことを要約すれば，語彙および話者を限定し，単語単位のパターンマッチという手法が使えるように問題を限定したとき，音声の特長パラメータとしては k パラメータを用い，k パラメータ空間で，動的計画法による非線形な時間軸の変換を行なって整合をとるという方法が現状では最もよい（簡単で確実な）音声認識法ということになる．

たとえ原理的にそうであるとしても，実際上問題がないわけではない．

その一つは k パラメータの計算精度の問題である．

k パラメータの計算は図 6.1 のような流れ図によって逐次的に行なわれるから，高次の k パラメータほど計算誤差の影響を受けやすくなる．ことに k パラメータの定義式からみてもわかるように，次数が高くなるほど，予測誤差量の平均値である分子，分母の値は 0 に近く不安定なものとなりやすい．したがって，p の値としては 6～12 が適当であるが，そのすべてを同一のウエイトで整合を計るのが最適であるかどうか問題であろう．

その二つは振幅情報（σ^2 または σ）の利用である．k パラメータはエネルギー的には正規化されているから，振幅の情報はすてられてしまっている．一つの単語内で正規化（たとえば最大値を 1 とする）された構振幅のパターン（時間的な変化）は，単語内の音節の成，特に有声・無声の別，鼻音やラ行音の存在などについてはかなり確実で有用な情報をもっているから，なんらかの形でこれを併用して利用することを考えるべきである．

その 3 は k パラメータの平滑化についてである．k パラメータ，ことに高次のそれは時間的に小きざみな変動を示す．これが音声認識にとって重要な情報かどうかである．このことは，高次の k パラメータの値を平滑化することによって，k パラメータによる音声合成方式（PARCOR 方式）の品質がどう変わるかによって評価することもできようが，認識実験によっても実験的に検証できよう．

8. Kalman フィルター理論による分析

k パラメータを利用した音声認識の問題点の一つとして抽出された k パラメータの平滑化ということを述べたが，この問題をもっときちんと理論的に取り扱

おうとするのが松井らによる Kalman フィルター分析である[10]．以下その考え方の要約を示す．

音声波形を内部状態 $\{\alpha\}_1^p$ によって記述される系の出力と考え，その間に次の観測式が成りたつと考える．

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} x_1, & x_2, & \cdots, & x_p \\ x_2, & x_3, & \cdots, & x_{p+1} \\ \vdots & & & \\ x_n, & x_{n+1}, & \cdots, & x_{n+p-1} \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_{n-1} \end{bmatrix} \quad (9)$$

$\boldsymbol{x}=\boldsymbol{H}\cdot\boldsymbol{\alpha}+\boldsymbol{\varepsilon}$ と略記する． (10)

ここで $\{\alpha_i\}_1^p$ は **6.** で述べた線形予測係数であり x_i の x_0〜x_{n-1} の n 個のサンプル区間では定常と仮定する．$\{\varepsilon_i\}$ は（予測）誤差ベクトルであり，平均値 0，共分散行列 R は既知であると仮定する．

ここで以後の数学的な処理を簡単にするために，

$$x_j=\sum_{i=1}^{p}\alpha_i x_{j+i}+\varepsilon_j \quad (i=0, 1, \cdots, n-1) \quad (11)$$

の予測表示において予測系 $\{x_{j+i}\}$ を Schmit の直交化に従って直交化し，そのときの予測係数（状態ベクトル）を $\{\kappa_i\}_1^p$ とすると，式(10)は次のように略記される．

$$\boldsymbol{x}=\boldsymbol{H}\cdot\boldsymbol{B}\boldsymbol{\kappa}+\boldsymbol{\varepsilon} \quad (12)$$

ここで $\boldsymbol{B}$ は直交変換のための変換行列である．

このとき，誤差 $\boldsymbol{\varepsilon}$ を2乗和が最小になるように状態ベクトル $\{\kappa_i\}_1^p$ を求めると，形式的には，

$\boldsymbol{\kappa}=(\boldsymbol{H}^T\cdot\boldsymbol{H}\cdot\boldsymbol{B})^{-1}\boldsymbol{H}^T\boldsymbol{x}$ と求められる． (13)

波形 x のサンプル値（観測値）$\{x_i\}$ からの具体的な計算手続きは松井らの論文にゆずるが，このようにして求められた $\{\kappa_i\}_1^p$ は，波形 $\{x_i\}$ を定常と考え，その相関係数 c_{ij} の値が $\Delta=|i-j|$ の等しいものはすべて等しいと考えたときに，式(7)によって定義される板倉の k パラメータに等しくなる関係にある．いいかえれば松井の κ パラメータは板倉の k パラメータを非定常な観測波形の場合に拡張したものということができる．

さてこのようにして抽出された κ パラメータ（状態 κ）は次の二次形式に従って時間的に推移するものと仮定する．

$$\begin{bmatrix} \kappa_1 \\ \kappa_2 \\ \vdots \\ \kappa_p \end{bmatrix}_{T+1} = 2\begin{bmatrix} \varphi & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi \end{bmatrix}\begin{bmatrix} \kappa_1 \\ \kappa_2 \\ \vdots \\ \kappa_p \end{bmatrix}_T - \begin{bmatrix} \varphi & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi \end{bmatrix}^2\begin{bmatrix} \kappa_1 \\ \kappa_2 \\ \vdots \\ \kappa_p \end{bmatrix}_{T-2} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix}_T \quad (14)$$

$\boldsymbol{\kappa}_{T+1}=2\boldsymbol{\Phi}\boldsymbol{\kappa}_T-\boldsymbol{\Phi}^2\boldsymbol{\kappa}_{T-1}+\boldsymbol{W}_T$ と略記する． (15)

ここで，$\varphi=\exp[-T/\tau]$，τ は時定数，$\{w_i\}$ は誤差ベクトルで平均値 0，共分散行列 Q は既知であるとする．

さらに $\boldsymbol{\lambda}_{T+1}=\boldsymbol{\Phi}\boldsymbol{\kappa}_T$ とおけば，

$$\begin{bmatrix} \boldsymbol{\kappa} \\ \boldsymbol{\lambda} \end{bmatrix}_T = \begin{bmatrix} 2\boldsymbol{\Phi} & -\boldsymbol{\Phi} \\ \boldsymbol{\Phi} & \boldsymbol{0} \end{bmatrix}\begin{bmatrix} \boldsymbol{\kappa} \\ \boldsymbol{\lambda} \end{bmatrix}_{T-1} + \begin{bmatrix} \boldsymbol{W} \\ \boldsymbol{0} \end{bmatrix}_{T-1} \quad (16)$$

κ の事前推定値を $\bar{\kappa}$，事後推定値を $\hat{\kappa}$ で表わすと式(15)の平均をとって，

$$\begin{bmatrix} \bar{\boldsymbol{\kappa}} \\ \bar{\boldsymbol{\lambda}} \end{bmatrix}_T = \begin{bmatrix} 2\boldsymbol{\Phi} & -\boldsymbol{\Phi} \\ \boldsymbol{\Phi} & 0 \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\kappa}} \\ \hat{\boldsymbol{\lambda}} \end{bmatrix}_{T-1} \quad (17)$$

さて，$\bar{\kappa}_T$，$\bar{\lambda}_T$ を $\hat{\kappa}_{T-1}$，$\hat{\lambda}_{T-1}$ から得た時点で，新しい観測を行なって，式(12)を得たとき，最も合理的な κ と λ の推定値は次式のような重みつき2乗誤差の総和を最小にするように選ぶというのが松井らの主張である．

$$J=\frac{1}{2}(\boldsymbol{x}-\boldsymbol{H}\cdot\boldsymbol{B}\cdot\boldsymbol{\kappa})^T\boldsymbol{R}^{-1}(\boldsymbol{x}-\boldsymbol{H}\cdot\boldsymbol{B}\cdot\boldsymbol{\kappa}) + \frac{1}{2}\begin{bmatrix} \boldsymbol{\kappa}-\bar{\boldsymbol{\kappa}} \\ \boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}} \end{bmatrix}^T\begin{bmatrix} M_{11} & M_{12} \\ M_{22} & M_{22} \end{bmatrix}\begin{bmatrix} \boldsymbol{\kappa}-\bar{\boldsymbol{\kappa}} \\ \boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}} \end{bmatrix} \quad (18)$$

ここで第1項は波形の予測誤差の2乗和であり，第2項は状態ベクトル $\boldsymbol{\kappa}$（と $\boldsymbol{\lambda}$）の予測誤差の2乗和である（$[\boldsymbol{M}]$ は $\boldsymbol{\kappa}$ の事前推定の共分散である）．

松井らはこのような考え方に基づいて，具体的に状態パラメータ κ の最適推定値を求める計算手続き（アルゴリズム）を与えるとともに，κ から α を求める手段をも示し二つの実験事実によって κ パラメータの妥当性を実証している[10]．

（1） 板倉の α パラメータからスペクトル包絡を求めるよりは，κ パラメータから α パラメータを求め，それから求めたスペクトル包絡のほうが，全体として音声を特長づけるフォルマントの動きを明瞭にかつ連続的に観測することができる．

（2） κ パラメータを板倉の k パラメータとして PARCOR 方式によって音声を合成したとき，p を大きくとれば原音声とほとんど同じ音声を再生することができる．

表 9.1　音声認識研究の内外の現状

研究機関＼項目	語彙	話者	パラメータ	認識率	目的	備考
電電公社通研	1桁数字音	男 24名	k パラメータ	99.8%		最尤推定，学習
KDD研究所	1桁数字音	男 10名	18チャネルスペクトル	誤：なし Reject：3%		書き替え規則の応用
NHK総合技研	5母音	不特定多数	10チャネルスペクトル	96%以上		主要因分析
日本電気中研	限定FORTRAN（約50語）	男 2名	8チャネルスペクトル	99.6%	音声による計算機入力	DPの利用* ほぼ実時間
富士通	制御語（約10語）	不特定多数	10チャネルスペクトル		機械の制御	万博用
日立中研	1桁数字音	男 1名	ケプストラム k パラメータ	99.8%		
東北大通研	都市名（13個）	男 1名	4チャネルスペクトル	96%	純研究	単語辞書の利用
京大工学部	1桁数字音	男 16名	18チャネルスペクトル	91～94%	純研究	シンボル系列の処理
ベル電話研	1桁数字音	不特定多数	音声エネルギー	100%	自動ダイヤル	特殊な方法**
IBM研究所	1桁数字音	男 11名 女 2名	8チャネルスペクトル	初回：87.9% 訓練：96.4%		パターンマッチ
RCA	数字音				小包区分機の制御	現在中止
ソ連科学アカデミー	ALGOL 60用 203語	男 2名	4チャネルスペクトル	95.2%	計算機入力	ノボシビリスク， DP使用
スタンフォード大学	192の限定文章	男 2名	3チャネルスペクトル と零交差回数	～85%	ロボットの制御	

* 個人別に認識
** 厳密には音声認識とはいえない

κ パラメータによる音声認識の試みはまだ発表されていないが，上記の（1）の実験事実からみて十分有効であろうと推定される．問題はどの程度の処理時間で κ パラメータを抽出できるかであるが，松井らの報告によれば，FFT によるスペクトル抽出や k パラメータの抽出と差がなく実用の域に達したものと報告されている．

9. 音声認識の今後の問題

一般的な音声認識の問題点および音声認識の具体例については，最近，藤崎の適切な解説があるので[11]それにゆずるとして，最近の発表をまとめて表9.1に示す．

さて，音声認識の実用化にあたっての今後の問題であるが，技術的にはなんといっても，話者や語彙の制限を少しでもゆるめるように努力することであろう．

押しボタン・ダイヤルやキイボードによる情報の入力は，実際やってみるとかなりわずらわしい作業であり，ことにちょっと複雑な内容になるとコードを表で引くのがまたいちいちたいへんな手間である．電電公社の計算サービスもその内容が高度になって実用的なものとなればなるほど，この入力のわずらわしさが壁となるであろう．たとえ十数字とアルファベットだけでも，不特定多数のユーザーの音声を確実に認識できれば，その実用上の効果は絶大であろう．一方語彙の増加に対しては，認識できる語彙を増すということのまえに，たとえ十語の認識であっても，そのほかの単語や雑音は必らず棄却し語認識しないということも実用上かなり重要なことである．

結論的にいって音声認識に対する実用上の要求もようやく顕存化しつつあるが，まだ技術が一歩及ばないというのが現状であるようにみうけられる．

10. 音声の発生，特に音声応答方式

計算機で処理された結果を音声の形で出力するのに必要な装置を音声応答装置といい，そのための音声の作り方を音声応答方式と呼ぶ．よく知られているように現在実用化されている音声応答方式は「録音編集方式」であり，あらかじめ録音されている人間の声を単語単位でつなぎあわせて音声を作り出している．この方式の欠点は情報量の多い音声波形を直接記憶しておくため，出力できる音声の内容（語彙）が少ないことであり，したがって用途も限定されてしまう．

この欠点を補なうために音声合成の原理を取り入れた音声応答方式がいくつか発表されているが[9),12)~14)]，装置が簡単で経済的なものは出力音声の品質が十分でなく，高品質の音声を出力できる方式は装置が複雑になり経済性の面に問題があり，まだ完全に実用化されるには至っていない．

先に述べた偏自己相関係数 k パラメータを用いる

PARCOR 方式[15]がいわゆる音声合成方式としては現在いちばん有望視されているが，まだ研究試作の段階である．

従来の音声合成方式の基本的な考え方は，合成という形で音声波形のもっている本質的な制約（任意の音波のなかである特定のものを音声波形たらしめている原因——音声の特長）を抽出し規則化して音声合成装置に内蔵させ，音声を作り出すのに必要な情報をできるだけ圧縮して記憶しておき，一定の記憶容量で実用上必要な語彙を作り出すことができるようにしようとするものである．

音声を作り出すのに必要な情報の記憶量と情報の処理量の積は定性的にいってほぼ一定であり，従来の音声合成原理による方式は必要記憶量を現状技術で可能な範囲に押えるために，合成という処理がかなり複雑なものとなっている．

しかし記憶素子，技術の進歩は計算機産業にささえられて急速に発展しつつあり，記憶容量についての技術的・経済的な制限はしだいに緩和されつつある．このような背景を考えるとき，将来の音声応答方式としては従来のものよりももっと記憶容量をふんだんに使った memory dependent な方式を考える必要があろう．

その一つの可能性として KDD 研の樽松，井上の研究がある[16]．その原理を以下に簡単に説明しよう．

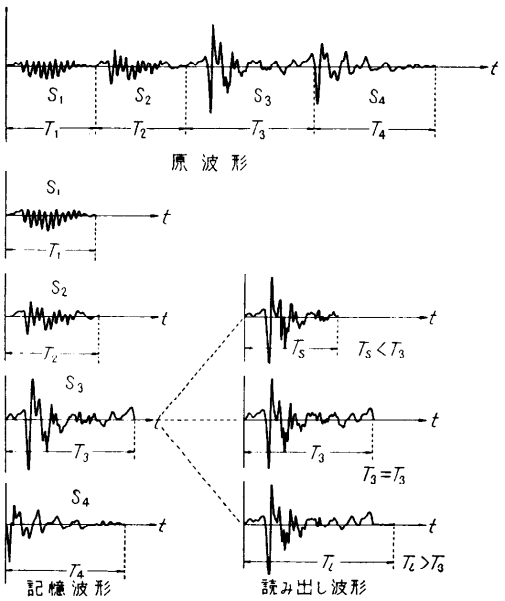

図 10.1　音声のピッチ単位素片を用いた音声合成の原理

まず音声波形，たとえば「サン」（数字音の「3」）を，まずピッチ周期ごとに区分し，図 10.1 に示すように1ピッチごとに分離して記憶しておく．再生（出力）にあたっては，外部からのピッチ制御情報によってこの記憶波形から順次必要な長さだけ読み出してつないでいく．このとき記憶されている1ピッチの波形の時間長 T_i に対して読み出すべきピッチ周期 T_s が短かいときは，その時点までで波形の読み出しを打ち切り，ピッチ周期 T_i が長いときは，T_i 以上の時点ではその波形の最終値（普通は零レベルになるよう T_i をとる）を保持する．このようにして出力波形のピッチ周期を外部からの情報によって適切に制御すれば，一つの音声波形「サン」から任意の単語イントネーションをもった発音「サン」を作り出すことができる．「3 3 3の3 3 3 3」といったような電話番号もこのようにして作れば一つの「3」から自然に作り出すことができる．

この考えをさらにおしすすめていくと，日本語の百音節を基本として記憶しておき，そのつなぎ合わせで任意の単語を自然性をもって作り出すことができる．

もちろん，より自然なものとするためには，基本を単音節ではなくて $V_1 \cdot C \cdot V_2$（V_1，V_2 は任意の母音，よう音もしくは無音，C は任意の子音）の音韻連鎖[17]ととることが望ましい．

この方式の問題点は，記憶される音声素片の長さがその最初の発音のピッチによって決まり，一定でないこと．したがって音声出力のピッチ制御に際しては記憶されている素片の長さと出力すべきピッチに対応する時間長の二つの情報が必要であり制御が複雑となること，素片の打ち切りまたは引き伸しによってピッチ制御を行なう結果，その波形ひずみによって極端なピッチ制御をすれば出力音声の品質が劣化することである．現状のままでは，許容できるひずみの範囲で制御できる時間長の幅は，記憶されている素片の時間長の±30%以内といわれている[16]．この二つの欠点を改良するくふうが当面の課題といえよう．

11. むすび

以上音声の認識および発生について，おもに実用の側面から従来の観点とはやや異なった見解を述べたが，音声の認識と発生は音声による情報の入出力手段として一対として利用されてこそ，真に実用上の価値を発揮することができる．音声応答は最近ようやく実

用化の機運が近ずきつつあるが，それらの実用化試験を通じて，改めて音声認識の必要性が再確認されつつあるといえよう．音声認識の研究はいわゆるパターン認識の一つとしても興味ある研究題目であり，今後，実用・学問の両面から盛んに研究されるであろう．

最後に音声の認識には，その話者の認識という別の側面があり，個人の確認の手段として注目されつつあることをつけ加えて，この解説を終わらせていただく．

参考文献

1) 板橋秀一，城戸健一："辞書と音形規則を利用した単語音声の認識"，日本音響学会誌，Vol. 27, No. 9, pp. 473～482 (1971).
2) 迫江博昭，千葉成美："動的計画法を利用した音声の時間正規化に基づく連続単語認識"，日本音響学会誌，Vol. 27, No. 9, pp. 483～490 (1971).
3) V. M. Velichko & N. G. Zagoruiko: "Automatic Recognition of 200 Words", Int. J, Man-Machine Studies, Vol. 2, pp. 223～234 (1970).
4) G. Fant: "Acoustic Theory of Speech Production", 's-Gravenhage, Mouton & Co. (1960).
5) J. L. Flanagan: "Speech Analysis, Synthesis and Perception" Springer-Verlag (1965).
6) 角川靖夫，中田和男："「合成による分析法」によるフォルマント周波数の抽出"，日本音響学会誌，Vol. 20, No. 1, pp. 1～3 (1964).
7) 板倉文忠，斉藤収三："統計的手法による音声スペクトル密度とフォルマント周波数の推定"，電子通信学会論文誌，Vol. 53-A, No. 1, p. 35～42 (昭和45年1月).
8) 板倉文忠："統計的手法による音声の特長抽出"東北大通研第8回シンポジウム論文集，音声情報処理，II-5 (1971年2月).
9) 板倉文忠，斉藤収三："最尤スペクトル推定法を用いた音声情報圧縮"，日本音響学会誌，Vol. 27, No. 9, pp. 463～472 (1971).
10) 松井英一，中島隆之，鈴木虎三，大村活："Kalman フィルタ理論による音声分析"，日本音響学会音声研究会資料 (1972年1月).
11) 藤崎博也："音声認識の諸問題"，日本音響学会誌，Vol. 28, No. 1, pp. 33～41 (1972).
12) 木村幸男，市川喜，中田和男他："音声応答装置"，情報処理学会誌，Vol. 12, No. 7, pp. 398～405 (1971).
13) 松井英一："音声素片のピッチ同期編集による多重化音声出力装置"，電気学会連合大会-2457 (43年).
14) 坂井利之他："零交差波による音声合成"，通信学会全国大会-161 (42年).
15) 板倉文忠，斉藤収三："PARCOR 形音声応答装置"，日本音響学会研究発表，3-2-4 (1970年5月).
16) 樽松明，井上誠一："ピッチ単位音声素片の録音編集による音声合成のシミュレーション"，日本音響学会研究発表，2-1-2 (1970年10月).
17) 斉藤収三，橋本新一郎他："音韻連鎖に着目した音声合成システム"，日本音響学会研究発表，1-3-16 (42年10月).

(昭和47年3月2日受 付)