パーソナルコンピュータを用いた音声情報処理システム

パーソナルコンピュータを用いた音声情報処理システム

高良富夫＊　亀山俊昭＊　屋宜盛俊＊

A Speech Information Processing System Using a Personal Computer

Tomio TAKARA, Toshiaki KAMEYAMA, and Moritoshi YAGI

Abstract:

Personal computers are more efficient than main frame computers in terms of portable and personal use, on-line processing, and conversational operation with a graphic display. To exemplify this, we present a new speech information processing system, which incorporates a speech analysis sub-system and a spoken word recogintion sub-system.

The speech analysis sub-system can be operated easily and repeatedly, because it adopts a light-pen and the menu method. It can also play back any segment of sound from any part of the speech. By using this function, we can get new knowlege of the auditory response to speech sound.

The spoken word recognition sub-system displays a running process of the system on the graphic display, and it makes it easier to understand the structure of speech recognition. The system is useful for demonstration of speech recognition and serves as a basis for developing practical systems, because it can recognize ten numbers, from zero to nine, uttered by males at a rate of almost 100%.

1．まえがき

近年あらゆる分野に電子計算機が導入され，大抵かつ高速な情報処理が行われている。音声情報処理の分野で電子計算機が使用され始めたのは比較的早く，1950年代初頭にはすでに初期の音声認識システムが発表されている。[1]

電子計算機を音声情報処理に利用する場合，その使用形態は大よそ次のようになる。

（i）オンライン処理

（ii）音声波形のグラフィック表示

（iii）対話形式による波形処理

（iv）音響実験室での使用

（v）長時間の独占的使用

これらの使用形態は，通常の汎用大型計算機の利用法にはなじまない。そこで，音声情報処理においては，従来，ミニコンピュータが多用されてきた。さらに最近は，パーソナルコンピュータが廉価になったことから，これが，ミニコンピュータに代わるものとして注

受付：1985年4月30日

＊琉球大学工学部電子・情報工学科

目されている。[2]

しかし，パーソナルコンピュータは，汎用大型計算機に比較して演算速度が遅く，しかも使用可能なメモリ量が格段に少ないという欠点がある。

ここに提案する音声情報システムは，以下のように，この欠点を除去する。まず，数値演算専用プロセッサi8087を内蔵することにより，一般の数値演算速度を向上させ，さらに，音声情報処理において最も演算時間を要する高速フーリエ変換（ＦＦＴ）を専用のハードウエアで行うことにより，演算時間を大幅に低減する。又，フロッピーディスクを介して，プログラムおよび大量のデータを大型計算機と相互に転送することにより，大型計算機の資源を有効に利用する。

このように本システムは，パーソナルコンピュータの長所を生かした上で，汎用大型計算機にも劣らない性能を発揮する。次章以下に，本システムのハードウエア構成，およびソフトウエアにより構成した音声分析サブシステムおよび単語音声認識サブシステムの概要を紹介する。

2．ハードウエア構成

図１に，パーソナルコンピュータを用いた音声情報処理システムのハードウエア構成を示す。このシステムの中心はパーソナルコンピュータである。音声は，マイクロフォンまたはカセットデッキから入力され，コントロールボックスを通過した後，ＡＤ変換器によりディジタル化され，パーソナルコンピュータに入力される。入力された音声は，パーソナルコンピュータおよびＦＦＴ専用ボードにより処理される。処理結果は，フロッピーディスクおよびプリンタに出力されるとともに，ＤＡ変換器，コントロールボックスを経て，スピーカまたはヘッドフォンから音として出力することができる。

各部の詳細を以下に述べる。

2.1　パーソナルコンピュータとその周辺機器

パーソナルコンピュータは，沖電気社製のif800モデル50を使用しており，これは，ＣＰＵとして16ビットのマイクロプロセッサi8086-2を内蔵している。クロック周波数は8 MH_zであり，これは，アセンブラでプログラミングすれば，音声のＡＤ／ＤＡ変換速度に十分追従できるプログラムを作成することができる。さらに，このパーソナルコンピュータは，数値演算プロセッサi8087－２を内蔵したマルチプロセッサ構成になっており，これにより，演算速度が約100倍高速化されている。

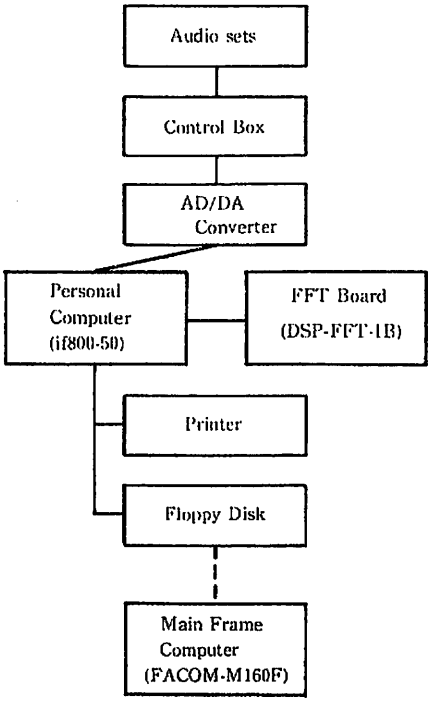


Fig. 1. System Hardware.

記憶容量は256ＫＢあり，本システムでは，さらに256ＫＢ増設して使用している。ＣＰＵは１ＭＢまで直接アクセス可能であるから，さらに512ＫＢ増設することができる。増設したメモリ部はメモリディスク（キャッシュメモリ）として使用できるので，これに処理の途中結果を格納すれば，処理速度が，フロッピーディスクを使用した場合に比較して格段に速くなる。

if800モデル50には８インチタイプのフロッピーディスクが２台実装されていて，同タイプのフロッピーディスクは，当学科の大型計算機ＦＡＣＯＭ－Ｍ－160Ｆにも装備されているので，両計算機で，プログラムおよびデータを共用することができる。これにより本システムは，音声情報処理のトータルシステムとして最大限の機能を発揮する。

プリンタは，24×24ドットの明瞭な漢字が出力できるものを使用しており，これは，プログラムおよびデータのプリント出力に使用するとともに，パーソナルコ

ンピュータをワードプロセッサとして使用して，論文作成等にも利用している。又，このプリンタにより，音声波形の図をハードコピーとして残すことができる。尚，このプリンタのプラテン幅が大型計算機のそれと一致することから，通常は大型計算機の大量の使用済不用紙を活用することができ，省資源に大いに役立っている。

音声情報処理では必須であるグラフィックディスプレイ装置は，解像度が640×475ドットであり，この種のものとしては高解像度の部類に属する。音声波形等をこれに表示し，縮尺したハードコピーをとれば，論文原稿用図面として十分使用可能なものとなる。図4および図6はこれにより作成した。又，本システムにはライトペンが装備されており，これとこのディスプレイ装置を活用することにより，人間工学的で使いやすい対話型音声情報処理システムを構成することができた。これについては次章以下で述べる。

2.2　音響装置

音響装置は，マイクロフォン，カセットデッキ，アンプ，ヘッドフォン，およびスピーカから成る。

マイクロフォンおよびカセットデッキは，それぞれソニー社製ＥＣＭ－260ＦおよびＴＣ－ＦＸ505Ｒを使用している。カセットデッキは，リモートコントロールが可能であり，将来この端子を利用して，音声の自動入力を行う予定である。又，カセットデッキは，音声をマイクロフォンからパーソナルコンピュータへ送出するためのプリアンプとしても使用している。

アンプおよびスピーカは，それぞれＤＥＮＯＮのＰＭＡ－930およびＳＣ－Ｃ9であり，ヘッドフォンは，パイオニア社製のＳＥ－ＤＪ１を使用している。これらは，音声分析および合成の結果と，音声の聴取結果とを比較するために使用する。ヘッドフォンには接話型マイクロフォンが実装されており，これは，高雑音下での音声認識の実験で使用する。

2.3　コントロールボックス

音声入力端子および出力端子を切替えるため，コントロールボックスがある。この中には，ＡＤおよびＤＡ変換器のそれぞれ入力部および出力部に必要な低域通過フィルタが内蔵されている。又，ＤＡ変換出力端子は，２個用意されており，この出力信号により，音声波形をオシロスコープ上に高速表示することができる。

2.4　ＡＤ／ＤＡ変換器

パーソナルコンピュータのＩＯスロットには，ＡＤ変換器およびＤＡ変換器が内蔵されている。量子化精度は12ビットであり，音声の研究には十分な精度である。一般に，音声信号に対して標本化周期は100μs以下である必要があるが，本システムの標本化周期は，ソフトウエアにより指定でき、ＡＤ変換では最小60μsまで，ＤＡ変換では20μsまでこれを下げることが可能である。

2.5　ＦＦＴ専用ハードウエア

音声情報処理では，信号の分析法としてスペクトル分析を用いることが多く，これを高精度に行うためにはフーリエ変換が必要となる。フーリエ変換の演算に要する時間は多大なものであり，実験時間の大部分はこのために費されると言っても過言ではない。従って，

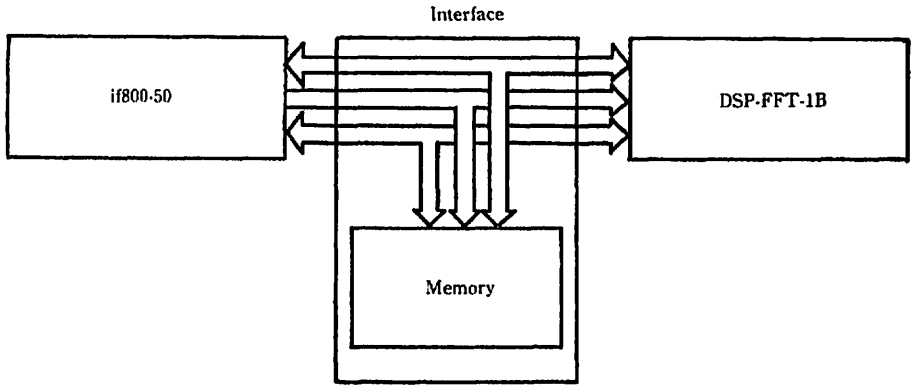


Fig. 2. Interface System.

いかにフーリエ変換を高速に行うかが，音声情報処理の研究の生産性を向上させる上で非常に重要である。

フーリエ変換を高速に行うアルゴリズムのひとつとして，高速フーリエ変換（ＦＦＴ）があり，これをプログラミングすることにより，この問題はかなり軽減される。ここでは，これを利用するとともに，さらに高速にするため，ＦＦＴ専用のハードウエアを使用している。採用したハードウエアは，米国ＤＳＰ社のＤＳＰ－ＦＦＴ－１Ｂであり，これは，16ビット演算精度で1024点のＦＦＴを8.57msで実行する。この演算処理速度は，当学科の大型計算機の演算処理速度に比較しても約10倍速く，これは，音声分析においては，十分な実時間処理速度であるということができる。

バス構成およびデータ転送方式が異なること等から，ＤＳＰ－ＦＦＴ－１Ｂは，このままではパーソナルコンピュータでは使用できない。そこで図２に示すようなインターフェース回路を製作した。ＤＳＰ－ＦＦＴ－1Bは、独立のプロセッサとして主体的に動作するので，パーソナルコンピュータとのデータのやりとりは，インターフェース回路に内蔵されたメモリを介して行われる。このメモリは，容量が16ビット×１Ｋ語であり，512点複素フーリエ変換の１回分のデータを格納することができる。

3．　音声分析サブシステム

音声認識や音声合成を行う場合，音声区間を検出することが必要になる。しかし，これを自動的に精度よく行うことは，一般に困難である。本サブシステムでは，音声研究の基礎的データを得るため，音声区間を視察により決定し，音声データをフロッピーディスクに格納することを第一の目的とする。これに加えて，音声区間の任意の部分を切り出して，拡大，聴取することができる。

3.1　システム構成

マイクロフォンまたはカセットデッキから入力された音声は、５ kHz 低域通過フィルタを通過した後，ＡＤ変換器によりディジタル信号に変換される。このとき，アナログ電圧－５Ｖ～５Ｖが数値－2048～2047へ変換される。これをパーソナルコンピュータのメモリへ格納した後，種々の処理が行われる。

処理された音声信号を実際の音として聴取する場合は，パーソナルコンピュータからＤＡ変換器に信号が送られ，アナログ信号に変換される。変換された信号は５ kHz 低域通過フィルタを通過することにより，スムージングされ，スピーカまたはヘッドフォンから出力される。

このサブシステムのメインプログラムは FORTRAN 言語で作成し，基本的な入出力プログラムはアセンブラ言語で作成した。

音声処理のフローチャートを図３に示す。

入力としてはＡＤ変換入力とフロッピー入力がある。フロッピー入力は，以前に格納したデータを処理する

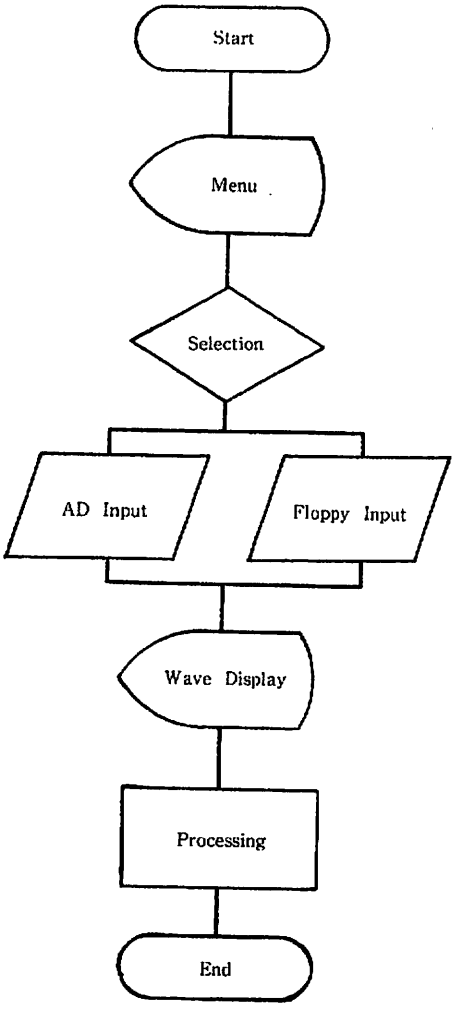


Fig. 3. Flow Chart of Speech Processing.

ためにある。入力された波形がディスプレイに表示されるとともにメニューが現れる。メニューをライトペンで選択することにより，種々の処理を行うことができる。

メニューの内容は，ハードコピー，拡大，全波形，格納，ＤＡ変換，およびＡＤ変換である。

「ハードコピー」は，ディスプレイ画面をプリンタにコピーするもので，縮小，標準，拡大の３種から選択することができる。

「拡大」は，音声波形の任意の区間を時間軸方向に拡大するものである。ディスプレイ上に表示された音声波の拡大したい部分にライトペンをあて，始点，終点の順で指定する。ディスプレイ上には左右２つの波形が現れ，拡大された波形は，始点・終点が指定された波形の反対側に現れる。

「全波形」は，拡大のくりかえしにより画面から消えた元の入力波形を表示する。

「格納」は，拡大された波形をフロッピーディスクに格納する。

「ＤＡ変換」は，音声波形をスピーカに出力する。出力の仕方は次の３種から選択することができる。すなわち，拡大された区間の音声出力，全波形の音声出力，および拡大された区間のくりかえし出力である。「くりかえし出力」により，音声波形の任意の区間の１周期分を切出し，これがどのように聴こえるかを調べることができる。

「ＡＤ変換」は，データを新たに入力したい場合に選択する。

以上のように本サブシステムは、すべてメニューで選択でき，メニュー選択および音声区間の指定はライトペンで行うので，非常に使いやすい。従って，くりかえして使用することができ，これにより，音声の聴こえに関して新しい知見が得られるものと期待できる。

3.2 使 用 例

音声分析サブシステムを使用した例を図４に示す。音声データは／aia／と発声したものであり，サンプリング周波数は10kHzとした。図の番号に従って説明する。

(1) 右側の図は，入力した／aia／の波形である。時間方向は下から上へ向かう。尚，左右両図の中央の太い線は，ライトペンの反応をよくするためのものである。左側の図は，右側の図の／a／の部分を拡大したものである。この図が表示された後，「ＤＡ変換」の「拡大された部分の音声出力」を選択し，これを

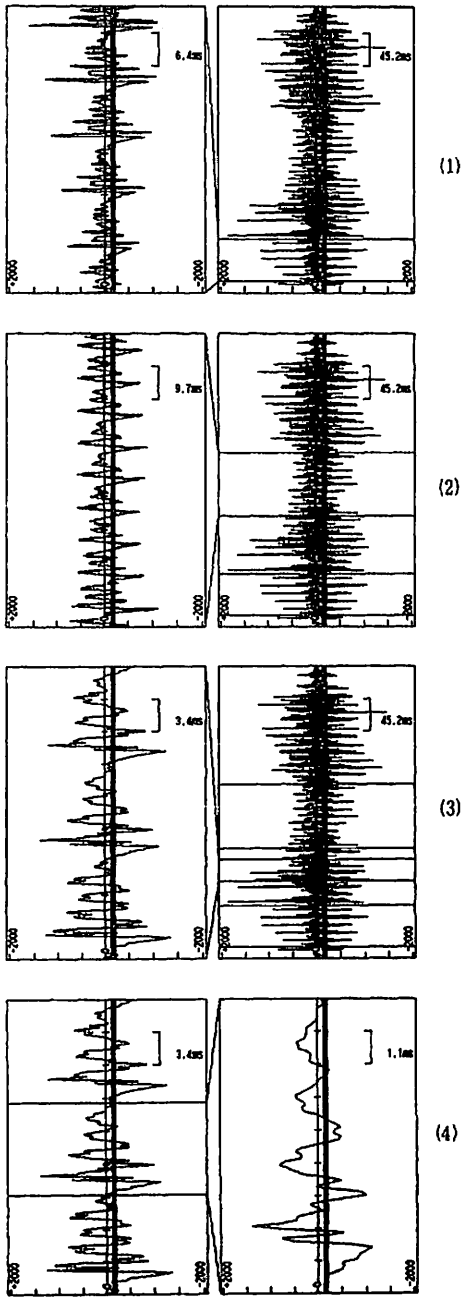


Fig. 4. Examples of Display of the Speech Analysis Sub-system.

聴取したところ，確かに／a／と聴こえた。

(2) 同様に左側が／i／の部分を拡大したものである。この部分は，確かに／i／と聴こえた。

(3) 左側の図は，／a／と／i／の中間部を拡大したものである。この部分は／a／でも／i／でもないように思われる。

(4) (3)の左側の波形の1周期分を切出し，拡大したものが右側の図である。このとき「DA変換」の「拡大された区間のくりかえし出力」を選択し，これを聴取したところ，／e／と聴こえた。この結果から，／aia／と発声したとき，／a／から／i／へ移行していく部分に／e／の性質をもった波形が現れることがわかる。もちろん，／aia／全体を聴取したときは，どこにも／e／は聴こえない。

この例から分かるように，聴覚的には聴こえないが物理的に波形が存在することがある。逆に，聴覚的には聴こえるが，物理的に波形が存在しない場合もありうる。これが，音声自動認識を困難にしている最大の原因であると著者は考えている。

4．　単語音声認識サブシステム

語いを限定し単語ごとに区切って発声した音声を，自動的に認識するシステムを単語音声認識システムという。ここでは10数字単語を語いとする単語音声認識システムを作成した。

このサブシステムは，音声認識のデモンストレーションのために作成されたものであり，音声認識の各過程の中間結果がディスプレイ上に表示される。これにより，音声認識の仕組および当研究室の研究内容が，直観的に理解できる。

4.1　システム構成

単語音声認識サブシステムの仕様および動作のフローチャートをそれぞれ表1および図5に示す。

語　　い	0－9の10数字
話　　者	不特定多数話者
発 声 法	孤立発声
分 析 法	ケプストラム分析 メル・ソーン・スペクトル分析
識 別 法	DPマッチング

Table 1. Specifications of the Word Recognition Sub-System.

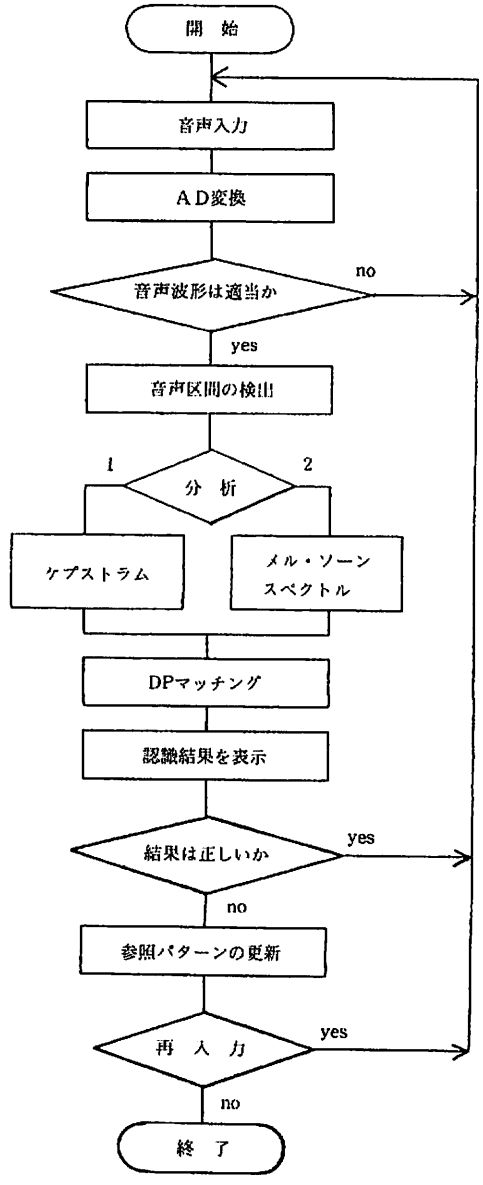


Fig. 5. Flow Chart of the word Recognition Sub-System.

入力された音声波は，遮断周波数が5 kHzの低域通過フィルタを通過した後，サンプリング周波数10kHzでAD変換される。このとき，音声波形の振幅が適当な範囲になければ，再び音声を入力する。

次に，入力されたデータから音声の存在する区間の始点および終点がRabinerらの方法[3]により検出される。

音声区間の音声波は，25.6ms長，20.0ms間隔のブラックマン窓で切出され，切出された部分（フレーム）ごとにケプストラム分析またはメル・ソーン・スペクトル分析が行われる。ケプストラムとは，対数スペクトル振幅の逆フーリエ変換である。メル・ソーン・スペクトル[4]とは、振幅スペクトルの周波数軸を音の高さの心理尺度であるメル尺度に変換し，振幅軸を音の大きさの心理尺度であるソーン尺度に変換した新型のスペクトル・パラメータである。これは，より人間の聴覚特性に適合したスペクトル・パラメータであり，当研究室オリジナルのものである。

識別法としては，テンプレートマッチング法のひとつであるDPマッチング法[5]を用いる。DPマッチングの結果は棒グラフで表示される。

認識結果は，数字で示される。認識結果が正しくない場合は，入力された音声を修正用パターンとして利用し，参照パターンを修正する。これにより，次回以後，同様の入力パターンに対して正しい認識が行われる。

このサブシステムは，11名の男性話者が発声した10数字単語音声に対して試験された。[6]その結果によれば，6名以上の話者の音声から参照パターンを作成すれば，認識率はほぼ100%となる。尚，動作時間は実時間の約30倍であるが，現在まだFFT専用ハードウエアを使用していないので，これを使用すれば，さらに実時間に近づくものと考える。

4.2 使用例

単語音声認識サブシステムを使用した例を図6に示す。

このシステムでは，数字を／it∫i／，／ni／，／san／，／jon／，／go／，／roku／，／nana／，／hat∫i／，／kju／，／rei／のように区切って発音することにする。この例では，／roku／と発声した。

以下，図の番号に従って説明する。

(1) これは，入力された音声の始点と終点，すなわち音声区間を検出している時の図である。

図では省略したが，この前に，音声振幅が自動的に評価されていて，振幅が大きすぎる場合は〝大きすぎます〟と，又，小さすぎる場合は〝小さすぎます〟というメッセージが現れ，音声は入力されない。

図は上から，音声波形，音声パワーの時間変化，零交差数の時間変化を表している。音声波形を囲んでいる長方形は，自動的に検出された音声区間を示している。これは，パワーと零交差数を利用して検出される。

(2) これは，検出された音声区間をメル・ソーン・スペクトル分析法により分析している時の図である。左図は，検出された区間の音声波形であり，右図は，1フレームごとのメル・ソーン・スペクトルである。

(3) これは，DPマッチング法により，入力パターンと10種の参照パターンそれぞれとの距離（類似度）を計算している時の図である。0，1，2，・・・の順序で距離が計算され，計算され次第，棒グラフで表示される。最後に10個の距離の中から最小のものが検出され，最小の部分は黒くぬりつぶされる。この例では〝6〟の参照パターンと入力パターンとの距離が最小であった。

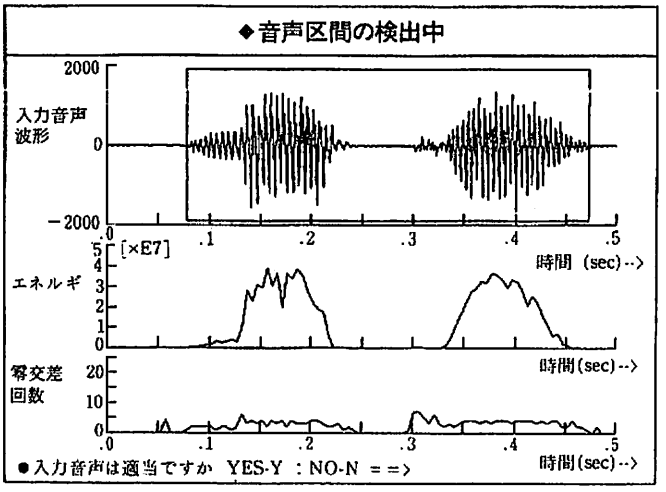
(4) 最後に，認識結果が表示される。ここでは〝6〟が表示され，正しく認識できたことがわかる。

このとき，〝正しい認識結果は得られましたか。というメッセージが現れる。これに，もし〝NO〟とキー入力で答えると，次に〝入力した音声は何ですか。。というメッセージが現れる。これに，正しい数字をキー入力することにより，その数字の参照パターンが先ほどの入力パターンで修正される。

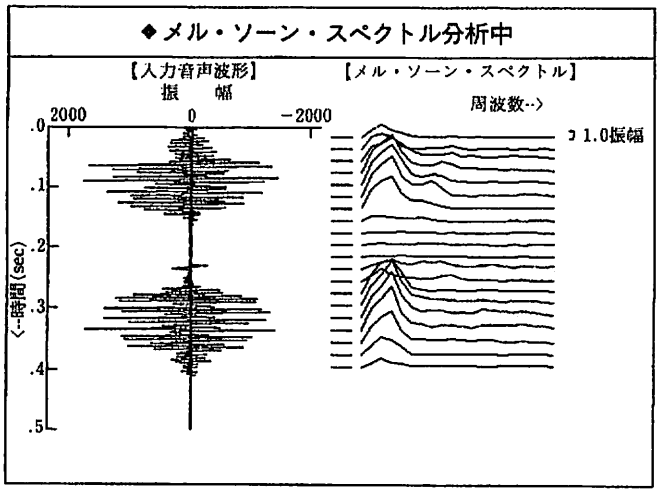
5. むすび

パーソナルコンピュータを用いた音声情報処理システムを提案した。

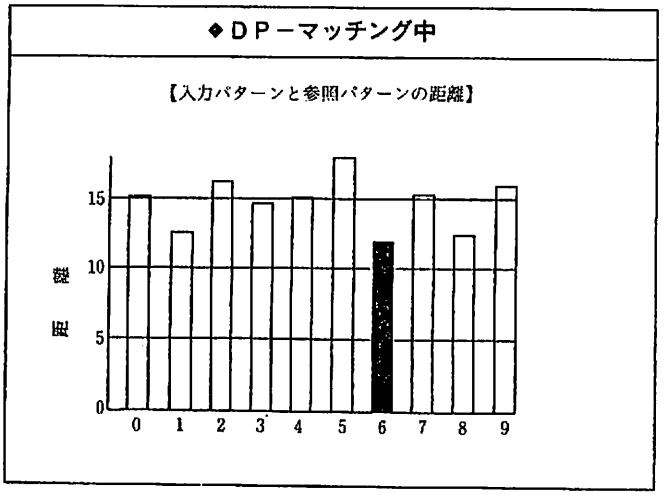
一般に，パーソナルコンピュータは，オンライン処理，グラフィック表示の対話的処理，可動性，パーソナル・ユースなどにおいて，汎用大型計算機より優れている。この実例として，本システムの音声分析サブシステムおよび単語音声認識サブシステムを示した。又，パーソナルコンピュータは，演算速度および記憶容量において，汎用大型計算機に劣ることから，本システムては，数値演算プロセッサおよびFFT専用ハードウエアを使用し，かつフロッピーディスクを経由して大型計算機と結合することにより，この欠点を除去した。


◆音声区間の検出中
入力音声
波形
2000
0
−2000
.0
.1
.2
.3
.4
.5
時間 (sec)-->
[×E7]
エネルギ
時間(sec)-->
零交差
回数
時間(sec)-->
●入力音声は適当ですか YES-Y ：NO-N ==>

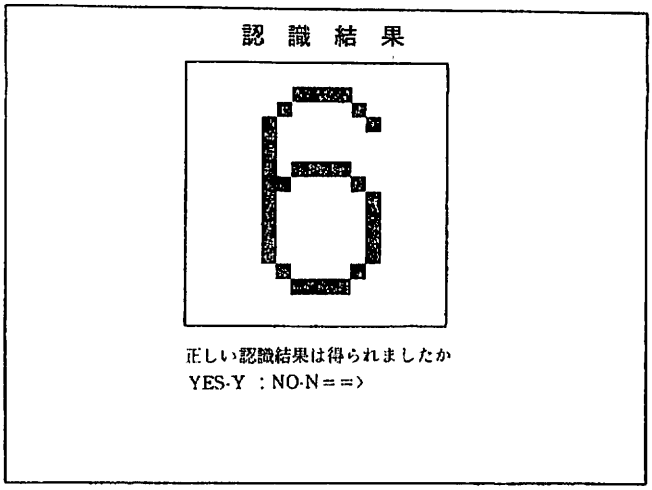
(1)


◆メル・ソーン・スペクトル分析中
【入力音声波形】
振 幅
【メル・ソーン・スペクトル】
周波数-->
2000
0
−2000
1.0振幅
<--時間(sec)

(2)


◆ＤＰ－マッチング中
【入力パターンと参照パターンの距離】
距離
15
10
5
0
0 1 2 3 4 5 6 7 8 9

(3)



(4)

Fig. 6. Examples of Display of the Word Recognition Sub-System.

ここで示した音声分析サブシステムは，メニュー方式およびライトペンの採用により，人間工学的で非常に使いやすいものとなっていて，これを使用することにより，音の聴こえに関して新しい知見を得ることができる。又，単語音声認識サブシステムは，グラフィック表示の使用により，音声認識システムの仕組みを理解することが容易である。又これは，男性話者の音声に対して，ほぼ100%の認識率を示すことから，デモンストレーション用として有用であるとともに，パーソナルコンピュータを用いた実用的な音声認識システムの研究の基礎にもなる。

ここでは音声情報処理の分野のうち音声分析と音声認識についてのみ述べたが，他の分野として音声合成がある。これに関して本システムでは，琉球方言の音声を合成することができるが，これについては別報[7]を参照されたい。

謝辞　本研究を行う機会を与えて下さった本学電気系学科の諸氏に感謝致します。

文　献

(1) Moore, R.K.: "Evaluating Speech Recognizers". IEEE Trans. Acoust., Speech, & Signal Process., ASSP－25,2,pp.178-183(1977-04).

(2) 城戸健一："音響学とコンピュータ"，日本音響学会誌，41,1,pp.30-33(1985-01).

(3) Rabiner, L.R. & Sambur, M.R.: "An Algorithm for Determining the End points of Isolated utterances", Bell Syst. Tech. J., 54,2, pp.297-315(1975-02).

(4) 高良・今井："メル・ソーン・スペクトルを用いる母音識別"，信学論 (A),J65－A,8,pp.818-825(1982-08).

(5) Sakoe, H. & Chiba, S.: "Dynamic Programming Algorithm Optimization for Spoken Word Recognition". IEEE Trans. Acoust., Speech, & Signal Process.,ASSP－26, 1,pp.43-49(1978-02).

(6) 亀山俊昭："パーソナルコンピュータを用いた単語音声認識システム"，琉球大学工学部卒業研究論文(1985-03).

(7) 高良・平良・今井："琉球方言音声の合成"，信学論 (D)，J68-D,9(1985-09).