音声制御カラーＴＶ

菅谷昭次、西村　賢、清水哲雄、杉浦洋治
三洋電機　開発研究所

はじめに

従来、音声認識装置と言えば、大型コンピュータやミニコンピュータを使用した大がかりな装置であり、しかも、話者限定の認識装置として、産業用に一部実用化されたにすぎない。しかし、最近のマイクロコンピュータやＬＳＩの技術の進歩はめざましく、音声認識装置の民生機器への応用を可能としたいと筆者らは考え、その一例として、標記ＴＶセットの音声制御装置を開発した。民生機器への応用という観点から、小型で、低価格の装置とする必要があり、技術的には、使用するメモリが少く、効率の良いソフトウエアを開発する事と、回路構成部品を極力少なくする事に主眼をおいた。試作装置は話者限定の単語音声認識装置であり、これにＴＶセットを制御する命令音声の特徴パラメータをあらかじめ登録しておく。次に同じ音声が入力された時、登録されている音声の特徴パラメータと比較して入力音声を認識し、その内容に従って赤外線を送信し、ＴＶセットを制御する。開発装置は試作品であるが、将来的には民生機器として商品化を可能とする事を目標としている。

装置の概要、構成

音声によって、このＴＶセットを動かす為には、あらかじめ操作者の制御命令音声を登録しておかなければならない。制御できる内容は、

- 電源の入・切
- 音量の調整
- ミューティングの入・切
- チャンネルの変更

である。

命令音声は上記各制御内容を指示する単語（ノ6語）と、その動作を促す単語「ＯＫ」により構成されている。即ち「電源」「ＯＫ」という発声により、電源の［入］又は「切」の制御が出来る。音量の増加は、「ＵＰ」，「ＯＫ」という発声により始まり、「ＳＴＯＰ」と言う発声により停止する。音量の減少も同様に、「ＤＯＷＮ」，「ＯＫ」という発声により始まり、「ＳＴＯＰ」という発声により停止する。但し、「ＳＴＯＰ」という発声は、応答動作の迅速性を保障する為、その音声信号のレベルのみが判定される。「ＭＵＴＩＮＧ」，「ＯＫ」という発声により、ミューティングの［入］又は［切］が制御される。

チャンネルについては、例えば「３」，「ＯＫ」という発声により、ノ～ノ2までのチャンネルが指定される。

このシステムは、「ＯＫ」、「電源」、「ＵＰ」，「ＤＯＷＮ」、「ＭＵＴＩＮＧ」，「ノ」～「ノ2」に相当するノ7語を一人の操作者について登録しなければならない。

登録言語に関する制約事項は、これらの言語時間長が、ノ　6秒以内である事だけで、操作者の好きな言語が選択できる。

登録する際、データの圧縮処理が行なわれる。登録回数は、ノ回以上で、2回

目以降は学習処理が行なわれ、変動の小さい特徴パラメータが有効に認識演算処理に利用される構成となっている。音声の始端部分は、循環処理により、重みづけられ、登録される。このようにして作られた登録パターンは、認識率に好ましい結果を与えるものであり、又、記憶メモリの容量は比較的少なくてすむ。

登録人数は、試作装置においては2名であり従って、全登録語数は8ケ語である。認識処理時間は実際に記憶をしている語数により異なるが、最大の場合でも0.5秒である。登録パターンメモリはRAMであり、バックアップ電源を持っていないので、電源が失なわれると登録パターンも失なわれることに留意しなければならない。

認識モードに於いても、登録モードと同様、入力単語の時間長は1.6秒以内でなければならない。この時間は、テレビジョンの前記の制御を命令する単語を通常の速度で発声するのに、充分な時間である。入力音声信号パターンが登録パターンのどれかに特定出来た時、装置は前面パネルの対応するランプを点灯する。

この装置は第一図のブロック図に見られるように、入力音声を分析して音声の特徴を抽出するアナログ的な回路部分、デジタルコード化された音声の特徴パラメータにもとづき認識処理するCPU及びこれらに入力制御部、表示部、テレビジョンとのインタフェースを付加した構成となっている。以下、これらの構成要素（ハードウェア）と、主としてCPUを中心とする処理プログラム（ソフトウェア）について順次その内容を説明する。

音声入力部

音声の入力は装置に組み込んだマイクロフォンに直接話しかける事により行なわれる。我々の希望する音声入力部の特性としては、話者の声のみを一定レベルで出力し、周囲の雑音は出力しないという相反する特性を要求する。そこで、マイクとしては、距離に対する減衰効果の大きい接話マイクを選び、バッファアンプとして、非直線AGC回路を採用する事により、上記特性に近いものを得るよう構成している。しかし、残念ながら、接話マイクは距離減衰特性が、周波数特性を持ち、話者とマイクの距離の変動が単に音声信号の大小のみならず、特徴パラメータの相対変動を引き起こし、認識率に悪影響を与えていると考える。

特徴抽出回路

音声信号の中から、言語を構成する音韻的特徴を抽出する方法としては、これまで自己相関関数、線型予測係数、周波数スペクトル、フォルマント周波数など、多数のパラメータが知られ、それぞれの性能が論じられている。筆者らは、音声認識装置の民生機器への応用という観点から、コストパーフォーマンスを重視し、比較的単純なハードウェアで実現できるフィルタにより音声信号の周波数スペクトルを分析する方法を採用した。100 Hzから5.0 KHzまでの周波数帯域を8個のフィルタにより分割している。減衰特性は24 dB／octのアクティブフィルタである。フィルタの数は更に多いほうが、認識率の点からは好ましいが、フィルタの出力を時分割処理する為に順次切り替える、アナログマルチプレキサのチャンネル数が8の倍数であること、及びCPUの処理データ量を可能な限り必要最小限におさえる為に、上記の個数とした。

波形等化方式

同一話者の同一言語音声であっても、発声の都度その時間軸、信号振幅とも、変動するのが普通であり、それぞれについて、何らかの正規化が必要である。この装置では、時間軸の正規化については、単語音声の開始点から終端までの時間軸を等分割する方法をとっている。この方法は、ダイナミックプログラミング手法に比べ、性能的には若干劣るが、処理時間とバッファメモリ容量の点でマイクロコンピュータ処理に適している。単語音声の開始点及び終端検出は、フィルタ出力を参照する事により、CPUが行なっている。

信号振幅の正規化は、8つのフィルタ出力の加算平均値と、各フィルタ出力値の比を演算し、8ビットのフィルタ出力値を2ビットにデータ圧縮する事と併せて行っている。

CPUシステム及びソフトウェア

この装置の全ての制御、演算処理はCPUを中心に行なわれる。CPUは、8ビットの比較的高速のものであればどれでも使用できる。我々は、インテル社の8085Aを採用した。これに、PROM2KB(8755A)、RAM2KB(8185X2)、I/Oポート(8155)とでシステムを構成した。CPUのマスタクロックは6.1MHzである。1.5KBのプログラムの内容は次の通りである。

- モードの判定
- 各クロック制御とデータの取込み
- 登録パターンの学習を含む登録処理
- 時間軸の正規化を含む入力データ処理
- 認識演算処理
- 入出力シーケンスコントロール

本装置に高い認識率と速い処理速度をもたらし、かつ経済的な価格を可能とする為、音声信号をスペクトル分析する8個のフィルタの機能を最大限にまかす認識演算処理プログラムを開発した。以下そのまな内容について説明する。

(1)データの取込み

8つのフィルタ出力は10ミリ秒おきにサンプリングされる。8ビットのA/D変換器を使用している為、1つのフィルタ出力に対して、1バイトのデータが対応するが、8つのフィルタ出力値の加算平均値と、各フィルタ出力値との比を演算し、8ビットのフィルタ出力値を2ビットにデータ圧縮する。従って、8つのフィルタ出力は、2バイトのデータに演算加工され、循環レジスタに常時入力されている。循環レジスタの記憶容量は4サンプル分である。このレジスタは音声検出回路の僅かな動作遅れを修正し、音声の先行子音部を確実に把握するのに必要である。バッファレジスタには、この循環レジスタの分は含めず、最大1.6秒間160サンプル点の320バイトのデータが取込まれる。この入力データは時間軸の正規化により、20サンプル点がバッファレジスタから取出され、1単語につき、24サンプル点48バイトのデータが特徴パターンとなる。この特徴パターンの抽出方法は、登録、認識両モードで共通している。34語の登録の為、登録パターンメモリーとして、163バイトが必要であり、320バイトのバッファレジスタと合せると、1952バイトとなる。

（２）登録時の学習機能

既に登録されている単語を再び登録した場合、入力パターンと、登録パターンとの間で学習処理が行なわれ、この学習された結果が新たな登録パターンとして、記憶される。学習処理により、変動幅の大きい特徴パラメータは、認識演算処理に際し、入力パターンの特定に作用する割合が、小さくなるようデータ加工される。

（３）入力パターンの特定（認識演算処理）

入力パターンと登録パターンの照合は、各対応するサンプル点での各対応する特徴パラメータの各２ビットを演算し、その総和を求める事によりなされる。この計算において、８つの特徴パラメータのデータが２バイトのデータとなっているが、それを並列計算するプログラムにより、演算処理されるので、認識演算処理の高速性を実現した。

結び

現在の音声認識の研究は、話者限定の単語認識装置については、高度な水準に達していると考えられるが、家庭内の機器を音声により制御するという民生用の観点から、コスト、サイズ、認識性能のバランスを求めた研究は少なかった。そこで筆者らは、特徴抽出のハードウェアを可能な限り小さくし、又、このハードウェアに依り得られた特徴パラメータを、少ないメモリ容量にて、最大限に生かす認識処理プログラムを開発する事により、片手に持てる程小さく、低価格な構成が可能となる音声認識装置を開発した。これにより、音声認識装置の民生機器への応用の見極めがついたと筆者らは考える。この装置は、協力的な話者であれば、認識率は、ほぼ９５％程度であり、音声入力回路に改善を加える事により、再に認識率の向上を目指すが、民生機器への応用という観点から、認識率が１００％に近いという事よりもコンパクトで低価格な構成が、再に要求されると考える。

［参考文献］

* J.A.Clark,"A spoken word recognition system for aiding the severely paralyzed," Proceedings of the COMPCON SPRING '79, pp. 19-22,1979.;

* B.Georgiou,"Give an Ear to Your Computer," BYTE, pp.56-91,June,1978.;

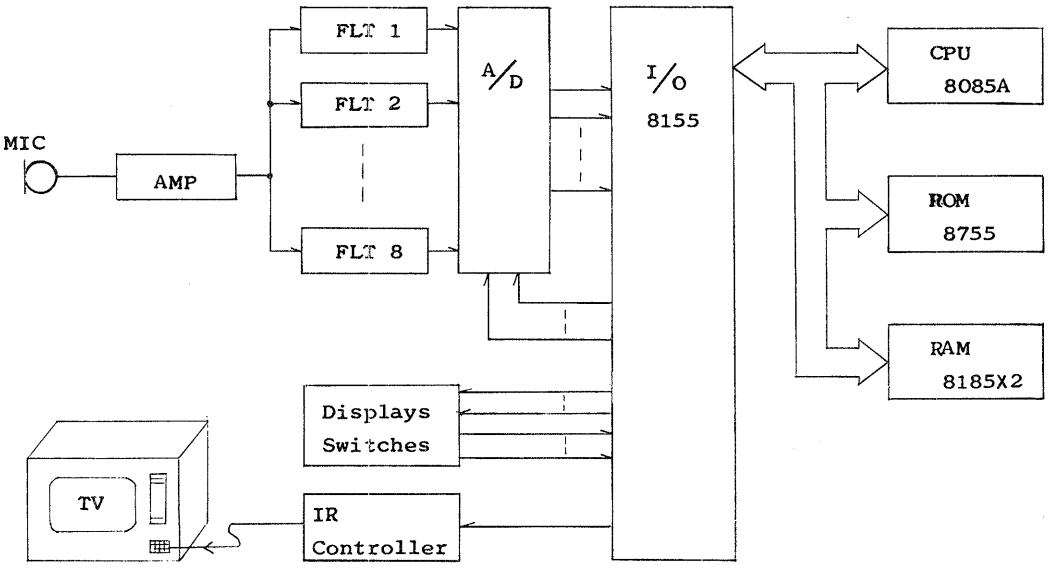
MIC
AMP
FLT 1
FLT 2
FLT 8
A/D
I/O
8155
CPU
8085A
ROM
8755
RAM
8185X2
Displays
Switches
TV
IR
Controller

Fig 1