E4　音声識別の一方式

石井威望（東京大学）
堀川　甫・阿部行信・浦野由多加（富士通）

1　まえがき

文字・図形・音声などパターン情報処理は電子計算機の今後の利用と技術開発にとって重要な課題であるといえる．

人間と機械との対話を考えるとき音声情報の利用は，それが人間にとってもっとも自然でより能率の良いものであり，しかも既存の伝送，交換設備にとっても受け入れやすいものであるだけに，きわめて有効な手段といえよう．

音声情報処理としては，音声出力と音声入力がある．音声出力に関しては，音声合成や音声編集による計算機出力として技術の開発が進み実用化の段階に達している．特に電話，テレビ電話，有線テレビ（CATV），教育用機器（CAI）への情報出力として大きな効果が期待できる．一方，音声入力に関しては，パターン認識という大きな課題をかかえ医学，物理，電子工学など各分野での積極的な研究が行なわれているが現在のところ満足できる結果は得られていない．しかし，この音声認識の研究は，人間－機械系，音声タイプライタ，自動通訳装置，声紋識別など，また波形情報処理として脳波や心音，心電図の解析，その他に応用できるものであり多くの効果を得るものとして期待できる．

今回，音声認識の研究の1ステップとして100単語程度の識別の可能な一方式を実験したので報告する．

2．レジスタード方式

人間が音声を識別する過程と仕組みは，まだよく解明されていない．しかし音声は，時間軸における周波数成分とエネルギーの変化であり物理量としてとらえることができる．人間による識別は，音声の物理的な特長を抽出し言葉の文法や概念との結びつきを学習によって記憶しておき処理しているものと思われる．したがって機械による識別でも，この物理量を何らかの形に分解し，特長を抽出して記憶させておくことは必要であるといえる．

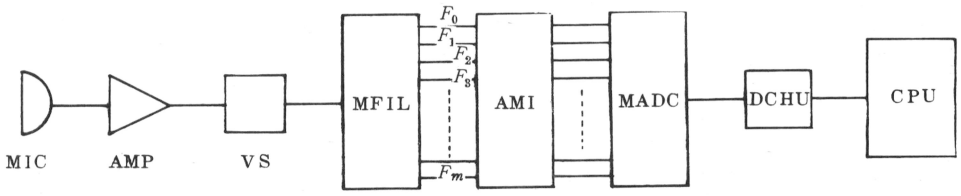
ところが，多数の人間が発声する同一の言葉の中に持つ共通の特長とは何か，またその差異は何か，同一の人間が多数の言葉を発声したときにその違いをどのように抽出すればよいか，多数の人間が多数の言葉を発声したときどのように対処すべきか，さらに言葉が複雑に組み合わさり文章となったときどのような処理が必要となるか，など音声の識別には解決しなければならない多くの問題が残っている．

このため今回の実験においては，発声する人間と識別する言葉を限定しておき，その言葉の

特長を抽出しパラメータとして登録しておく．識別は登録してある人間の登録されている言葉を対象としてのみ行なうというものである．これを「レジスタード方式」と名づけることにする．このように登録された人間だけが所定のサービスを受けられる形態も，システムとして効果のあるものと考えられる．

3. 音声入力処理

この実験におけるシステムのブロックダイヤグラムを図−1 に示す．



MIC ：マイクロホン
AMP：増巾器
VS ：音声スィッチ
MFIL：多チャネル帯域フィルタ
AMI ：多チャネル音声蓄積装置
MADC：多チャネルA／D変換装置
DCHU：データチャネル装置
CPU ：中央処理装置

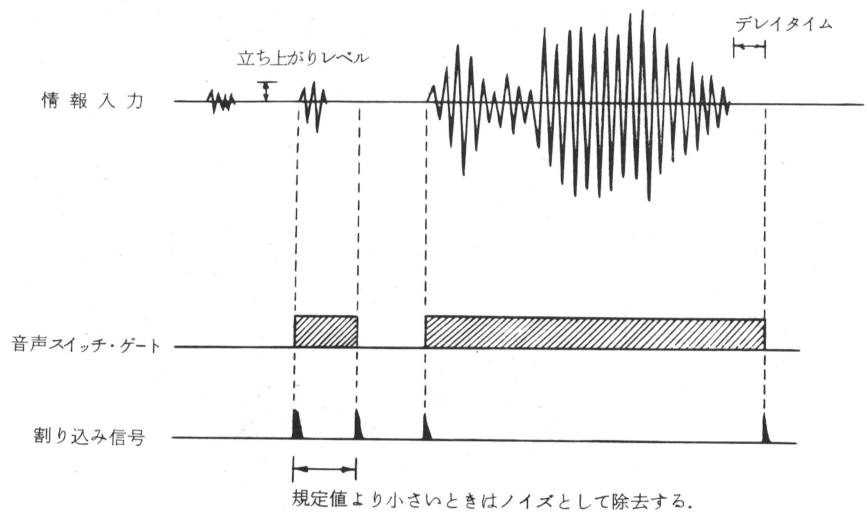
図−1　ブロックダイヤグラム

1. 音声情報部の抽出

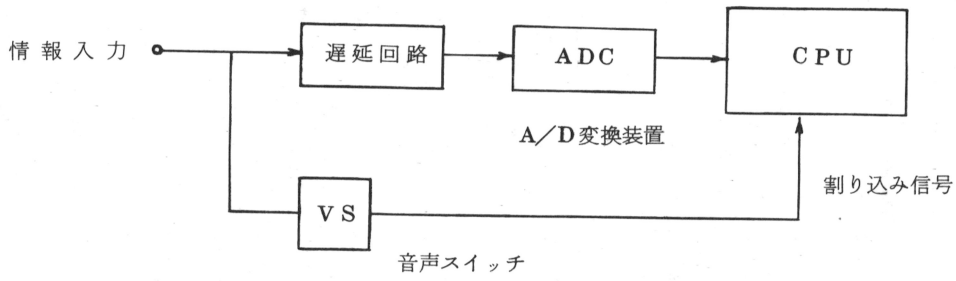
不定の時刻に入力される情報を計算機で処理する手段として，音声が発声された開始点と終了点を検出し，その時間帯だけ入力処理を行なうために音声スイッチ（VS）を設けた．

音声入力があると，音声レベルがある規定値以上になった点を開始点として計算機に割り込みをかける．この割り込みにより計算機は音声入力があることを知り，A／D変換装置に起動をかけて入力情報を収集する．そして規定値以下のレベルがある一定時間継続したことにより終了を知り情報収集を停止する．

なお，入力継続時間の判定をすることにより，クリックやせきなどの雑音を除去することも可能である．また立ち上がり点の頭部情報が欠けるのを防ぐため図−2 (b) のように遅延回路をそう入することも考えられる．



図－2 (a)　音声部の抽出



図－2 (b)　音声部抽出のブロック図

4. 情報の前処理

音声情報などアナログ量を計算機で処理しようとするとき考えなければならないことはその情報の前処理であろう.

原情報を直接，計算機に入力するよりもあらかじめ前処理を施すことにより多面的に分離したり，計算処理がしやすい形に変形するなどにより能率の良いものとすることができる．例えば，

1. 帯域フィルタによる周波数成分の分離.
2. 非直線増巾器により波形に歪(ひずみ)を与える.
3. 積分器により情報の圧縮を行なう.
4. 変調により波形の性格を変える.
5. その他各種解析装置による処理.

などが考えられる．

このように音声などアナログ情報を扱うときにはその処理に有効な前処理を施すことの重要性が強い．

したがって方式決定のためのシュミレーションの段階においては，数値フィルタや数値微分などのプログラム手法を導入して前処理の方法を早く見つけることが必要であるといえる．

今回の実験においては帯域フィルタ（MFIL）により周波数成分を分離し，その情報を積分器（AMI）により圧縮する方法を採用した．

5. A／D変換

A／D変換で問題となるのはデータのサンプリング密度と変換精度である．

データのサンプリングについては普通，入力データの持つ周波数の2倍の量子化が必要である．したがって，波形変化の特定部分の検出を高精度で行ないたいときや無情報部や雑音部などの時間帯のみサンプリング値を低くしたいといった場合もあり，任意に選択できるようになっていることが望ましい．

A／D変換装置にはサンプリング値が固定のもの，外部からの同期信号によるもの，プログラム制御によるものなどがあり目的によって使い分けることが必要である．

またマルチプレクサによる多点入力の走査が行なえることも必要条件である．

変換精度としては，8～10ビット（0.4～0.1％）程度が普通である．

6. 前処理方式の決定

以下に今回の実験で採用した前処理の方式について述べる．

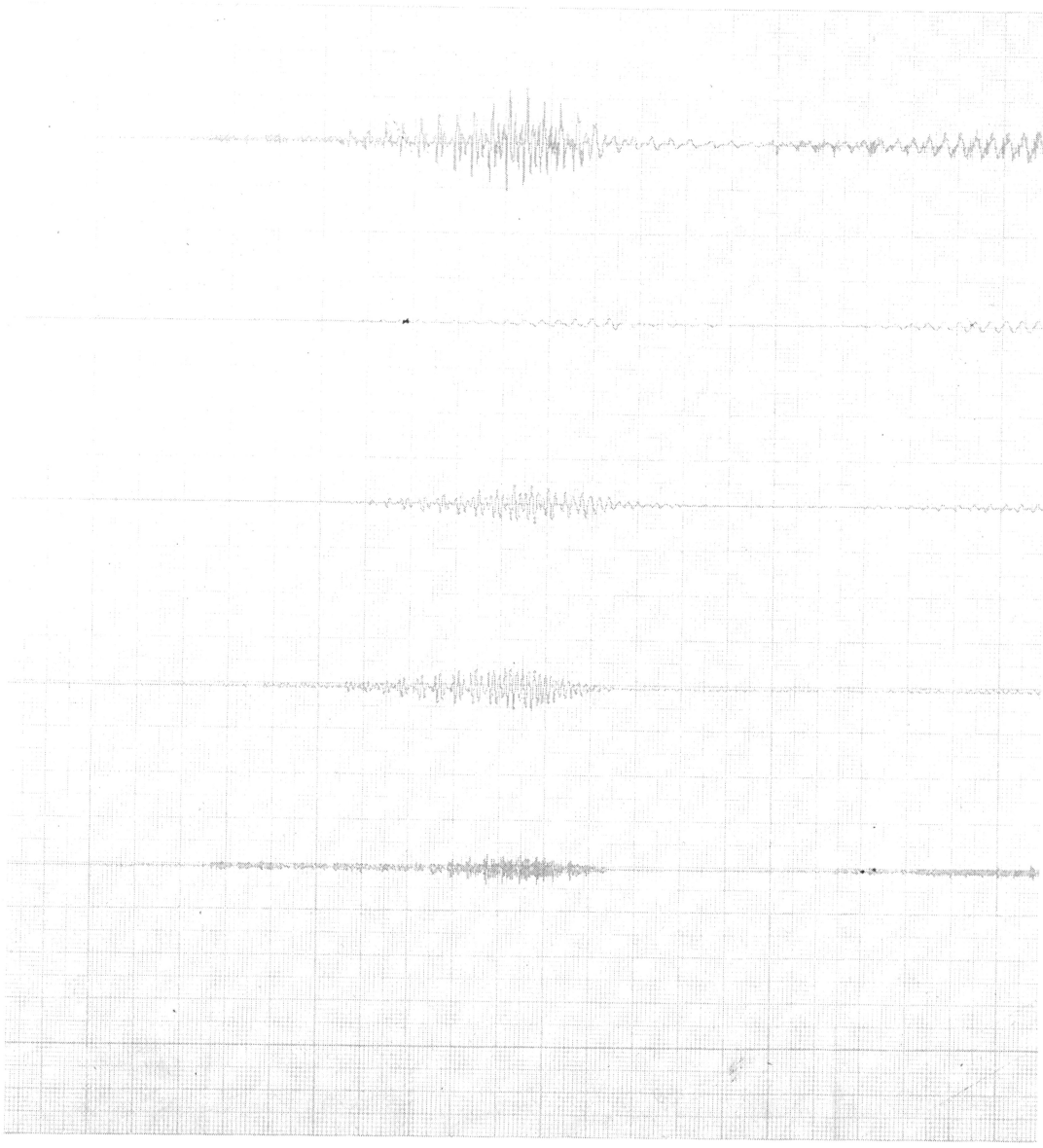
図－1のブロックダイヤグラムにおける，帯域フィルタ（MFIL）を通った後の音声情報の波形をA／D変換しX－Yプロッタで表示したものを図－3に示す．図－3の波形出力は上から図－4のような周波数成分をとり出したものである．

すなわち，F_0の出力波形はF_1からF_4の全帯域をカバーしている．

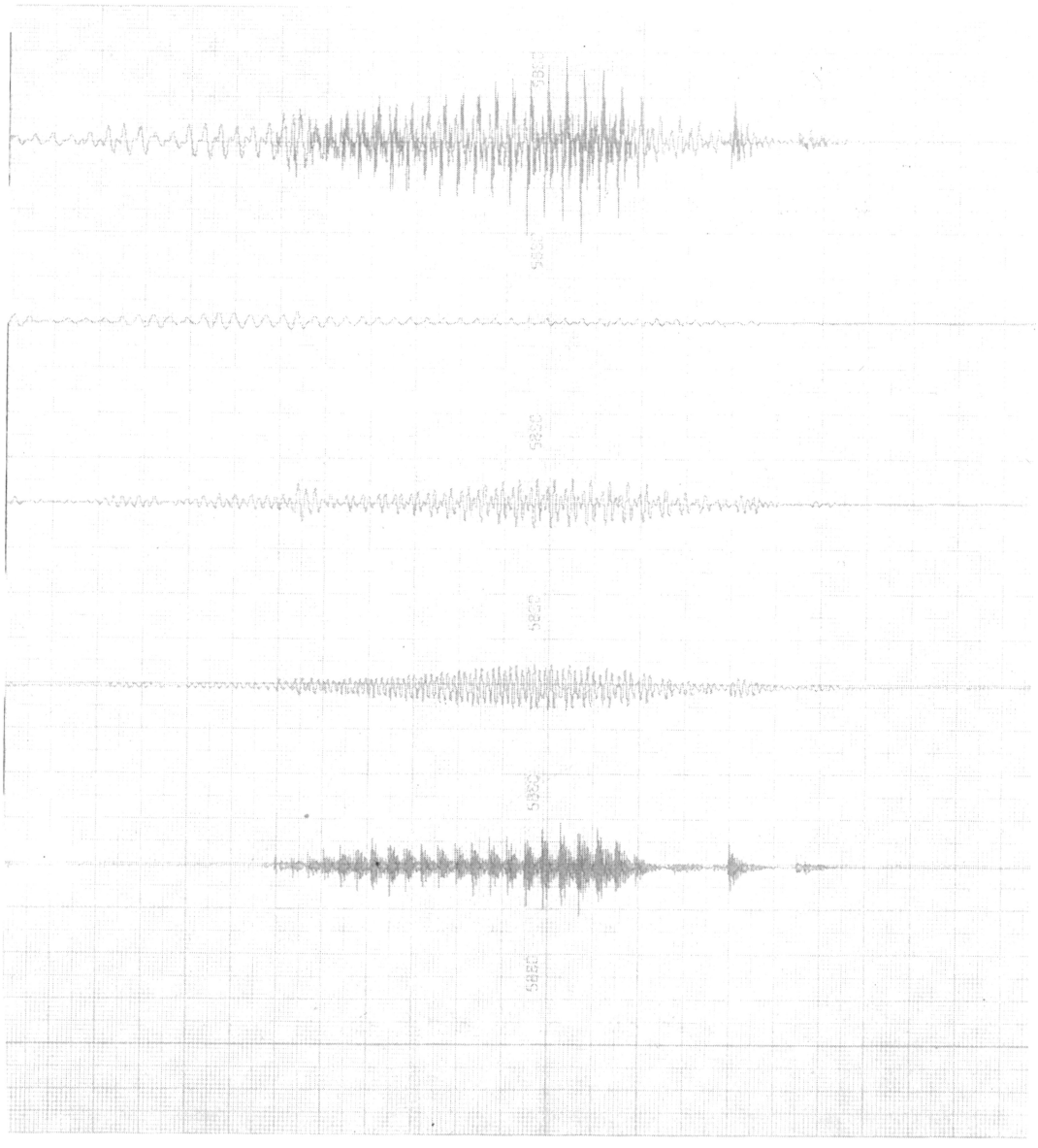
この波形データは，4000Hzまでの周波数成分を含んでいるため，8000Hzのサンプリングでで A／D変換を行なう必要がある．したがって例えば5チャネルの情報量は8000×5データ／秒の大きさとなる．このようにデータ量の増加は，計算機のデータ転送速度，記憶容量，二次記憶のアクセス・タイム，プログラムの処理時間などの制約から多くの問題が生ずる．この情報量を圧縮する手段として積分器を想定し，その積分装置の設計に次のような方法をとった．

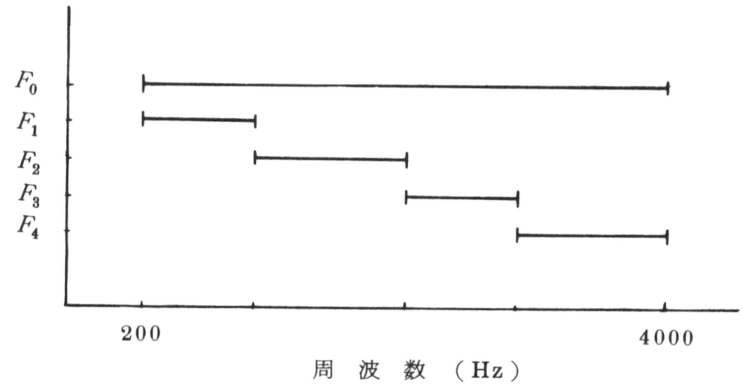
積分器からの出力波形を図－5に示す．この積分時間Tとその積分値E_{ij}（$i=0, 1, \cdots m$ $j=1, 2, \cdots, n$）をA／D変換し積分値をリセットする時間の決定が必要となった．このため多くの音声情報について積分値を変化させた図－6のデータを収集した．

左のグラフは出力レベルE_{ij}の各チャネルの表示であり，右のグラフは$F_1 \sim F_m$の出力値を



図－3　音 声 波 形





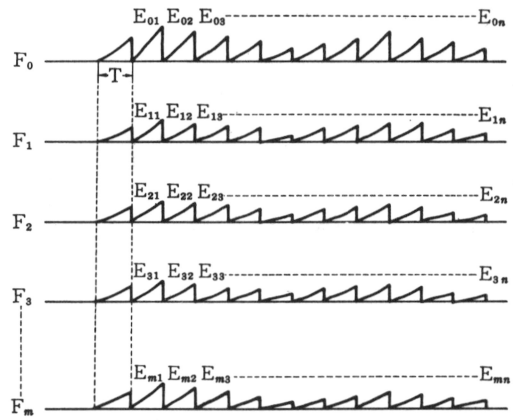
図－4　周波数帯域分離

F_0の出力値で正規化したものでE_{ij}/E_{0j}の変化を示す．

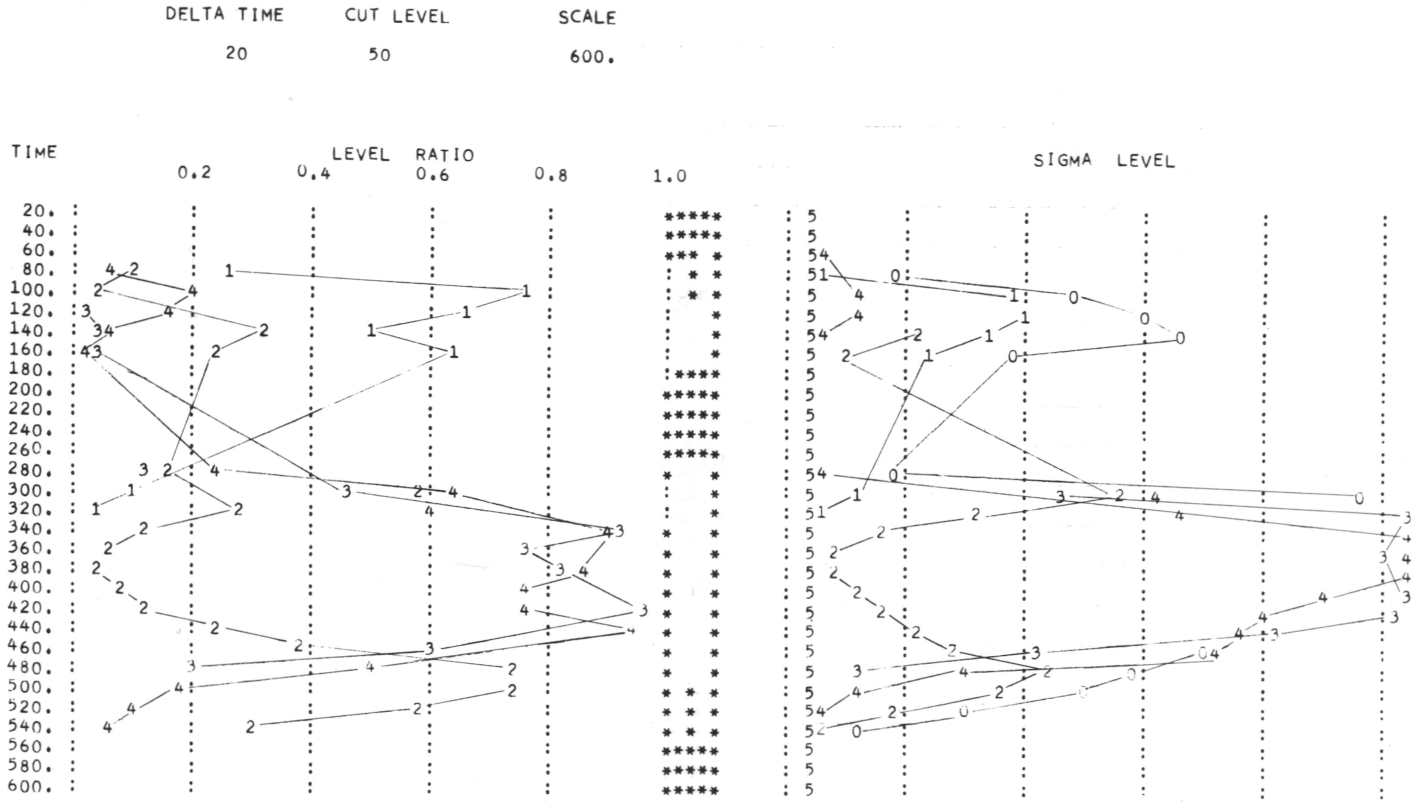
このグラフ出力を判定し積分装置の特性を決定する資料とした．

またこのデータは，音声情報の周波数成分の分布を知るための貴重な資料となった．

計算機による波形の観察として，プロッタやグラフィック・ディスプレイによる方法があるがこのようなグラフ出力も簡易で効果のある手段である．



図－5　多チャネル積分器出力波形



(a) 正規化レベル　　(b) 絶対レベル

図－6　音声レベル変化

7. パラメータの検出

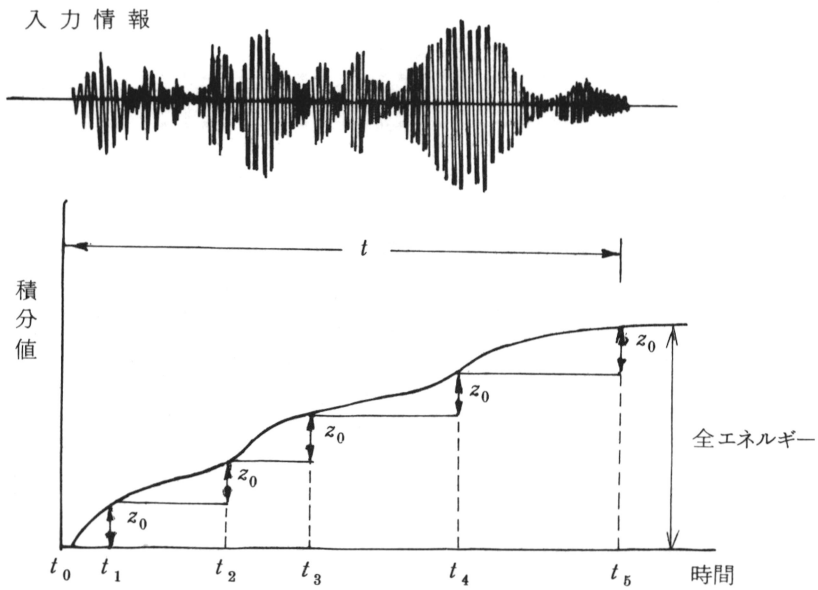
帯域フィルタの各チャネルの出力値は音声蓄積装置（AMI）で積分され，マルチプレクサで走査しA／D変換装置（MADC）でディジタル量に変換され計算機に読み込まれる．

$$\begin{vmatrix} E_{01}, E_{02}, E_{03}, \cdots\cdots\cdots\cdots E_{0n} \\ E_{11}, E_{12}, \cdots\cdots\cdots\cdots\cdots \\ \vdots \\ E_{m1}, \cdots\cdots\cdots\cdots\cdots\cdots E_{mn} \end{vmatrix} \qquad (1)$$

このデータは計算機に一時記憶され，つぎの計算を行なう．まずF_0経由のデータを全時間帯について積分を行ない規定の分割数kで分割する．

$$Z_0 = \frac{1}{k} \cdot \sum_{j=1}^{n} E_{0j} \qquad (2)$$

再びF_0経由のデータの積分を行ない，この値がZ_0の整数倍となる時刻点を求めてゆく．これは$k-1$点あり時間軸はk分割されたことになる．この様子を図－7に示す．



図－7　エネルギー分割（$k=5$の場合）

他のチャネル$F_1, F_2, \cdots F_m$の出力値については上記で求めた分割時刻でそれぞれn分割し分割時間帯ごとに積分を行なう．このようにして(1)式は時間軸方向（横方向）がk個に圧縮され次の形となる．

$$\begin{vmatrix} Z_{01}, Z_{02}, \cdots\cdots Z_{0k} \\ Z_{11}, \cdots\cdots \\ \vdots \\ Z_{m1}, \cdots\cdots Z_{mk} \end{vmatrix} \qquad (3)$$

ここに，

$$Z_{01} = Z_{02} = \cdots\cdots = Z_{0k} = Z_0$$

である．

しかし，このままでは発声の際のレベルの大小がそのまま数値全体の値に影響するので，上記の Z_0 を用いて正規化を行ない（4）式の形にしてパラメータ化を完了する．

$$\begin{vmatrix} Z_{11}/Z_0, Z_{12}/Z_0, \cdots\cdots Z_{1k}/Z_0 \\ Z_{21}/Z_0, \cdots\cdots \\ \vdots \\ Z_{m1}/Z_0 \cdots\cdots Z_{mk}/Z_0 \end{vmatrix} \qquad (4)$$

このようにして検出されたパラメータを図－8に示す．なお図－7における次の情報もパラメータの一つとして使用できる．

$$\left| t_1 - t_0/t, \quad t_2 - t_1/t, \quad \cdots\cdots \quad t_k - t_{k-1}/t \right| \qquad (5)$$

記憶装置にはこのパラメータを音声内容と対応づけた辞書として記憶させておく．

8．パラメータの照合

記憶装置にはあらかじめ言葉に対応した基準パラメータを辞書の形で登録しておく．このとき前記2.で述べた識別語の制限や話者の制限を適用し，識別内容や話者単位の辞書を多数用意しておけばよい．例えば対話する話者番号の指定によりその利用者個有の辞書の中から識別を行なうことにより識別精度を高くとることができる．

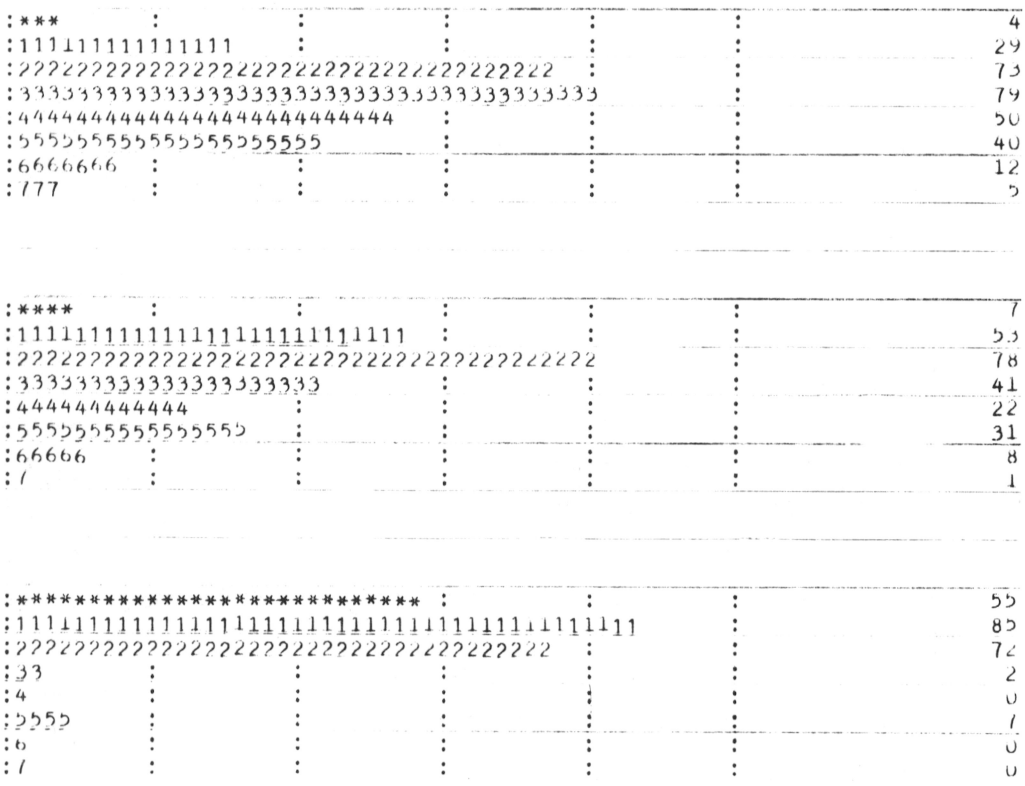
パラメータ照合の方法には，相関による方法，パラメータのベクトル空間の距離を求める方法，対応座標ごとの差の絶対値を求めその総和による方法などが考えられるが今回の方式においては，自乗偏差法，平均偏差法について検討した．

すなわち，基準パラメータを

$$P_{ij} = \begin{vmatrix} P_{11}, P_{12}, \cdots\cdots P_{1k} \\ P_{21}, \cdots\cdots \\ \vdots \\ P_{m1}, \cdots\cdots P_{mk} \end{vmatrix} \qquad (6)$$

とし，識別パラメータを

```
          0.2       0.4       0.6       0.8       1.0

:*******  :         :         :         :         :          13
:111111111111111111111111111111111111111:         :          77
:2222222222222222222222222222222222222222         :          79
:333333333333333333 :         :         :         :          34
:44444444 :         :         :         :         :          14
:555555   :         :         :         :         :          11
:666      :         :         :         :         :           5
:7        :         :         :         :         :           0


:***      :         :         :         :         :           5
:11111111111111111111111111111111111111111111     :          86
:2222222222222222222222222222222222222  :         :          72
:3333333  :         :         :         :         :          12
:44       :         :         :         :         :           2
:5555555555555      :         :         :         :          24
:66666666666        :         :         :         :          21
:77777    :         :         :         :         :           9


:**       :         :         :         :         :           3
:111111111111111111111        :         :         :          41
:22222222222222222222222222222222222222222        :          81
:3333333333333333333333333333333        :         :          61
:44444444444        :         :         :         :          21
:555555555555555555 :         :         :         :          35
:66666666666        :         :         :         :          20
:7777777  :         :         :         :         :          13


:**       :         :         :         :         :           3
:1111111111111      :         :         :         :          25
:22222222222222222222222222222222222222 :         :          74
:33333333333333333333333333333333333333333        :          81
:444444444444444444444444444  :         :         :          52
:555555555555555555 :         :         :         :          35
:666666   :         :         :         :         :          11
:77777    :         :         :         :         :           9


:****     :         :         :         :         :           6
:111111111111111111111111     :         :         :          47
:2222222222222222222222222222222222222  :         :          73
:33333333333333333333333333333:         :         :          57
:44444444444444444444         :         :         :          39
:555555555555       :         :         :         :          23
:666      :         :         :         :         :           5
:77       :         :         :         :         :           3
```



図－8　パラメータ・パターン

$$Q_{ij}=\begin{vmatrix} Q_{11}, Q_{12}, \cdots\cdots\cdots\cdots Q_{1k} \\ Q_{21}, \cdots\cdots\cdots\cdots\cdots \vdots \\ \vdots \\ Q_{m1}, \cdots\cdots\cdots\cdots\cdots Q_{mk} \end{vmatrix} \qquad (7)$$

とすると，その識別偏差は，

$$\eta=\Sigma\,(P_{ij}-Q_{ij})^2 \qquad (8)$$

あるいは，

$$\eta=\Sigma\,|P_{ij}-Q_{ij}| \qquad (9)$$

で求められる．

登録されている基準パラメータ群のすべてについてこの偏差値を計算し，その値が最小値をとりかつ一定のいき値（threshold）に入っているものを識別パターンとする．もし偏差値がこのいき値を越えている場合には，判定を保留し識別ができなかったものとして再発声をうながす．これはノイズ入力があった場合や発声者が登録語にない言葉を発声したときなどに備えるためであり，再発声に際しては発声者も心理的に慎重になることが期待できるので，いき値を多少ゆるめる処置をとることも考えてよい．

さらに識別がうまくいったとき，そのときのパラメータで基準パラメータを修正，補正することも可能であり，一種の学習機能といえることができる．この補正の方法の一つとしてつぎの実験を行なった．

$$P'_{ij}=\frac{(P_{ij}+Q_{ij})}{2} \qquad (10)$$

9．セグメンテーション

今回の方式の特長として，音声情報の持つエネルギーを均等に分割数 k により配分したことがあげられる．

音声情報を時間軸に分割する方法として理想的な手段は，音素や音韻を取り出して処理することである．このことをセグメンテーションという．このとき発声が音韻ごとに区切ってされればよいが，単語や文章として発声された情報の中からこの要素を抽出することは容易でない．これは連続した情報の中で一つの要素から次の要素に変化するとき，その過度部（わたりといわれる）が影響を与え，また同一要素であっても接続している前後の要素からの影響を受けていることがセグメンテーションを困難にしているものと考えられる．

したがって，今回の方式では分割数 k という定数によってセグメンテーションの効果を代行したが，この値は現在のところ言葉のモーラ数（音韻の数と考えてよい）を目安として決定している．例えば識別対象の言葉の平均モーラ数の 1.5 倍から2倍程度に選んだとき 良い結果が得られている．

他の何らかの方法でこのモーラ数が自動検出できるならばその値を分割数として利用することにより識別語数を多くすることが期待できる．また式(5)の時間軸要素のパラメータを使用して一次識別を行ないその中から二次識別を行なうなど段階的に処理を行なうことも有効である．

このように，音声識別や波形識別の方式は多くの手段や方法を多面的に組み合わせて行なうことがより効果的といえる．

10. 本方式の識別能力

この方式の識別確度を判定するため多くの実験を行なった．

例えば，百人一首の上五文字を用いた発声実験では，まぎらわしい音韻構成をもつ「あまのはら」と「わたのはら」，「みかのはら」などを除いて，95～99％の識別確度を得た．これは識別対象となる単語を，識別し易さという立場で選択することによって100個程度の単語識別が可能であることを示唆している．また発声の速度をある程度早くしたり遅くしたりした場合にも十分な識別確度を得ることができ，エネルギー分割のセグメンテーションが効果をあげていることが立証できた．

また電報用語（アサヒノアなど），数字語，会社名，人名，県名などについても実験を行なったがほぼ良い結果を得ている．幸いに人間は一つの言葉をいくつもの表現で話すことが可能であるので，例えば「トクシマ」と「フクシマ」が誤りやすいのであれば「阿波トクシマ」と発声を変えることにより救うことができる．

このように現在の技術においては人間がある程度，計算機が識別しやすいような発声の方法を学習して対話してゆく必要があると思われる．

今回の方式は音声情報のエネルギー分布に主眼をおいているので，エネルギーの小さい子音部が母音部に吸収されてしまい，いわば母音列による識別といえる．これを改めるため非直線増巾器を用いて子音部のエネルギーを増巾して処理することを考慮中である．

11. おわりに

以上，音声識別の一方式について報告したが，まだその緒についたばかりの実験であり十分なものではない．しかし，この実験で気付いたことは認識という技術を計算機で処理するには，あらゆる手段や方式を組み合わせることが必要であるということである．

そして，この技術を完成させるためには物理，医学，電子工学，情報工学など広い分野での研究と協力が必要であろう．

おわりに本方式の実験に当たって多大の御指導と御助言をいただいた富士通(株)第1交換技術部次長 枝川洋氏に深謝します．

参考文献

石井威望，堀川甫，阿部行信，浦野由多加：「Register方式による単語識別の一方法」
日本音響学会　2-2-21，1969年10月

枝川洋，伊藤時生，堀川甫，阿部行信：「音声情報処理へのアプローチ」
FUJITSU　Vol.21　No.3　1970

枝川洋：「音声識別」　データ通信　1970年6月

阿部行信：「コンピュータによる音声識別」　bit　1970年8月

松山辰郎，阿部行信：「百人一首とコンピュータ」　bit　1970年1月