U.D.C. 801.4:534.784:681.32.056

# 最近の音声認識技術

# Recent Speech Recognition Technologies

市川　熹*　*Akira Ichikawa*
畑岡信夫**　*Nobuo Hataoka*
北爪吉明**　*Yoshiaki Kitazume*
小松昭男**　*Akio Komatsu*

音声認識の主要技術課題である，(1) どこででも(使用環境条件の緩和)，(2) だれもが(話者の拡大)，(3) 連続的に発声した(自然な発声)，(4) どのような音声でも(語彙の拡大)認識できる装置を，(5) 小形かつ低価格に実現する諸技術について，日立製作所での開発状況を報告する。

まず，(1) と (2) が問題となる不特定話者電話音声認識方式について検討した結果，大局的特徴抽出法と標準パターン学習法を開発し，音声ダイヤル装置を試作した。次に，(3) 自然に発声した，(4) 任意の音声を仮名文字列に変換する連続音韻認識手法について論じた。ここでは，連続パターンマッチング法を用いた。最後に，音声認識装置の各処理について検討し，(5) の立場からLSI化手順の一案を提案した。

## 1　緒　　言

音声は人間にとって最も自然で便利な情報発生手段である。音声による情報発生速度は，キーボードのような打鍵による情報入力手段に比べ2～4倍と言われている。このような音声による入力装置を利用すると，特別の訓練なしに，動き回ったり，他の物を見たりしながら容易に情報を入力することができる[1]。最近の半導体技術，エレクトロニクスの発達と，それに良く整合した音声処理技術の開発により，音声入力装置もようやく実用化の段階に入った。

本論文では，音声認識技術の主要課題を明らかにし，それに対する最近の日立製作所での技術開発について述べる。

## 2　音声認識技術の開発動向

通常広く音声認識という場合，言葉としての音声認識だけでなく，だれが話した声かを認識する，いわゆる話者認識[2]なども含まれるが，ここでは紙面の都合上言葉としての音声の認識に対象を絞る。

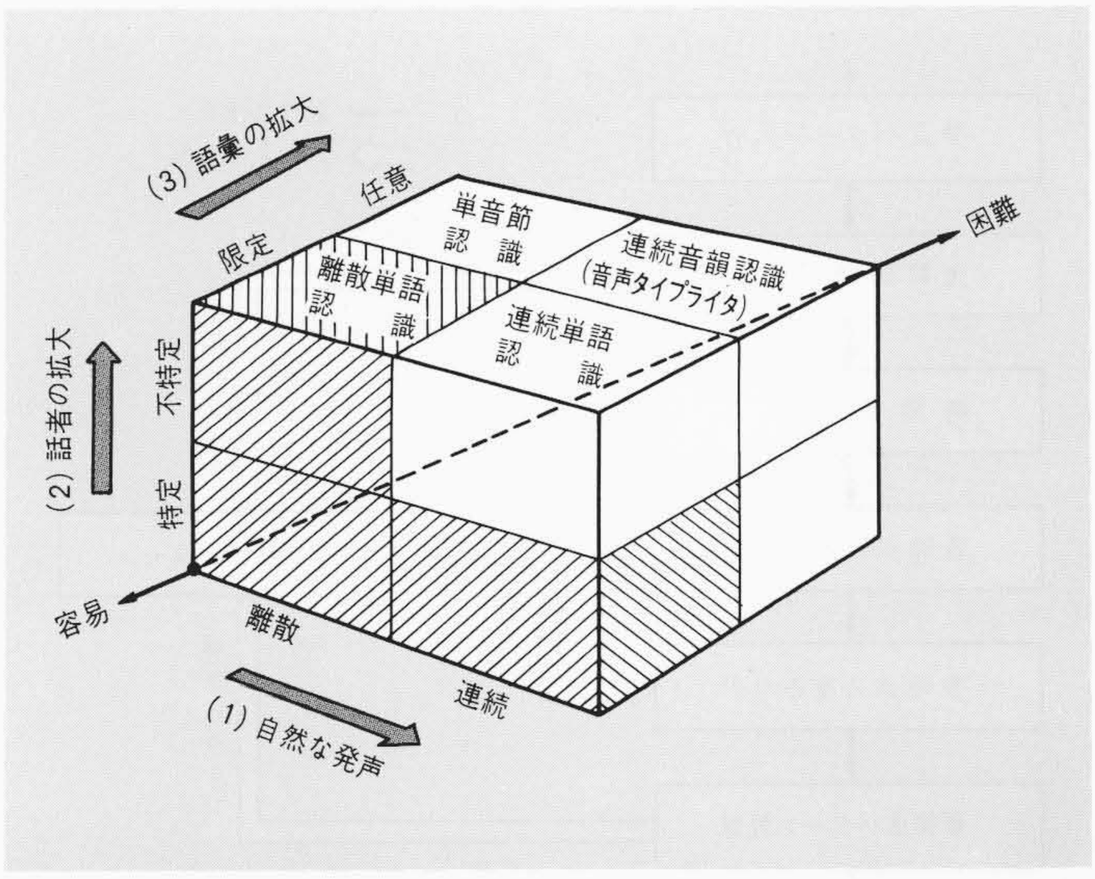



図1　音声認識の複雑さ要因と各方式　このほかに，(4)「環境条件」を考慮する必要がある。右上の組合せほど技術的に困難である。斜線部分の方式が実用化されている。

音声認識の各方式の関係を**図1**に示す。同図中の三つの軸は，音声認識の複雑さを示す三つの要因を示す。発声法，話者，対象語彙に対する制限の大小で分類して示した。このほかに第四の要因として使用環境(周囲雑音，マイクロホン，電話など伝送系条件など)がある。

**図1**中ハッチした部分が実用の段階にある。まず実用化されたのは特定話者の離散発声限定単語認識である。最近では限定された単語を数個連続発声した音声をも認識する装置や，不特定話者用の離散発声限定単語認識の一部が実用化段階に入りつつある。しかし，前者では連続発声とは言え，かなり丁寧に発声する必要がある。特に連続発声が要求される数字音声では，単語を構成する音節数が少ないため，連続発声による変形が大きく認識率が低下する。また後者についても，認識可能な語彙を変更するためには多数の話者の大量なデータ処理による標準パターンの作成が必要であり，制約は大きい。

このように，話者の制約を少なくする技術(話者の拡大)並びに音声入力の良さを生かすために連続発声した音声を認識する技術(自然な発声)，認識可能な語彙を増やす技術(語彙の拡大)及び使用環境条件の緩和技術の開発が課題となっている。

更に，装置実現のためのハードウェア技術としては，ＬＳＩ技術との関連を見落とすわけにはゆかない。これには三つの側面がある。

その第一は，音声入力装置がその性格上必ず人がついて利用されるため，高い稼動率が期待できず，コスト的条件が厳しいという点である。低コスト化の手段としてＬＳＩ技術が注目されるゆえんである。その第二は，音声処理技術がＬＳＩとの整合性が良いという点である。第三は，処理内容が高度化するにつれて専用ＬＳＩによる処理能力の増大を図らざるを得なくなるであろうという点である。このように，ＬＳＩ技術をどう取り入れてゆくかも今後の大きな課題である。

## 3　不特定話者電話音声認識

電話音声を条件としているため，単に話者が不特定であるというだけでなく，電話系を経た音声であることも考慮する必要がある。電話系は帯域が制限されている(300～3,400Hz)

* 日立製作所中央研究所 工学博士　** 日立製作所中央研究所

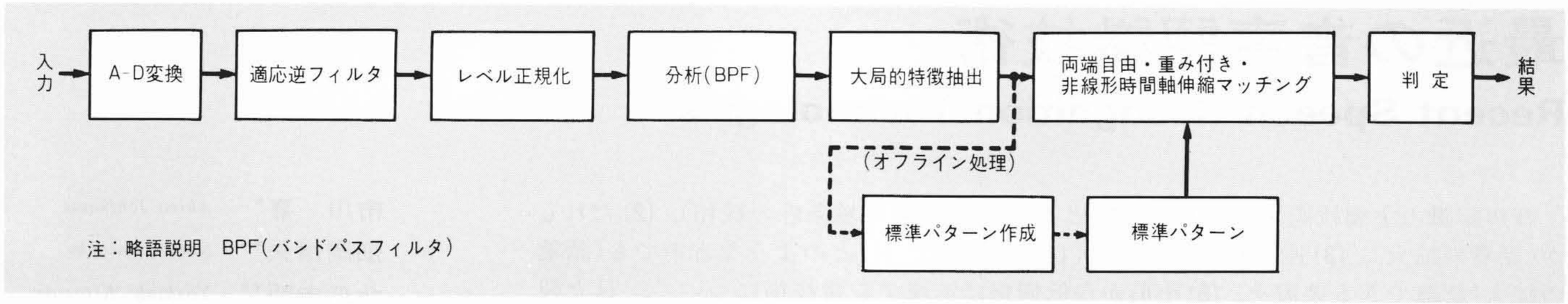



図2 不特定話者電話音声認識装置の処理の流れ 点線は標準パターン作成時のルートである。

こと，S/N(信号対雑音比)の条件が不利なこと，送話器のひずみが大きいことなど認識に不利な条件が多い。言い換えれば，前章で述べた音声認識の複雑要因のうち，第二及び第四の要因に課題のある装置である。

### 3.1 構 成

離散発声限定単語方式とし，パターンマッチング法による認識方式を採用した。**図2**に処理の流れを示す。

電話機から入力された音声はA-D変換器でディジタル信号に変換される。入力された音声は，発声者の個人差や電話系によりスペクトルの概形(全体的な傾斜)が異なるので，一次の適応逆フィルタ[3]によりほぼ一定(平坦)となるようにそろえる。次にレベルの正規化を行なう。音声の内容によってレベルが変動することに対処し，以降のディジタル演算処理精度を確保することを目的としている。適応逆フィルタ処理及びレベル正規化処理は，25ms分のデータ(分析区間長さ)をもとに12.5ms(分析時間間隔)ごとに実行している。

分析部は$Q=5$程度のバンドパスフィルタ群である。各フィルタは300～3,400Hzの電話帯域に対数間隔に16チャネル配置されている。実際には1個のディジタルフィルタを時分割で用いている。$Q=5$と低い値に取った理由は，ホルマント周波数などの周波数軸上の個人差をある程度吸収することと，次に述べる大局的特徴抽出時に音声の特徴を確保するためである。フィルタ群の出力は検波後低域通過フィルタを通して平滑化し，12.5msごとに特徴として抽出される。

音声の入力が終わり，フィルタの分析処理が終了すると，その結果を用いて次に説明する大局的特徴抽出法により，入力音声の特徴を求める。

まず，フィルタ群の$j$チャネル目の$i$番目の分析時点の出力値を$A'_{ij}$とし，次式に示すような非線形変換を行なう。

$$A'_{ij}=\log\left(1+\frac{A_{ij}}{A_0}\right) \quad \cdots\cdots(1)$$

この処理は$A_0$を定数として，$A_0$よりも大きい値に対しては対数的圧縮が行なわれるが，$A_0$よりも小さい値に対してはほぼそのままの値が保たれる変換である。この変換は大振幅部分の圧縮特性により，大振幅のパルス性雑音や入力音声のダイナミックレンジのばらつきの影響を抑える一方，微小振幅入力に対しては線形を保つことにより，不要なS/N劣化を防止する特性となっている。

次に，各フィルタのチャネル$i$ごとにその出力の最大値で正規化を行なう。

$$A''_{ij}=\frac{A'_{ij}}{\max\limits_{i}\left[A'_{ij}\right]} \quad \cdots\cdots(2)$$

この処理は，音声が本来時間構造をもったパターンである点に着目した処理となっている。バンドパスフィルタの$Q$が低いため隣接チャネル間の出力は相関が高く，この処理によりスペクトル構造は悪影響を受けることはない。これに対し，周波数軸方向(チャネル番号$j$方向)に正規化すると時間構造に大きな悪影響を与える。

このようにして得られた入力音声の特徴パターンは，標準パターンと，語頭・語尾の位置にある程度の幅を許した両端自由非線形時間軸伸縮(NL)マッチングを行ない，各標準パターンとの類似の程度が評価される。両端自由とした理由は，電話系を経た音声は回線雑音などを受け音声区間の切り出しが不安定になりやすいため，その悪影響を軽減するためである。

各標準パターンは，あらかじめ多数の発声者により発声された音声から作成される。**図3**に標準パターン作成手順を示す。

まず認識すべきすべての単語について，任意の一発声ずつを大局的特徴抽出法により分析し，初期標準パターンとして登録する。これを第1回目の既学習パターンと見なす。2回目以降の学習パターンの処理は次のようになる。各パターンは大局的特徴抽出が行なわれた後，既登録標準パターンと非線形時間軸伸縮マッチングを行なう。学習回数による重み付きの時間軸対応を行ない，新標準パターンの時間構造を求める(時間構造の学習)。この対応関係に従い，既登録標準パターンと追加入力パターンの各特徴から学習回数を重みとした重み付き平均を求め，新標準パターンの特徴とする(特徴の学習)。このとき，併せて新標準パターンの各時点$t$ごとにそれまでに学習したパターンのばらつきの程度を分散$\sigma_{it}^2$の形式で求める。更に，他のカテゴリー(単語)の標準パターンとのマッチングを行ない，他パターンとのマッチングのばらつきの度合を分散$\sigma_{mt}^2$の形式で求め，分散比$\sigma_{mt}^2/\sigma_{it}^2$の関数として

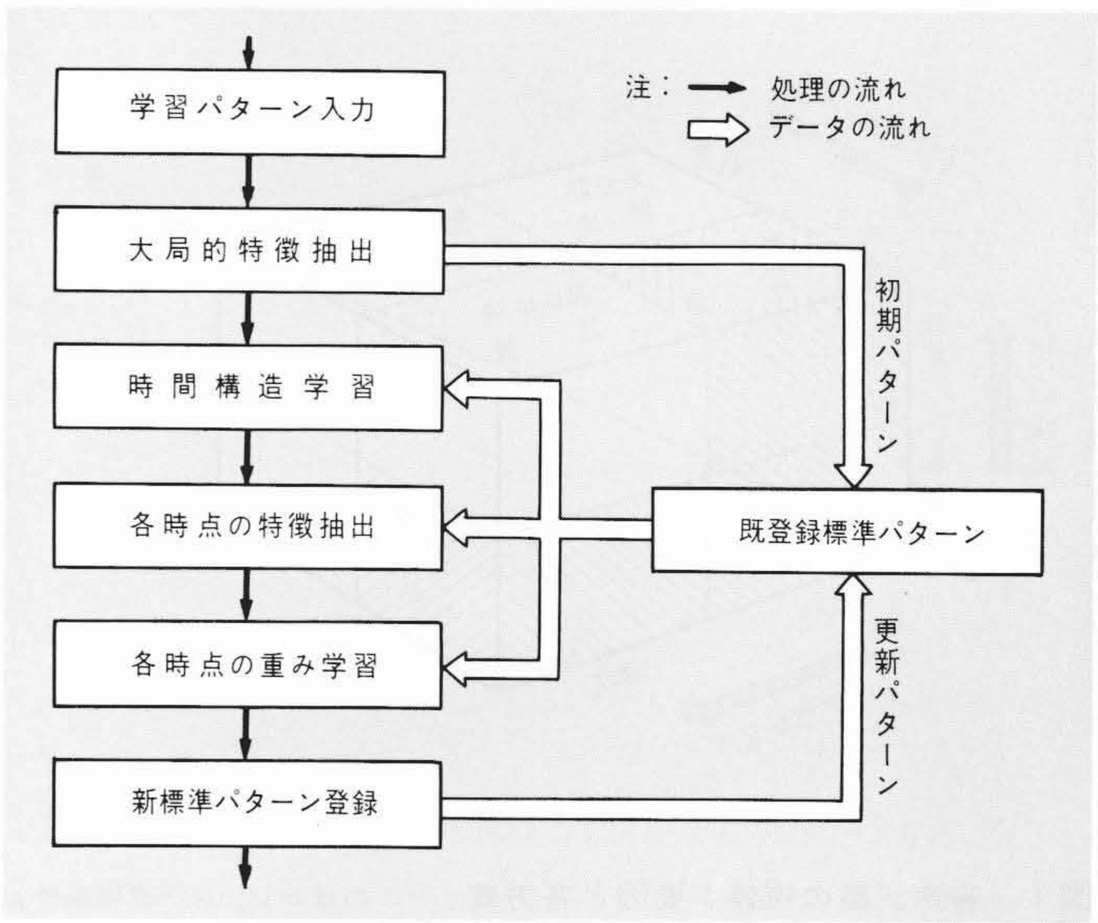



図3 標準パターン作成手順 時間的パターンである音声の性質を考慮した特徴抽出を行なった後，発声ごと，話者ごとに異なる音声の時間構造と各時点の特徴を学習してゆく。更に，単語間の区別に有効な程度を各特徴点ごとに評価し，重み係数とする。各特徴量と重み係数を，学習した時間構造の順に並べ，標準パターンとする。

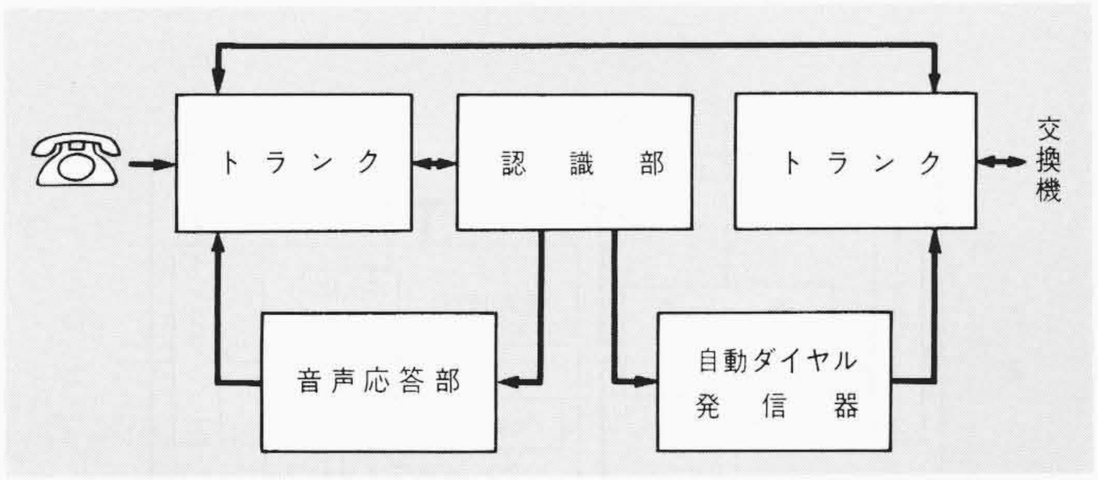



**図4　音声ダイヤルの構成図**　不特定話者電話音声認識装置の一応用例である。

標準パターンの各時点 $t$ の重み $w_t$ を求める(重みの学習)。$w_t$ は，各時点 $t$ が他のカテゴリーとの認識区別に寄与する度合となっている。

**図2**での両端自由非線形時間軸伸縮マッチングでは，この重み $w_t$ を考慮して入力パターンと標準パターンのマッチングが行なわれる。

標準パターンは認識率を上げるために，一つの単語に対し複数個用意することも可能である。

### 3.2　音声ダイヤルへの応用

音声ダイヤル装置に応用した例を**図4**に示す。音声ダイヤルは，電話番号をダイヤルする代わりに相手先名を音声で電話機から入力することにより，自動的に電話のかかる装置である。送話器を取り上げると，まず自動的に音声認識部に接続される。ＡＲＵ(音声応答部)からの誘導音声「相手先をどうぞ」に従い相手先名を入力すると，認識結果が制御部経由でＡＲＵに送られる。ＡＲＵは認識結果の確認を求める音声「……ですね」を出力する。結果が正しければ「はい」を，誤りであれば「いいえ」を入力する。音声認識部は「はい」と認識すると自動ダイヤル発信装置を起動し，相手先電話番号を交換機に向け発信し，トランクを相手方と通話可能な状態とする。「いいえ」と認識した場合は，ＡＲＵから「もう一度相手先をどうぞ」という音声を出力し，入力を待つ状態にもどる。

このように音声ダイヤルを使うと，一々相手の電話番号をダイヤルする(コード変換する)必要がなく，電話をかけたいと思う相手先名をそのまま発声すればよい。その意味で，音声入力の最も良い面を生かしたシステムの一例と言えよう。

## 4　連続音韻認識

連続音韻認識は音声認識の複雑さ要因のうち，第一(発声法)と第三(語彙の拡大)に課題のある方式である。

音声入力の利点を生かしながら任意の内容を入力するためには，自然に発声された音声を音韻のレベルで認識する技術が必要である。これが連続音韻認識と呼ばれるものである。単語を単位とした連続音声の認識方式では，あらかじめ登録してある単語を組み合わせた音声しか認識できない。また音節を単位とした方式でも，離散発声を前提としたものは音声入力の利点である使いやすさが発揮できない。

### 4.1　連続音韻認識の困難さ

連続音韻認識が困難な主な理由を挙げると次のようになる。

まず第一に，連続音声中では音と音の境界が不明確な点である。口や喉，舌などの発声器管は，物理的・生理的制約から階段的には変化できず連続的に動くため，そこから生成される音波である音声もまた連続的に変化せざるを得ない。

第二の理由は，各音はその前後にくる音の種類や発声の速度で，その物理的性質が相互に重なり合うほど大きく変化する点である。音が脱落したり(母音がなくなる無声化現象など)，無声子音が有声音に変わるなどの現象も現われる。

第三の理由は，各音韻は100ms程度の非常に短い継続時間しかなく，類似した音を区分するだけの十分な情報を得ることがなかなか容易でないという点である。

更に，方言などの習慣の差によっても同一音韻に対する物理音響現象に差が現われ問題を複雑にしている。これらの問題が複雑に絡み合って，連続音韻認識の実現を非常に困難なものにしている。

### 4.2　音声タイプライタ

このような問題を克服し，自然に連続的に発声された任意内容の音声を認識できる装置こそ，音声タイプライタと呼ばれるにふさわしいものと言えよう。しかし，専門家の間ではその実現は21世紀に入っても困難ではないかと言う見解が一般的である。

この夢の実現への挑戦の第一歩とも言うべき試作装置について紹介する。この装置は1980年秋東京で開催された「日立技術展」で，音声タイプライタプロトモデルとして公開されたものである。

この装置では，前後の音による影響を考慮した音声の単位を標準パターンとして準備し，入力音声に対し連続的にマッチングさせながら，自然に発声された連続音声中の音韻を認識してゆく構成となっている。音韻が前後の音により影響を受けるということは，逆に前後の音にもその音の情報の一部が存在していることを示している。したがって，前後の音の影響を考慮した単位を標準パターンに選ぶということは，連続音韻認識の困難さの第二及び第三の原因に対処することを意味している。ここでは，母音－子音－母音(ＶＣＶ)[4]単位や子音－母音(ＣＶ)単位などを状況に応じて用いている。

連続パターンマッチングは，入力音声を構成する音の境界を意識することなく，連続的に処理してゆくので，連続音韻認識の第一の困難さを避けることができる。ここでは音韻の性質により連続ＤＰ(ダイナミックプログラミング)マッチング法[5]を変形した手法のほかに連続線形マッチング法を提案し，この両者を使い分けて最適な認識が行なえるように制御している。**図5**は入力音声/akaneiro/に対し，標準パターン/aka/を連続的にマッチングさせた結果を示している。マッチング結果は，入力パターンに沿って連続的に出力され，入力に同一パターンが生じた時点で値が小さくなっているこ

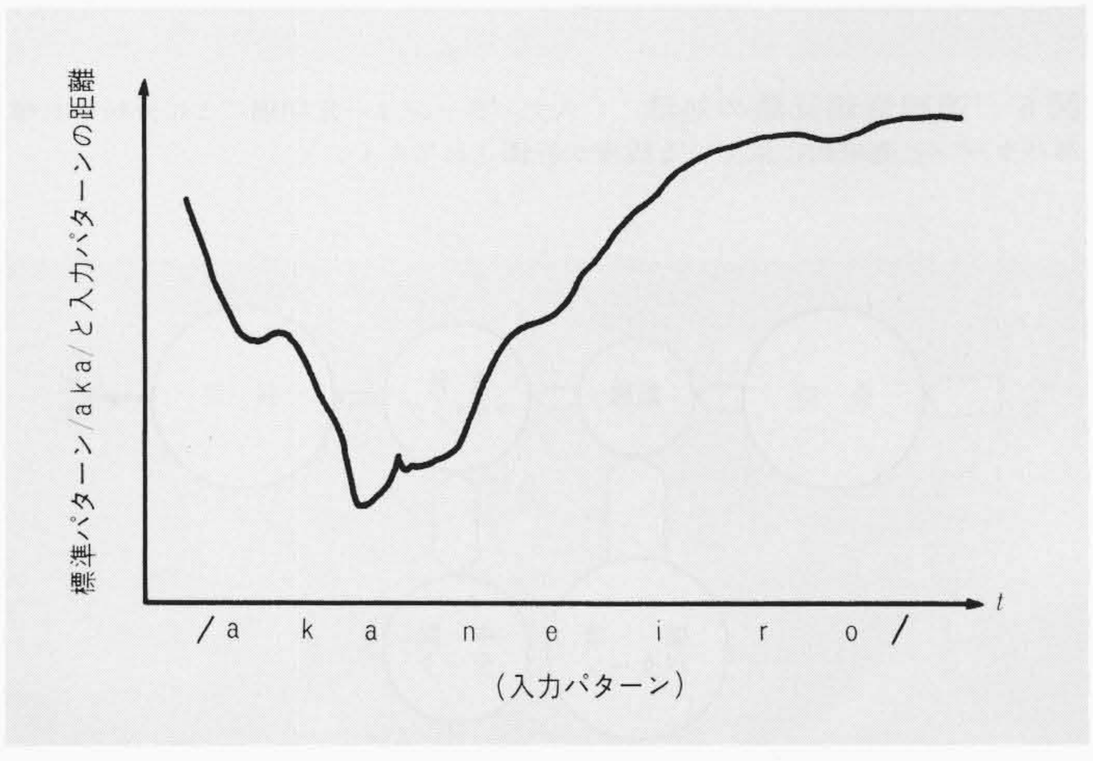



**図5　連続マッチング部出力の例**　入力パターン/akaneiro/の2番目の/a/が入力されたとき，標準パターン/aka/との距離が最小となり，入力音声中に/aka/が存在していたことが分かる。

とが分かる。

装置の処理の流れを**図6**に示す。入力音声は分析部で音声の特徴パラメータに変換され，距離演算部で標準パターンの各部分との距離が計算される。この結果は連続マッチング部に送られ，ここで最適な対応を考慮した入力パターンと各標準パターンの類似の程度が入力の時間に沿って連続的に評価されてゆく。判定部では，連続マッチング結果が一定以上類似していると判断される候補の中から，前後関係などを考慮して最適な音韻を選択し認識結果として出力する。

この装置は，まだ話者は特定の人に限定されているが，ほぼ実時間で90%程度の高い音韻認識率が得られている。しかし，90%の音韻認識率でも10音韻(ほぼ仮名5文字)の単語としての認識率に換算すれば，35%(≒$0.9^{10}$)程度に低下する。更に長期にわたる改良研究が必要と思われる。

なお，これらの技術は連続単語認識装置にも適用することができる。

## 5 音声認識装置のLSI化

**図7**は**図6**の連続音韻認識処理を例に，各部の処理の複雑さを円の大きさで，入出力データの量をデータの流れの幅で大まかに描いたものである。同図から分かるように，入出力のデータ量は相対的に少ないが処理が複雑な部分(分析部と判定部)，入出力データ量が多いが処理内容が比較的単純な処理の繰り返しである部分(距離演算部とマッチング部)の二つに性格分けされることが分かる。前者は汎用プロセッサ向きの，後者は専用LSI向きの性質と言えよう。この点を考慮すると**表1**に示すような構成を考えることができる。分析部は信号処理用汎用プロセッサDSP(Digital Signal Processor)，判定部はマイクロプロセッサが適している。マッチング部は専用ＬＳＩがよい。**図8**は連続非線形マッチング用ＬＳＩの構成例である。連続非線形マッチングのほかに連続線形マッチングも処理可能な構成を想定している。

分析部はディジタル電話のように，音声合成器と対で用いるシステム用には，ピッチ周期など音源パラメータの分析も同時に行なう必要がある。この場合は，内部ＲＡＭ(Random Access Memory)容量の制限などの問題も生じ，専用ＬＳＩを検討する必要が生ずる[7]。この専用ＬＳＩは音声認識装置の分析部としても利用が可能である。

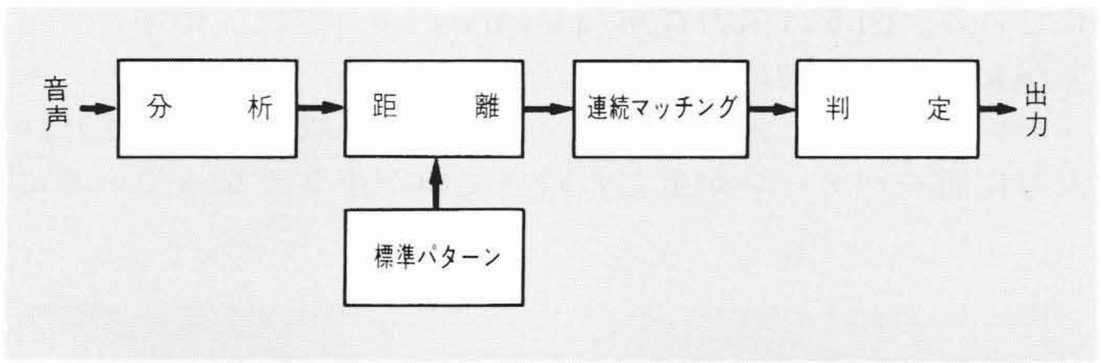



図6　連続音韻認識の処理　入力パターンは一定間隔ごとに分析され標準パターンと連続的に似ている程度が評価されてゆく。

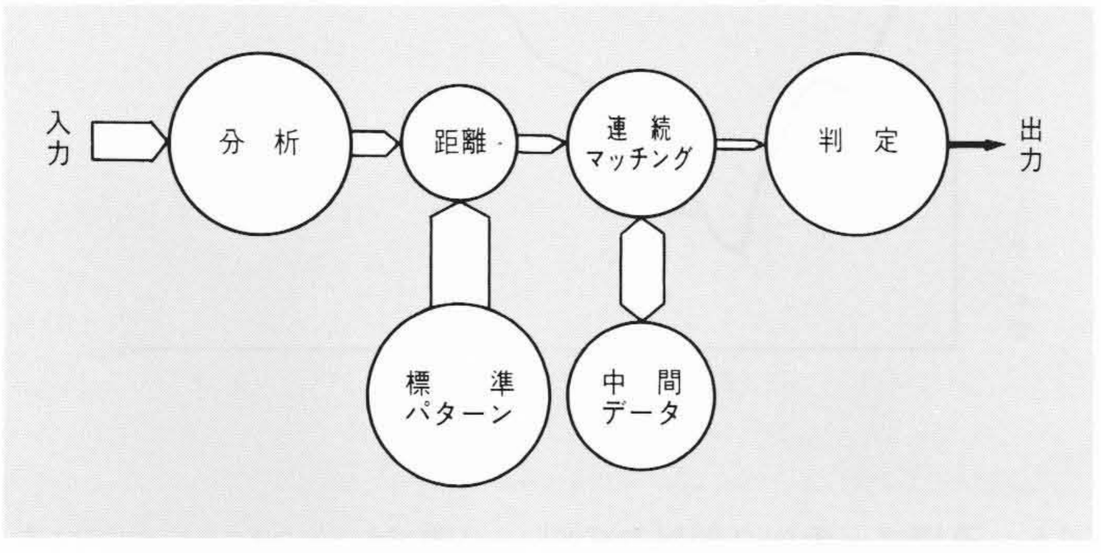
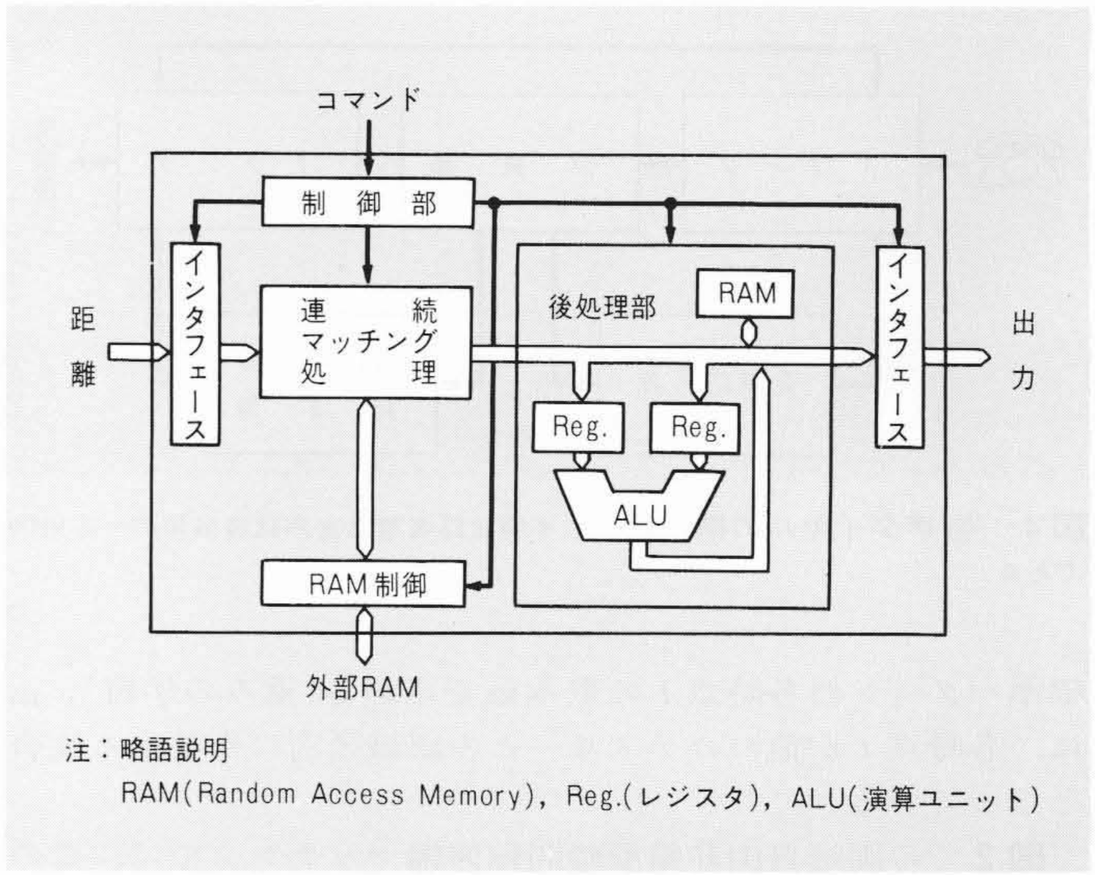



図7　連続音声認識の各処理の性格　各処理部の円の半径は処理の複雑さの程度を，矢印の幅はデータの流れの量を定性的に示している。




注：略語説明
RAM(Random Access Memory)，Reg.(レジスタ)，ALU(演算ユニット)

図8　連続非線形マッチング用専用LSIの構成例　連続マッチング部で入力パターンと標準パターンの最適な対応づけなどを行ない，後処理部ではパターンの長さの相違による結果の正規化などを実行する。

表1　認識装置のLSI化の一形態　図7から各部の性格を考慮し，LSI化の形態を検討した一例である。

| | 分析部 | 距離部 | マッチング部 | 判定部 |
|---|---|---|---|---|
| 性格 | 複雑な処理<br>一様なデータの流れ | 単純な処理<br>大量データ | 一様な処理<br>大量データ | 複雑な処理<br>データ量は少ない。 |
| 実現方法 | 信号処理用<br>マイクロプロセッサ | 専用LSI化 | 専用LSI化 | マイクロプロセッサ |

## 6 結　言

音声認識では，「使いやすさ」という原点を無視したシステムは，音声入力のもの珍らしさがなくなるにつれて存在価値はなくなる。また，低価格小形システムであることを本質的に要求される宿命をもっている。これらの要求に本当に応ずることのできる装置の実現には，まだまだ研究開発が必要であるが，実現の暁には真に効率的なマンマシンシステムが出現し，その発揮する効果は大きい。

## 参考文献

1) 新美：音声認識，情報科学講座E19-3，共立出版（昭54-10）
2) 市川，外：電話音声を対象とした話者照合，日本音響学会誌，35 (2)，(1979-2)
3) 中島，外：適応逆フィルタ法による声道断面積関数の推定，日本音響学会音声研究会資料（昭48-2）
4) 中津，外：ＶＣＶ音節を単位とした連続単語音声の認識，日本音響学会研究発表会講演論文集，2-2-18（1974-10）
5) 岡：連続ＤＰを用いた連続単語認識，日本音響学会音声研究会資料，S78-20（昭53-6）
6) A. Ichikawa et al.：Conceptual System Design for a Continuous Speech Recognition LSI, ICASSP 81 E5 (1981-3)
7) 浅田，外：Le Roux型格子法によるPARCOR音声分析認識装置の試作，電子通信学会電気音響研究会，EA80-81（1981-2）