マイクロコンピュータを用いた音声認識装置

マイクロコンピュータを用いた音声認識装置

A Speech Recognition Device with Microcomputer

高橋文彦* ・ 黒田芳朗* ・ 福岡克仁**
Fumihiko TAKAHASHI* ・ Yoshiro KURODA* ・ Katsuhito FUKUOKA**

(昭和57年4月15日受理)

あらまし

本研究は，マイクロコンピュータによる音声認識装置の試作と，その実験結果に関するものである。マイクロコンピュータには市販のＳＯＲＤ　Ｍ223Ｖ（2.5ＭＨｚ，64ＫＢ）を使用，特徴パラメータにはＲＡＲＣＯＲ係数を用い算出はソフトウエアで行っている。パラメータの個数を種々変更して認識実験を行った結果，単一話者の母音認識ではＰＡＲＣＯＲ係数は6個（1～6次）あれば，ほぼ100％の認識率（所要時間　約6秒）が得られ，標準パターンの標本数を多く（1母音当り10標本程度）とりさえすれば，2個（1～2次）でもほぼ100％の認識率（所要時間　約2秒）が得られることが判明した。

1．まえがき

人間と機械のコミュニケーションシステムが，最近大きくとりあげられるようになってきたが，コミュニケーションの手段として一番便利なものは何といっても音声である。他の装置のように操作に熟練する必要もないし，同時に手足，目などを使って別の仕事をすることもできる。情報入力速度は，キーボードの約3倍と言われている。このように考えてみると，音声合成は機械から人間への働きかけであり，一方音声認識は，逆に人間から機械への働きかけと言うことができる。音声認識の研究は，合成の研究よりも古くから行なわれており，より高い認識率をより速く達成するために，研究が続けられている。

われわれは大型計算機を利用して音声認識の基礎的研究を続けてきたが，今回その実用化をめざしてマイコンによる音声認識装置を試作し実験を行ったので報告する。

2．装置の概要

装置の構成は，マイクロコンピュータＳＯＲＤ Ｍ223Ｖを主体にし，これに音声入力用のマイクアンプ，Ａ／Ｄコンバータを接続したものである。補助記憶としてフロッピーディスク装置2台をもち，出力用のプリンターも接続されている。マイクアンプには，サンプリングのためのクロックパルス発生器や信号取込みのためのスタートパルス発生器が組み込まれており，発声レベルのめやすのためにレベルメータを取付けている。A/D変換部は12 bit で，10KHzでサンプリングされマイクロコンピュータにとり込まれる。CPUのクロックは2.5MHzで，記憶

* 電気工学科 (Department of Electrical Engineering)
** 東京大学工学部電子工学科 (Department of Electronic Engineering, Tokyo University)

容量はＲＡＭエリアで64ＫＢである。フロッピーディスク装置は，Y–E Data 174D型で両面倍密度・1MB，プリンターはIBMのゴルフボール型である。これらの外観写真を**図1**に示す。**図2**はシステムのブロック図を示したもので，**表1**は各部の性能をまとめたものである。

図1　音声認識装置の外観

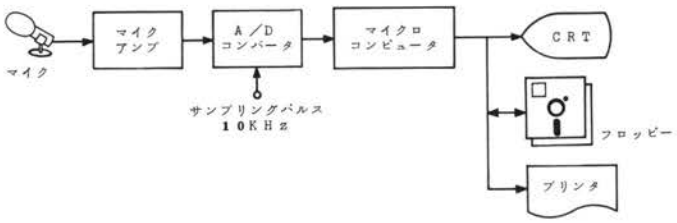


図2　システム構成

マイクロコンピュータ SORD M223 Mark V	
CPU	Z–80 (2.5MHz)
RAM	64kバイト
フロッピーディスク装置	両面倍密度 1Mバイト×2台
入出力装置	12インチグリーンモニタ (キャラクターディスプレイ) プリンター，キーボード
アナログ入出力ユニット	SORD HC–AIO
A D 変換部	12ビット
使用モード	外部トリガモード (10KHz)
D A 変換部	8ビット×2
マイクアンプ	
ゲイン	80dB (MAX)
フィルター	5KHz ローパスフィルター
付属回路	クロックパルスジェネレータ スタートパルスジェネレータ ヘッドフォーンアンプ レベルメーター　　等

表1　構成機器の性能

3．認識の方法

音声認識の方法は，まず標準音声パターンを登録しておき，テスト音声の入力があると，先の標準パターンとテスト音声パターンを比較して最も類似した声に判定する方法である。音声の特徴パラメータには PARCOR 係数を用い，判定は統計的手法によるベイズの判別得点方式によって行っている。

マイクからの音声入力は，マイクアンプ内で 20dB／dec のプリエンファシスを行った上，5 KHz のローパスフィルターを通してA／Dコンバータに入り，ここで 10 KHz でサンプリングされて計算機に入力される。計算機内では予めプログラムによって指定された個数だけデータをとり込み，次にこれらデータ列の切り出しむらを平滑化するために，時間窓としてハミング窓（hamming window）をかける。ハミング窓は次式で表わされる。

$$W_h=0.54+0.46\cos\left\{2\pi\left(i-\frac{n+1}{2}\right)\frac{1}{n-1}\right\} \qquad (3.1)$$

$(i=1.2.\cdots\cdots n)$　iはデータ番号，nは切出個数

この処理を施したあとで，自己相関係数を必要な次数だけ求め，これから PARCOR 係数を同じ次数だけ算出する。このアルゴリズムは板倉の方法による。このようにして各音韻毎に標準パターンとすべき音声を何回かづつ入力して，これらから求めた PARCOR 係数群を分散共分散行列の形に再整理して音韻毎に登録しておく。同時に PARCOR 係数の平均値も各次数毎に算出して登録しておく。

このような準備が出来上った段階でテスト音声パターンが入力されると先の手順でPARCOR係数群がまず求まる。これをベクトル的にXで表現すると，この音声が第 i 母集団(音韻 i)に属すると仮定した場合の判別得点は，

$$S_i=-\frac{1}{2}\log_e\left|\sum_i\right|+(X-\mu^i)'\sum_i^{-1}(X-\mu^i)+\log_e\pi_i \qquad (3.2)$$

ここで $\sum_i$ は第 i 母集団の PARCOR 係数の分散共分散行列であり，μ^i は同じく，平均値をベクトル的に表わしたもので，π_i は音韻 i の先験確率である。実際の計算では先験確率は同じと仮定して，さらに共通因数−1/2を取り去って，

$$S_i'=\log_e\left|\sum_i\right|+(X-\mu_i)'\sum_i^{-1}(X-\mu_i) \qquad (3.3)$$

を最小にする i をもつ母集団（音韻）に属すると判定する。

4．音声認識プログラム

4．1　プログラムの構成

この実験では，音声認識を 3 段階に分けて行っている。まず第 1 の段階は，標準音声パターンを収録することである。ここでの作業内容は，マイクから入力した音声信号をA／D 変換して，フロッピーディスク上にラベルをつけて書き込むだけで，システムはレコーダとして動作するだけである。

第 2 の段階はデータ解析である。先に記録しておいた音声データを用いて，PARCOR 係数を求め，これからさらに各音韻グループ毎に PARCOR 係数の平均値 μ_i と分散共分散行列 $\sum_i$ を求める。解析した結果はフロッピーディスク上にラベルをつけて登録される。

第3の段階ではテスト音声を入力しPARCOR係数を求めてから，先の解析結果と照合して，判別得点法により判定を行う。

便宜上，以上の3段階の処理過程をそれぞれ，

PASS 1　〈データ収録〉
PASS 2　〈データ解析〉
PASS 3　〈音声認識〉

と名づけ，以下に詳述する。なお，全体の処理の流れを**図3**に，それぞれの段階の内部の流れを**図4**に示した。

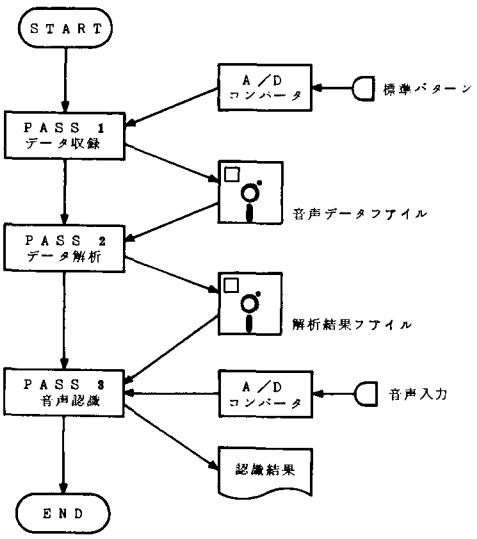
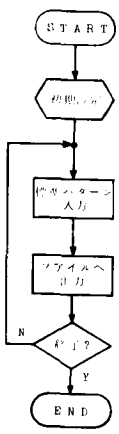
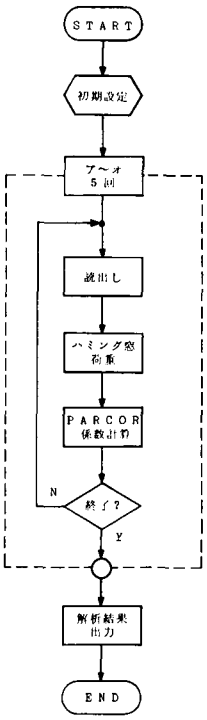


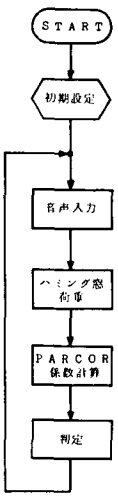
図3　音声認識のゼネラルフロー



(a) PASS 1
〈データ収録〉



(b) PASS 2
〈データ解析〉



(c) PASS 3
〈音声認識〉

図4　各段階での詳細フロー

4．2　PASS1〈データ収録〉

ここでは，認識作業の基準になる標準音声パターンを得るための標準音声データを収録する。プログラムを起動して，作成すべきファイル名，データのフレーム長さ，データの個数を入力すると，音声データの入力待ち状態になり，ここでマイクに向けて発声する。音声のレベルが飽和レベルの約75%を越えたところで自動的にトリガされ取り込みが開始される。この動作の完了を表示するために，入力したデータが，**図5**のようにＣＲＴ上に表示される。同時に，フロッピーディスク上にも，ラベル，フレーム長さを付記して，このデータが記録される。

以下同様に，例えば，／ア／，／イ／，／ウ／，／エ／，／オ／，／ア／，……と発声していけば，音声データが順次フロッピーディスク上に記録され，所定個数を入力したところで，ＯＳにもどり，作業は完了する。

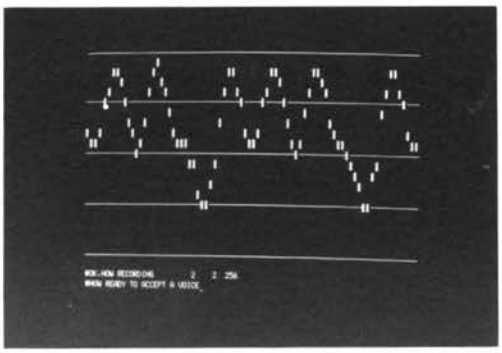


図5　ＰＡＳＳ１実行中のディスプレイ

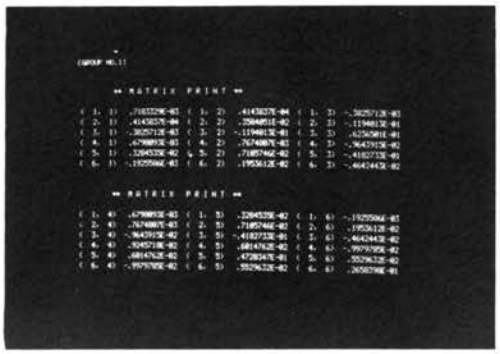
ＰＡＳＳ１は標準音声データの取り込みをできるだけ手早く終えられるように配慮したのでデータ処理などは，ＰＡＳＳ２に回した。従って50個の音声データの取り込みに必要な時間は５分～10分程度で十分で被験者の負担は軽い。

ＰＡＳＳ１を終了すると，ドライブ＃１のディスク上に“V0C”という払張子をもつ音声データファイルが完成する。このデータは整数型２バイトのデータ列で書き込まれているので，フレーム長256とすると，１データにつき512バイトとなり，フロッピーディスクは約１Ｍバイト／メディアなので，約2000個のデータが登録できるわけである。

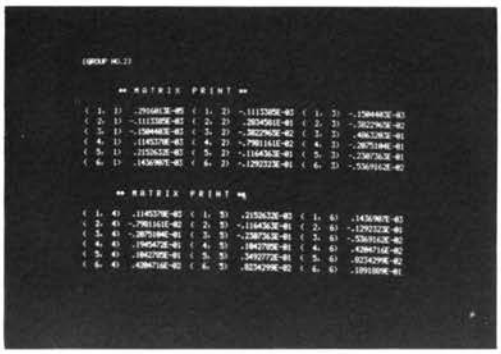
4．3　PASS2〈データ解析〉

ＰＡＳＳ２を起動し，解析すべき音声データファイル名，各音韻のデータ数，出力ファイル名を指定すると，以下の解析動作は終了まで一気に進行する。ＰＡＳＳ２の実行が完了するとドライブ＃１のフロッピーディスク上に“PAR”という拡張子をもつ解析結果のファイルが得られる。このファイルに入っているデータは，音声データの各音韻毎のPARCOR係数の平均と分散共分散である。始めのデータと比べると，非常にコンパクトなデータになっている。

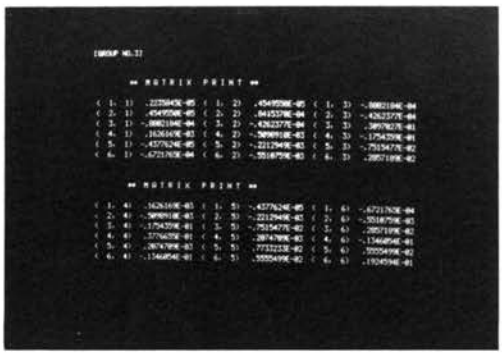
図6にＰＡＳＳ２実行中のディスプレイを示す。(a)～(c)は，それぞれ／ア／，／イ／，／ウ／，に対するPARCOR係数の分散共分散行列，(d)は／ア／～／オ／に対するPARCOR係数の平均値である。（各行が，／ア／～／オ／を，各列がPARCOR係数の次数を示す）



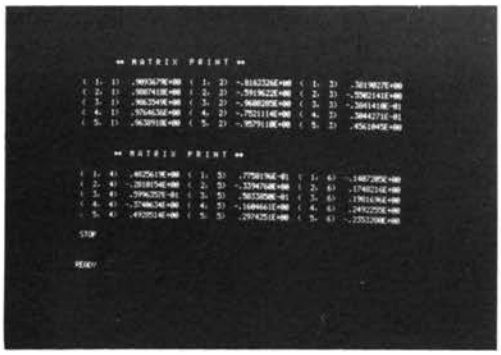
(a) ／ア／分散行列



(b) ／イ／の分散行列



(c) ／ウ／の分散行列

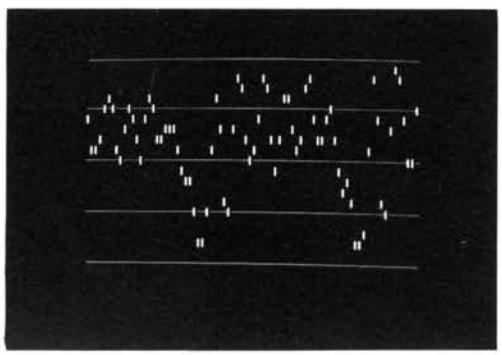


(d) PARCOR係数の平均

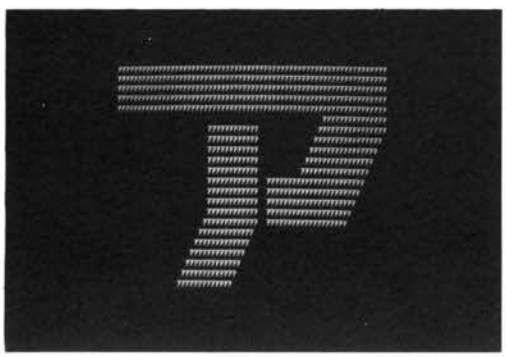
図6 ＰＡＳＳ２実行中のディスプレイ

4．4 ＰＡＳＳ３〈音声認識〉

ＰＡＳＳ３では，使用する解析結果のファイル名と，使用するPARCOR係数の次数を入力すると，テスト音声の入力待ちになる。音声が入力されると，その都度認識動作が行なわれ，ＣＲＴ上に認識結果が表示される。**図７・１，図７・２**にＰＡＳＳ３実行中のディスプレイを示す。

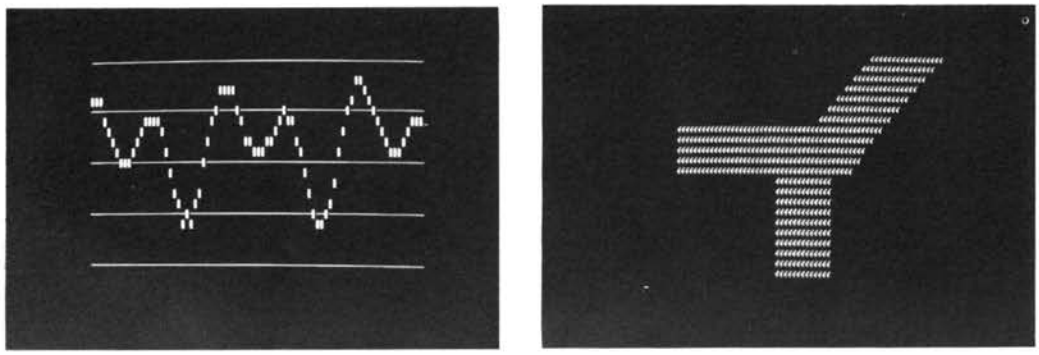


(a) 波 形



(b) 表 示

図７・１ ＰＡＳＳ３実行中のディスプレイ・／ア／



(a) 波　形　　(b) 表　示

図7・2　ＰＡＳＳ３実行中のディスプレイ・／イ／

5．評価実験

5．1　実験方法

試作した音声認識装置の性能を調べるために，次のような音声データファイルを作成して評価実験を行なった。

データ１（標準パターン）・単一話者
　ア，イ，ウ，エ，オ　各20音（全 100 音）
　フレーム長　256
データ２（試験パターン）・単一話者
　ア，イ，ウ，エ，オ　各10音（全50音）
　フレーム長　256

データ１で標準パターンを作成し，この結果を用いてデータ２を認識する実験を行ったものである。内容は以下の通りである。

(1) データ１（標準パターン）の個数を制限して，分散行列作成時のデータ個数が，認識率に与える影響を調べる。
(2) PARCOR 係数の個数を変化させて，これが認識率と所要時間に与える影響を調べる。
(3) データのフレーム長を制限して，フレームの長短が認識率と所要時間に与える影響を調べる。

5．2　実験結果

実験(１)および(２)の結果をまとめて，**図8**に示す。実験(３)については，フレーム長 128 以上になると，４次・２次とも 100 %の認識率を示すが，２次の場合フレーム長 100 で低下し始め，4次の場合はフレーム長75で低下し始める。所要時間は，ほぼフレーム長と，PARCOR の次数の積に比例して増加する傾向を示した。

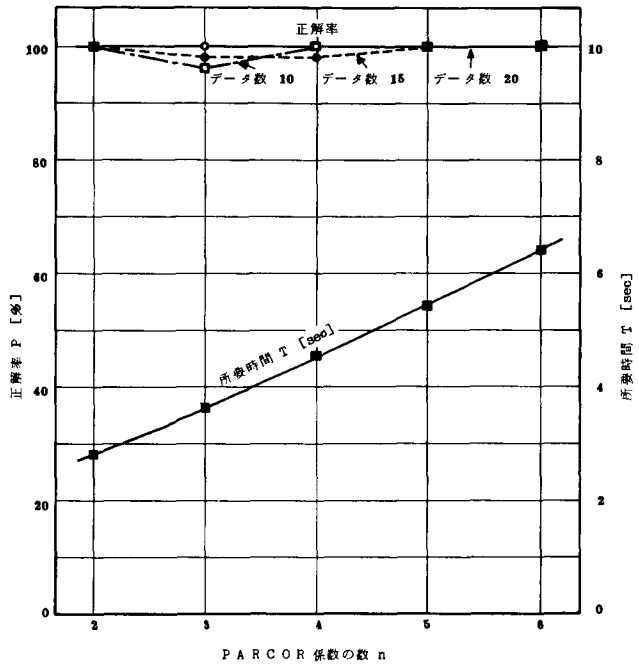


図8 標準用データ数・PARCOR次数と性能の関係

6. むすび

マイクロコンピュータによる音声認識装置を試作して良好な結果を得た。単一話者で母音を対象にした場合，標準用データ数を10個程度とれば，PARCOR 係数は4次くらいでも十分実用になることがわかった。また，このようにデータ数を10個程度とると非常に安定した標準パターンが作成でき，事実，数ケ月前に収録した音声を基準にしても，認識率は，ほぼ100％で変動は見られなかった。今後は PARCOR 係数の算出部をハード化して更に実験を進めるつもりである。

謝辞　日頃，適切なご助言と貴重な資料をご提供下さっている京都大学工学部の坂井利之教授，ならびに，筆者らが米国滞在中，多くの資料と有益なご助言を頂いた，南カリフォルニア大 SCRL (Speech Communication Research Laboratory) の所長 June E. Shoup 博士，脇田博士らに心からの謝意を表する。

文　献

(1) 宮川他　ディジタル信号処理　電子通信学会
(2) 中田　音声　日本音響学会
(3) ラオ　統計的推測とその応用　東京図書