零交差情報にとる音声認識

白石, 幹雄 / SHIRAISHI, Mikio / 渡辺, 嘉二郎 / 田中, 具治 / TANAKA, Tomoharu / WATANABE, Kajiro

(出版者 / Publisher)
法政大学工学部
(雑誌名 / Journal or Publication Title)
Bulletin of the Faculty of Engineering, Hosei University / 法政大学工学部研究集報
(巻 / Volume)
21
(開始ページ / Start Page)
95
(終了ページ / End Page)
106
(発行年 / Year)
1985-03
(URL)
https://doi.org/10.15002/00004063

零交差情報による音声認識

白石幹雄*・渡辺嘉二郎**・田中具治***

Automatic Speech Recognition by Zero-Crossing Information

Mikio SHIRAISHI*, Kajiro WATANABE** and Tomoharu TANAKA***

Abstract

A system for speaker-dependent isolated word recognition is described. The system consists of a single-board preprocessing hardware and a microcomputer system. In this hardware, speech signal is divided into two frequency bands, and zero crossings of speech signal in the respective frequency bands are counted. Recognition performance of the system was investigated. The recognion rate for a set of Japanese numeral word (0 to 9) which was uttered by 5 male speakers was 96.6 %. DP matching implementation was used in this experiment.

1 緒　　言

零交差分析は，音声のホルマント周波数を抽出する方法の一つで，簡単なハードウエアで行うことができる。また，コンピュータとのインターフェースに，A/Dコンバータを必要としない。音声認識装置に使われている音響分析の方法には，ほかに，帯域通過フィルタ群による方法や，線形予測分析，ウォルシュ・アダマール変換などがある。これらは，いずれもA/Dコンバータを必要とする。A/Dコンバータは，部品としては高価で，回路も複雑であり，音声認識装置をLSI化するときには，別のチップにして外づけされる例を見かける。

音声認識装置の音響分析部分に，A/Dコンバータのいらない零交差分析を用いることによって，

a) 装置の価格を低減でき，

b) 認識処理を行うマイクロプセッサ部分を含めて，装置全体を1チップLSI化するのに役立つ

と考え，零交差情報による音声認識を試みた。零交差情報によって数字音声の認識を行った例は，過去にいくつか報告されている[1], [2]。

* 法政大学工学部大学院工学研究科電気工学専攻（M2）
** 法政大学工学部電気工学科計測制御専攻
*** 法政大学工学部電気工学科電気電子専攻

2 零交差分析

音声の零交差波を作ると，エネルギーの強い周波数成分が強調される[3]。零交差波は，高増幅率の増幅器で入力信号をクリップして得られる，振幅一定の矩形波である。音声信号を$f(t)$とすると，その零交差波$z(t)$は

$$z(t)=\begin{cases} a, & f(t)\geqq 0 \\ -a, & f(t)<0 \end{cases} \tag{1}$$

のように表すことができる[4]。ここで，aは正の実定数である。Fig. 1に，母音 /e/ の波形とその零交差波とを示す。

音声母音には，その母音を特徴づける，エネルギーの強い周波数成分が，いくつか存在する。

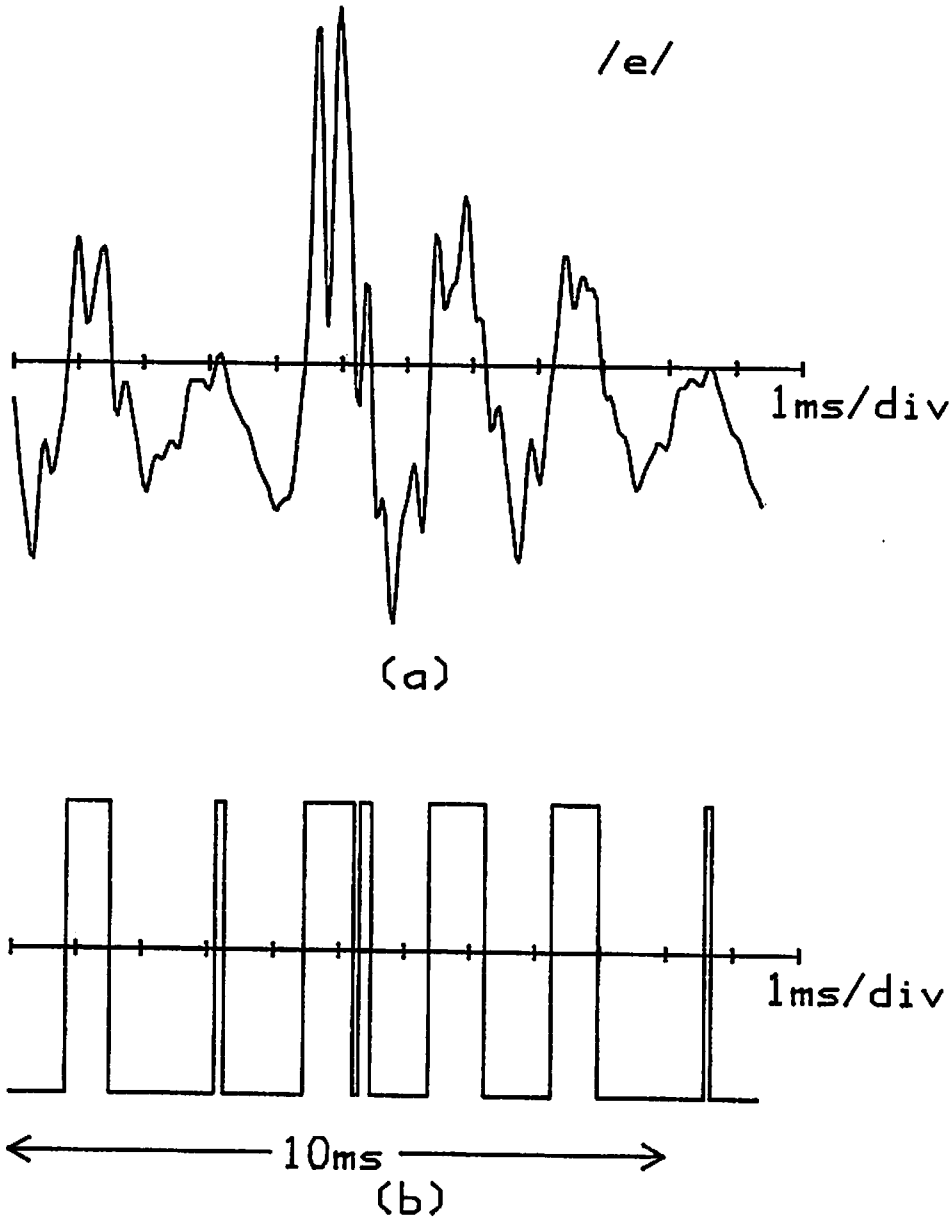

Fig. 1 (a) Wave form of /e/ uttered by a male speaker and (b) its zero-crossing wave.

これらをホルマントと言う。母音のホルマントは，周波数が低い方から，第1ホルマント，第2ホルマント，……と呼ばれる。これらの周波数は，それぞれ記号F_1，F_2，……によって表される。Fig. 2に母音ホルマントの例を示す。Fig. 2の曲線は，Fig. 1に示した母音/e/の波形をFFT（高速フーリエ変換）して得たスペクトルである。F_1，F_2，F_3に対応する周波数成分を太線で示した。FFTの結果から，$F_1=5.3\times10^2$Hz，$F_2=2.15$kHz，$F_3=2.77$kHz（名目上の周波数分解能は44Hz）である。

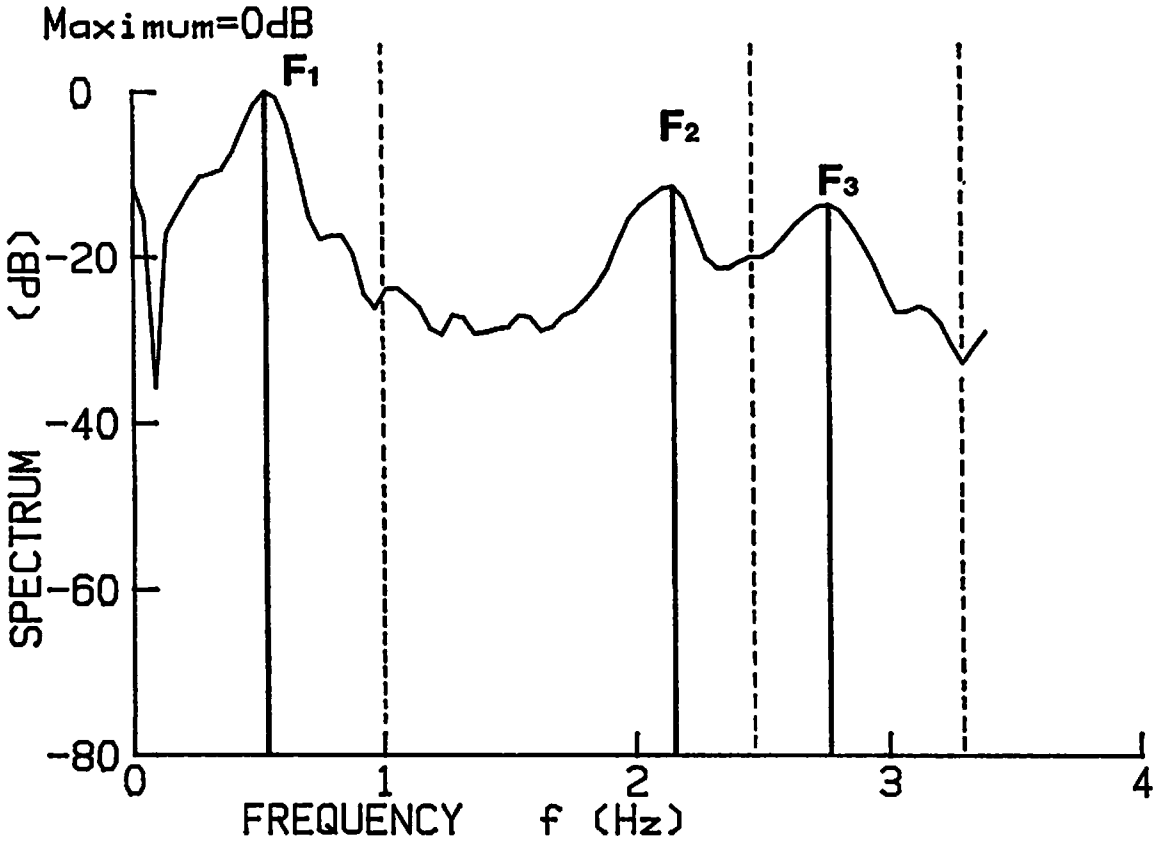

Fig. 2　Formants of /e/.

音声の零交差波が単位時間あたりに立ち上がる回数は，エネルギー最大のホルマントの周波数に近い。Fig. 1に示した10msの区間で，母音/e/の零交差波は6回，立ち上がっている。したがって，/e/の零交差波の単位時間あたりの立ち上がり回数は$6/(10\times10^{-3}\text{s})=6\times10^2\text{s}^{-1}$である。Fig. 2に示した/e/のホルマントの中で，エネルギーが最大のホルマントは，第1ホルマント（周波数$F_1=5.3\times10^2$Hz）である。零交差波の立ち上がり回数は，カウンタ回路によって，容易に計数できる。

Fig. 2の破線のように，音声信号をフィルタリングによって幾つかの周波数帯域に分けて，それぞれの帯域で信号の零交差波が立ち上がる回数を測定すれば，それぞれの帯域に存在するホルマントの周波数を知ることができる。このような方法でホルマント周波数を測定することをここでは，零交差分析と呼ぶことにする。

3 音声認識装置の構成

零交差分析装置を1枚のボード上に組み，これと8-bitマイクロプロセッサ（Z80）をCPUとするパーソナル・コンピュータ（NEC, PC8001）とを組み合わせて，音声認識装置を構成した。

音声認識装置には，実現のしやすさを考慮して，次の三つの制約を設けた。

1) 使用前に，話者が自分の声を装置に登録する。

2) 認識させたい単語は，一つ一つ区切って発声する（単語と単語の間に，無音の休止区間を入れる）。

3) 認識語彙数を十数語に限定する。

用途としては，次のようなものを想定している。

1) 音声による電話のダイヤリング

2) コンピュータのデータ入力端末

3) 患者の声による病室の制御（カーテンや窓の開閉，照明の点灯・消灯など）

4) おもちゃ

3.1 零交差分析装置

Fig. 3に，製作した零交差分析装置のブロック・ダイヤグラムを示す。Fig. 3の破線で囲んだ部分が，1枚の基板上に作られている。この装置は，

a) フィルタリングによって，音声信号を高低二つの周波数帯域に分割し，

b) それぞれの帯域で，音声波形の零交差波が一定時間内に立ち上がる回数をカウントする。

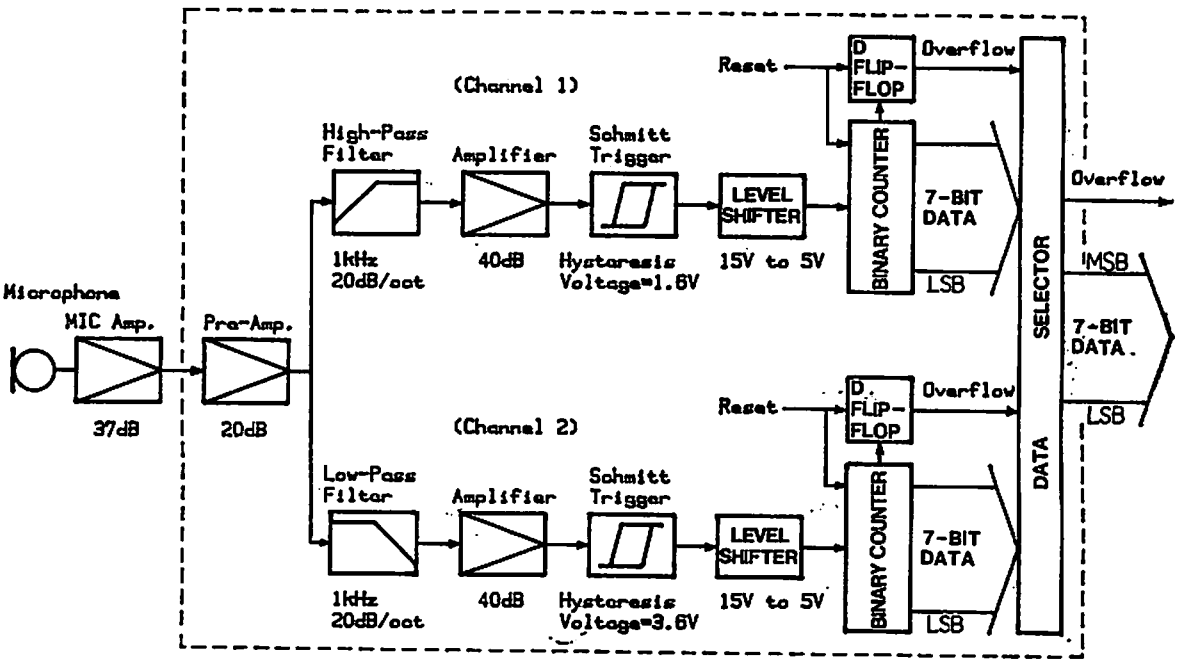

Fig. 3 Block diagram of the preprocessing hardware.

カウントしたデータは，二進7bitのディジタル信号として出力される。これらの値は，それぞれの帯域に存在する音声ホルマントのうち，最もエネルギーが大きいものの周波数に近い。以下でFig. 3について説明する。

マイクロホンによって電気信号に変換された音声信号は，マイク・アンプによって適当な電圧レベルにまで増幅される。マイクロホンには，接話型のもの（パイオニア，SE-DJ 1，周波数特性20～20000Hz，感度—85dB/μbal，双方向指向性）を用いた。マイク・アンプの電圧利得は37dB，周波数特性（帯域幅）は30Hz～10kHzである。音声信号を増幅するには，この帯域幅でも十分である。帯域幅を音声信号に合わせて制限することは，高い周波数成分を持つ雑音を減少させるのに有効である。マイク・アンプは，商用電源周波数の回わり込みを防ぐために電池駆動にし，外来雑音を減少させるために金属ケースに納めた。

マイク・アンプを出た音声信号は，プリアンプによって増幅され，高域通過フィルタと低減通過フィルタとによって，高周波数帯域と低周波数帯域とに分割される。プリアンプの利得は20.6 dB，周波数特性は66Hz～15kHzである。増幅する周波数帯域の下限を66Hzにしたのは，商用電源周波数(50または60Hz)の回わり込みを防ぐためである。フィルタは共に，遮断周波数1kHz，減衰特性20dB/octave，通過帯域での利得0dBの3次チェビシェフ・フィルタである。遮断周波数を1kHzにしたのは，母音ホルマントのF_1とF_2との境界がこの付近にあるからである。プリアンプの入力端子から，それぞれのフィルタの出力端子までの周波数特性をFig. 4に示す。

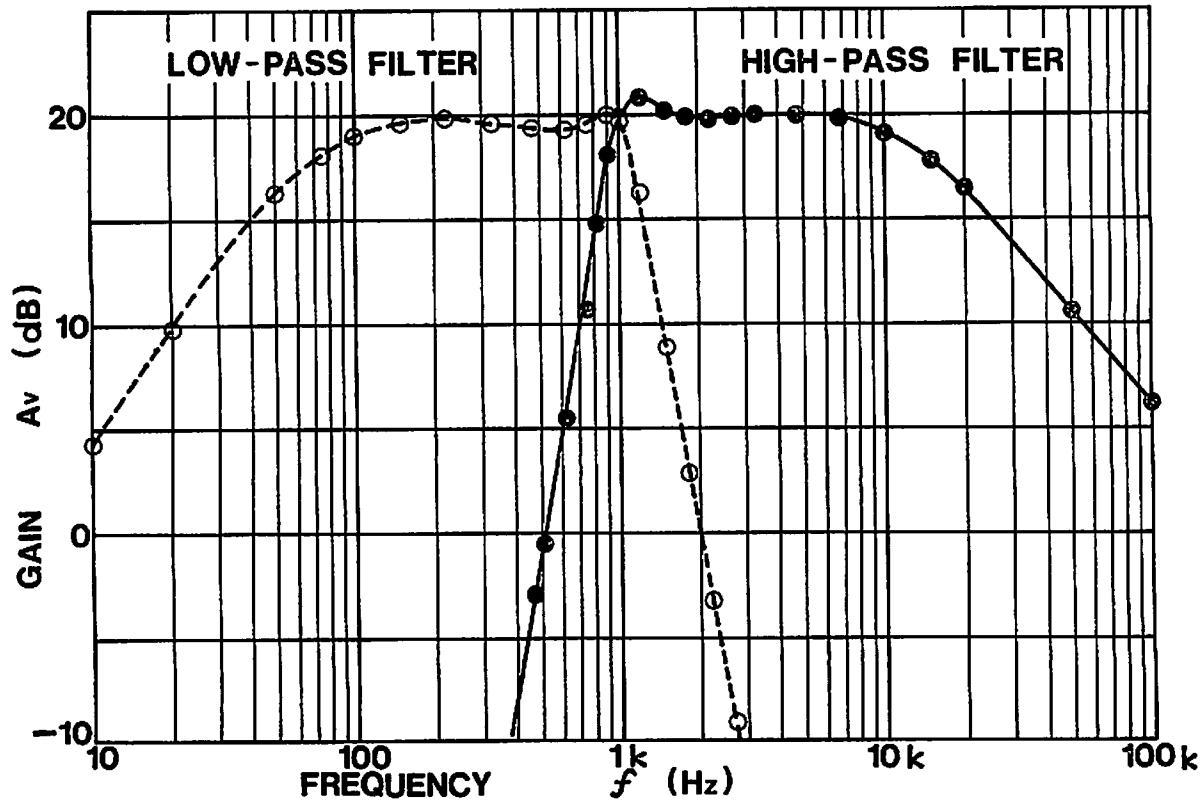

Fig. 4　Frequency characteristic of channel divider.

フィルタからの信号は，利得40dBのアンプによって増幅され，シュミット・トリガによって波形整形されて零交差波になる。このアンプでは，飽和が生じるように利得を大きくしてある。線形動作させたときの周波数特性は，設計値で15Hz～19kHzである。シュミット・トリガにはヒステリシス特性を持たせて，無入力時の雑音による誤動作を防いでいる。ヒステリシス電圧は，高域側（Channel 1）が1.6V，低域側（Channel 2）が3.6Vである。これらの値はそれぞれ，高域通過フィルタを通過した母音/u/の波形の波高値の1/10と，低域通過フィルタを通過した母音/i/の波形の波高値の1/10とを目安に設定した。それぞれのフィルタを通過した母音波形のうち，これらの値（高域では/u/，低域では/i/）が最小であることが，男性話者1名による実験の結果わかっている。

零交差波は，カウンタによってその立ち上がり回数がカウントされる。カウンタには8-bitバイナリー・カウンタを用いた。カウンタ出力の第0bitから第6bitまでの7bitsの出力をデータとして使用した。第7bitの出力は，その立ち上がりをDフリップ-フロップで記憶し，オーバーフローの検出に用いた。カウンタの出力とDフリップ-フロップの出力とは，データ・セレクタによって，Channel 1, Channel 2のうちどちらか一方の出力が選択され，パーソナル・コンピュータのI/Oポートに接続される。カウンタおよびDフリップ-フロップは，パーソナル・コンピュータからの信号でリセットされる。

3.2 認識動作

3.2.1 音声データの入力

上に説明した零交差分析装置から，一定時間ごとの零交差波の立ち上がり回数をパーソナル・コンピュータに入力する。この一定時間をゲート・タイム（gate time）と呼ぶ。入力動作はすべてソフトウエアで行う。ゲート・タイムもソフトウエアで決定する。

ゲート・タイムは10msに設定した。ゲート・タイムが短いと分析精度が悪くなる。ゲート・タイムが長いと，ホルマント周波数の変化が平均化されて測定される。ゲート・タイムを10msにしたのは，

a) 音声のピッチ周波数が100Hz以上なので，母音の1周期が10ms以内に納まること（下限），そして

b) 子音のホルマント周波数によって，カウンタがオーバーフローしないこと（上限）

からである。

Fig. 5に入力動作を行うサブルーチンのフローチャートを示す。Fig. 5のTIME DELAYのところで，ゲート・タイムが決定される。Channel 1のデータをChannel 2のデータよりも先に入力しているが，Channel 1のデータを読み込んでから，Channel 2のデータを読み始めるまでに要する時間は10.25μsである。この値はゲート・タイムの10msと比べて無視できる。したがって，両チャンネルのデータは同時に入力されると見なすことができる。

単語音声の始まりと終わりとは，自動的に検出される。単語の始まりは，4回連続して両チャンネルのデータのうち少なくとも一方が0でないデータが入力されたとき，その第1回目のデータを単語の先頭と見なすことによって検出する。単語の終了は，両チャンネルとも0のデータが入力されたときを第1回目とし，単語の先頭と見なせるようなデータが50回のあいだ入力されなかったとき，第1回目に入力されたデータの前のデータを単語の末尾と見なすことによって検出する。ゲート・タイムは10msだから，単語の継続時間長は30msより長くなければならず，一つの単語を発声した後に0.5s以上の休止区間を入れなければならない。入力できる単語全体の長さは1sに制限した。

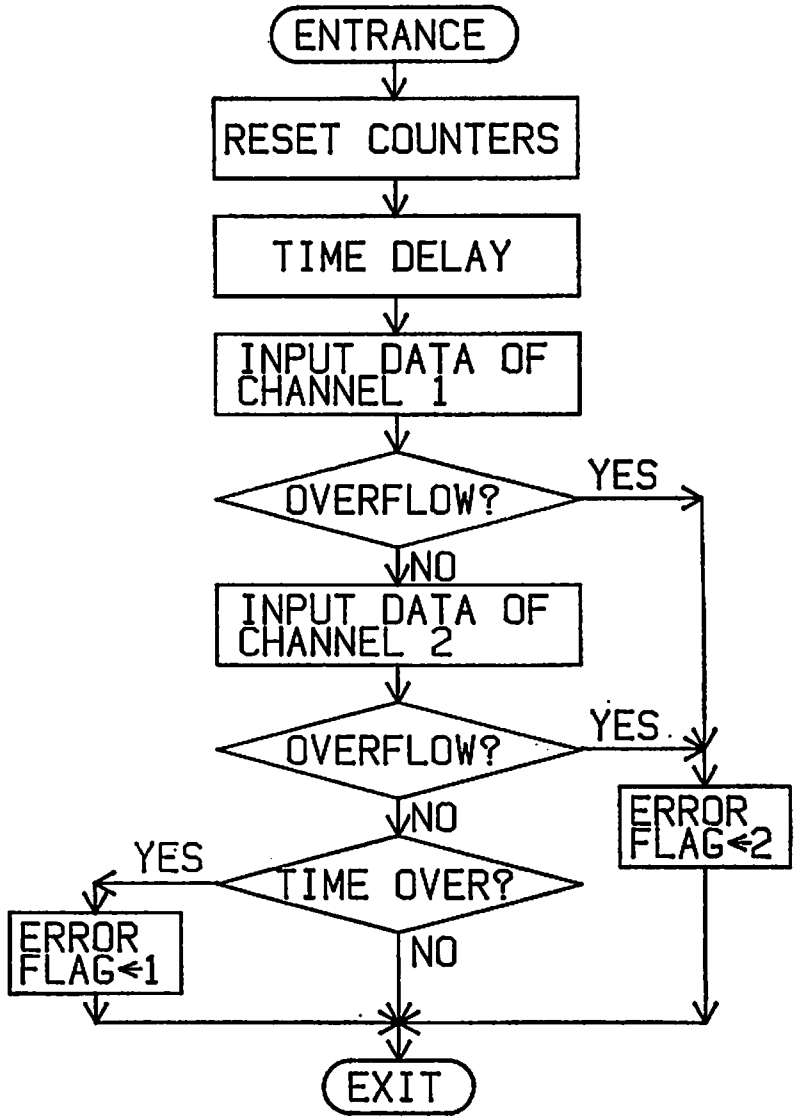

Fig. 5　Flow chart of data input subroutine.

3.2.2　単語音声の認識

認識はパターン・マッチングによって行う。ここで述べる音声認識装置には，教示モードと認識モードという二つの動作モードがある。教示モードは，認識語彙に含まれるすべての単語を話者の声で登録する動作を行うためのものである。このモードで，単語の標準パターンが作られる。認識モードは，認識動作を行うモードである。装置をこのモードにして，認識させたい単語を装置に入力すると，その単語のパターン（これを入力パターンと言う）と最も類似したパターンが，あらかじめ登録しておいた標準パターンの中から探し出され，これに対応する出力が得られる。モードの選択は，キー入力によって行う。

この音声認識装置は，次のような動作を行う。話者は，まず装置を教示モードにして，認識さ

せたい単語（またはそれに対応づけたい文字列）をキー・ボードから入力する。それからその単語を発声する。発声開始は，キー入力で装置に知らせる。音声データ入力完了は，beep 音（ベル）で装置が話者に知らせる。認識語彙に含まれるすべての単語を登録し終わったことをキー入力で，装置に知らせると，登録したデータ（標準パターンと文字列）は，ファイル・ネームとともに，フロッピー・ディスクに書き込まれる。

次に，装置を認識モードにし，登録データのファイル・ネームをキー入力すると，登録データがフロッピー・ディスクからメイン・メモリ内にロードされる。それから認識させたい単語を発声すると，パターン・マッチングによって探し出された標準パターンに対応する文字列が，CRT 画面上に表示される。

パターン・マッチングには，DP マッチング[5]を用いて，時間軸の正規化を行っている。使用している DP マッチングは，端点固定の平行四辺形 DP である。パスの傾きは 1/2～2 のあいだに制限されている。

いま，入力パターンと標準パターンの一つとをそれぞれ

$$X=x_1x_2\cdots x_i\cdots x_I, \tag{2}$$

$$Y=y_1y_2\cdots y_j\cdots y_J \tag{3}$$

で表す。x_i と y_j とは特徴ベクトルで，それぞれ

$$x_i=(x_{i1},\ x_{i2}), \tag{4}$$

$$y_j=(y_{j1},\ y_{j2}) \tag{5}$$

のように表される。ここで，x_{i1} と y_{j1} とは Channel 1 のデータ，x_{i2} と y_{j2} とは Channel 2 のデータである。

DP 漸化式には次のものを用いた。

$$g(i,j)=\min\left\{\begin{array}{l} g(i-1,\ j-2)+2d(i,\ j-1)+d(i,\ j) \\ g(i-1,\ j-1)+2d(i,\ j) \\ g(i-2,\ j-1)+2d(i-1,\ j)+d(i,\ j) \end{array}\right\}. \tag{6}$$

初期条件は

$$g(1,\ 1)=2d(1,\ 1), \tag{7}$$

パターン間距離は

$$D(X,\ Y)=g(I,\ J) \tag{8}$$

である。

式(6)，(7)の中の $d(i,\ j)$ は，ベクトル間距離である。ベクトル間距離には，

$$d(i,\ j)=\sum_{q=1}^{2}|x_{iq}-y_{jq}|\,w_q \tag{9}$$

を用いた。ここで w_q は正の荷重係数である。荷重係数 w_q を掛けるのは，次の理由による。

特徴ベクトルの要素間には，

$$x_{i2} \leqq x_{i1},\quad y_{j2} \leqq y_{j1} \tag{10}$$

の関係が，ほぼ成り立つ。それは，Channel 1 のデータは高域通過フィルタ，Channel 2 のデータはこれと遮断周波数が等しい低域通過フィルタを通過した音声波形の零交差波が，一定時間に立ち上がる回数だからである。したがって，それぞれの要素の変化率が等しいと仮定すれば，式(9)の計算で，すべの q に対して $w_q = 1$ とすると，$q = 2$ の要素のほうが，$q = 1$ の要素よりも $d(i, j)$ の値に及ぼす影響が大きくなってしまう。そこで，$w_1 < w_2$ であるような荷重係数を乗じて，両要素がベクトル間距離に与える効果の大きさの違いを補正する。Fig. 6 に母音/e/の特徴ベクトルを示す。

認識モードでは，入力パターンと各標準パターンとのパターン間距離が計算される。その中でパターン間距離の値が最小の標準パターンが探し出され，これに対応する文字列が結果として表示される。

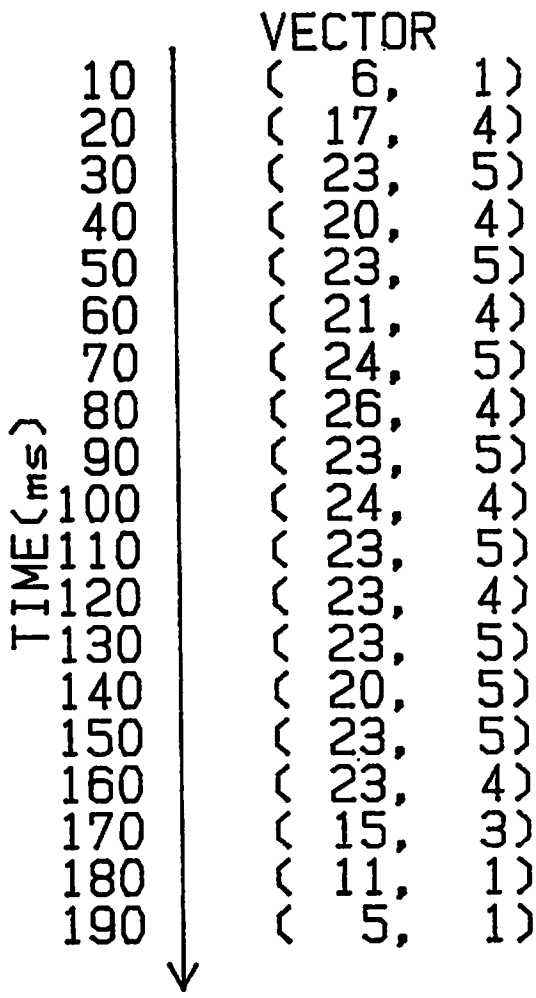

Fig. 6　Characteristic vectors of /e/.

4　認識率の測定

男性話者5名が発声した日本語10数字（ゼロ，イチ，ニ，サン，ヨン，ゴ，ロク，ナナ，ハチ，キュウ）について，認識率の測定を行った。実験は，次のようにして実行した。各話者に各数字を11回発声してもらい，これらを零交差分析して得た音声パターンをいったんフロッピー・ディスクに蓄えておいた。このうち最初に発声したパターンを標準パターン，他を入力パターンとしてマッチングした。このようにしたのは，式(9)の荷重係数 w_q が認識率に及ぼす影響を確かめるためである。荷重係数 w_1 を1（一定）とし，w_2 が1，2，4，8の各場合について，同一のパターンを用いて認識率を測定した。音声データの入力は静かな部屋の中で行った。

話者別の認識率を Table 1 に示す。認識率の平均値は，w_2 が増加するにつれて高くなっている。$w_2 = 8$ のときの平均値は，$w_2 = 1$ のときのそれに比べて1.2%高い。$w_2 = 8$ のときのコンフュージョン・マトリックスを Table 2 に示す。Table 2 の中の——は該当無しを表す。数字音声/ゼロ/と/イチ/の認識率が，他のものと比べて低い。また，/イチ/は，/ハチ/に間違えられる

Speaker	Weight coefficient w_2			
	1	2	4	8
A	95	95	97	99
B	95	96	96	95
C	91	92	91	92
D	98	98	98	98
E	98	98	99	99
Average	95.4	95.8	96.2	96.6

Table 1 Recognition rate in percent for 10 digits as a function of the weight coefficient w_2 (See equation(9)).

傾向があることがわかる。認識率が高いのは，/ニ/，/ヨン/，/ロク/，/ハチ/，/キュウ/で，100%認識されている。

パターン・マッチングに要した時間は，1単語あたり平均1sである。

5 考　　察

認識率を測定する実験で，認識語彙として数字音声を選んだのは，次の二つの理由からである。一つは，第3節で示した想定する用途の1)と2)に使用するには，数字音声を認識する必要があるからである。もう一つは，今までに報告されている，零交差情報による音声認識の研究では，数字音声の認識を行った例が多いので，過去の例との比較がしやすいと考えたからである。

文献2)によると，ある特定の男性話者1名が1000回発声した数字音声に対して99.7%，男性20名の1000数字音声に対して97.9%の識別率が得られている。ただし，/ゼロ/ではなくて/レイ/を認識させている。今回の実験で得た，男性話者5名による認識率の平均値96.6%は，これらの値よりも低い。

Table 1 を見ると，$w_2=8$ のときの各話者の認識率は，最大値が99%，最小値が92%であり，その差は7%である。また，Table 2を見ると，誤りは各単語に分散しているのではなく，特定の単語に集中していることがわかる。認識率が最も低いものは/イチ/で，その値は88%である。このことから，本装置は数字音声を認識させるのには十分でないと考える。しかし，100%認識されている単語は五つあり，少語彙の単語認識に用途を限れば，実用になると思われる。

Table 1 から，w_2 の値が大きいほど，認識率の平均値が大きくなっているが，話者別に見ると，w_2 を大きくすることによって認識率が改善される場合と，そうでない場合とがあるのがわ

Table 2　Confusion matrix for 10 digits uttered by 5 male speakers ($w_2=8$).

Input	Output										Recog-nition rate(%)
	zero	itʃi	ni	san	jon	go	roku	nana	hatʃi	kju	—	
zero	45				1	2				2		90
itʃi		44							4		2	88
ni			50									100
san	2			47	1							94
jon					50							100
go					2	48						96
roku							50					100
nana						1		49				98
hatʃi									50			100
kju										50		100
											Average	96.6

かる。話者A，Eは認識率が改善されているが，他の話者では，あまり変化していない。荷重係数w_2の値は，1，2，4，8の4通りしか試していない。さらにw_2を大きくすると認識率が下がることが予想される。

女性の声については，実験を行わなかった。

6　結　　言

零交差情報によって離散単語認識を行う特定話者用の装置を，1枚の基板上に組んだ零交差分析装置と，パーソナル・コンピュータとを組み合わせて実現した。この装置は，零交差分析の結果得られる，音声のホルマント周波数F_1，F_2に対応するデータを用いて，パターン・マッチング（DPマッチング）によって認識を行う。男性話者5名の発声する数字音声に対する認識率の平均値は，96.6%であった。認識率の測定は，各話者に各単語を11回発声してもらい，最初に発声した1単語を標準パターン，他の10単語を入力パターンとして行った。

7　謝　　辞

話者として協力してくれた，田中研究室の木村君，桑原君，平野君，川島君，小林君，藤沢君に感謝する。

参 考 文 献

(1) 坂井利之，堂下修司：会話音声識別装置，信学誌，46，11，1696-1702 (1963).
(2) 加藤康雄，千葉成美，永田邦一：数字音声識別装置，信学誌，47，9，1319-1325 (1964).
(3) 大泉充郎，藤村靖：音声科学（東京大学出版会，1972）p. 203.
(4) 電子通信学会編：聴覚と音声（電気通信学会，1966）p. 323.
(5) 坂井利之：情報基礎学詳説（コロナ社，1983）pp. 186-190