ROBUST SPEECH RECOGNITION SYSTEM FOR AVIONICS

Arbind Kumar Singh*, S. Shivashankar* and S. Janarthanan*

Abstract

An isolated Spoken word Recognition system "SRIJAN" (सृजन) has been developed by Aeronautical Development Establishment (ADE). This speaker dependent and robust, system employs energy based end point detection, Linear Predictive Code (LPC) coefficients derived cepstral coefficients as feature vectors and Dynamic Time Warping (DTW) algorithm. The DTW algorithm offers better system performance by minimizing the effect of speaking rate variation. The optimum end point pair (start and end) obtained by taking the average of different end point pairs, resulting from the marginal variation of the lower and upper energy thresholds, results in the improvement of the system performance and reduces the computational complexities.

Many new techniques such as multiple reference patterns, averaged reference pattern, online and interactive online reference pattern updating, Cepstral Mean Subtraction (CMS), etc. have been implemented to enhance the recognition accuracy and simplify the training required.

Keywords : Isolated spoken word, Speech recognition, LPC, Cepstral coefficient, DTW, CMS

Introduction

Some of the modern aircraft's cockpit is equipped with Voice Command System (VCS) in order to reduce pilot's workload and to enhance the system throughput. The VCS offers a natural and quicker means of communication and provides and additional pilot vehicle interface. The Speech Recognition System (SRS) is core of the VCS, which is an interesting and a challenging topic among the engineers and scientists. The development of SRIJAN was to explore the feasibility of a cockpit direct voice command system.

The development of a highly accurate SRS would find an increased use in military, industries, medical field etc. Fig.1 illustrates the block diagram of a SRS, built as a pattern recognizer. The Linear Prediction (LP) analysis has been among the most popular methods for extracting spectral information (features) from speech. The LP analysis provides the coefficients of all pole filter which constitutes the mathematical modeling of vocal tract. The LPC coefficients [4, 6] provide an accurate method to parameterize a spoken word in time domain with less computational complexities. The LP analysis does not resolve the vocal tract characteristics. Since the laryngeal characteristics vary from person to person, and even from utterances within a person of the same words, LP parameters convey some information to a speech recognizer that degrades performance, particularly for speaker-independent system. The LP model is very useful tool to compute cepstral coefficients [3, 7] that are referred as LPC derived cepstral coefficients. The low time varying characteristics of the vocal tract can be represented by few cepstral coefficients, which further improves the system performance, as they are more reliable and robust compared to LPC coefficients.

The rest of the paper is organized as follows : The Speech Recognition System (SRS) contains brief functional description and explains the implementation of the system. The Robust Features of SRIJAN describes implementation of several techniques to enhance the system robustness and explains its multi-user adaptability mechanism. The Results and Conclusions includes the results of works carried out and the conclusion in brief.

Speech Recognition System (SRS)

The system/machine able to understand/recognize spoken words is known as SRS. The SRS can be catego-

* Scientist, Aeronautical Development Establishment, CV Raman Nagar, Bangalore-560 093, India
Email : arvindkums@yahoomail.com; shiva@ade.ernet.in; jana@ade.ernet.in

Manuscript received on 17 June 2003; Paper reviewed, revised and accepted on 08 Jan 2004

rized in different ways such as isolated / connected / continuous word SRS, speaker dependent / independent SRS etc. The detailed description of SRIJAN is covered as follows :

Functional Description of SRIJAN

The SRIJAN comprises of both Hardware (HW) and Software (SW). The block diagram of hardware is shown in Fig.2. The Hardware comprises of EZ-LAB Board (based on ADSP-21062, AD1847 Audio Codes etc) hosted on a Personal Computer. All functional Software modules are implemented in Assembly and 'C' languages.

Data Acquisition : The programmable AD1847 audio codec supports the conversion of analog speech input into digital data. The 8000 samples, obtained by 8KHz sampling frequency, for every spoken word is stored in the memory for further processing. Fig.3 is the plot of digitized data for the spoken word 'Left'. The digital data i.e. 128 samples (digitized data) have been grouped into a frame of 16 ms duration to satisfy the quasi-stationary property of the speech signal.

End Point Detection : The end point detection is performed to identify the beginning and end of the spoken word. From Fig.3, it is obvious that there exists background noise before and after the spoken word, which degrades the system accuracy [9] and increases computational complexities. An energy based algorithm [1], "The point where energy of the signal crosses (up) the lower threshold and before crossing (down) the lower threshold, it crosses the upper threshold", has been implemented as a SW module for end points detection. These thresholds are based on the silent energy and maximum energy of the signal and they are expressed as follows :

```
IE1 = 0.04 x (max_ener-av_sil_ener) + av_sil_ener
IE2 = 4 x av_sil_ener
if (IE1 > IE2) then IETL = IE2
else IETL = IE1
    IETU = 4 x IETL
```

where IETL and IETU are lower and upper thresholds respectively.

The marginal variation in the value of thresholds provides a different end point pair. The determination of optimum end point pair, by taking the average of previous

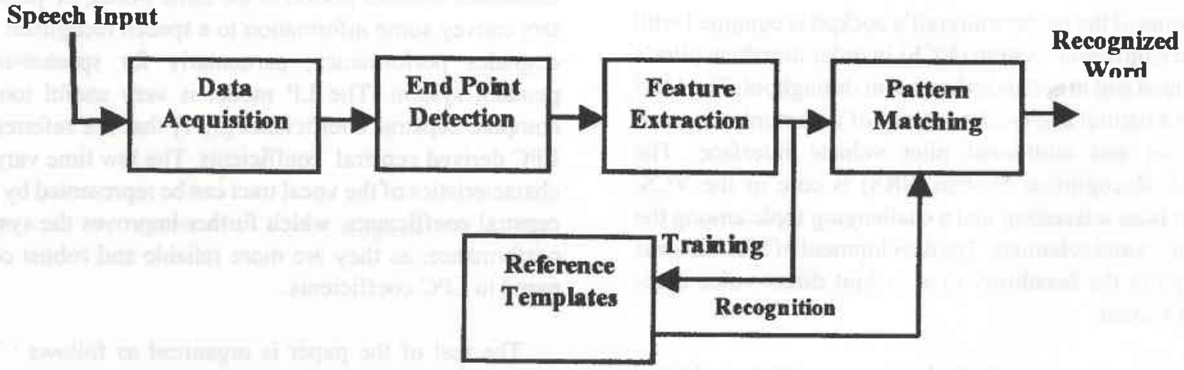

Fig. 1 Block diagram of SRS

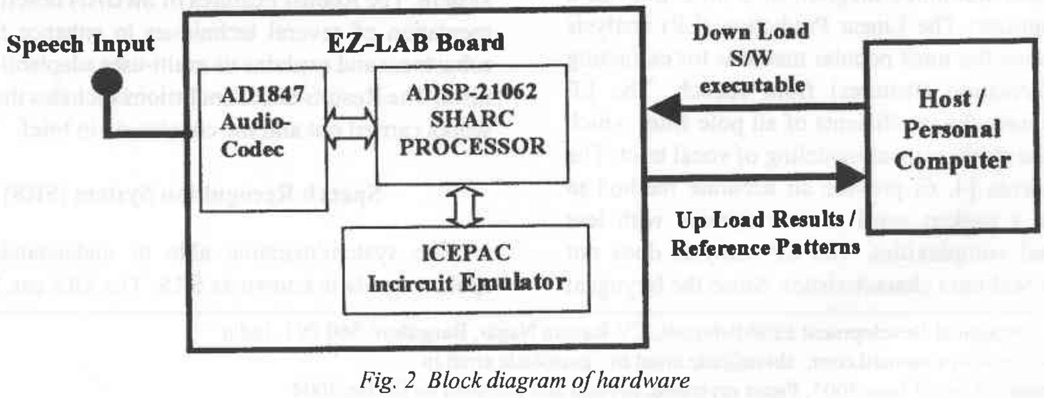

Fig. 2 Block diagram of hardware

two end point pairs, reduces the computational complexities, provides data compression and enhances the recognition accuracy.

The energy plot for the spoken word, 'Left', is given in Fig.4, which also illustrates the thresholds used for end point pair detection. The lower threshold IETL1 and upper threshold IETU1 produces first end point pair (NB1 = 20 and NE1 = 38). Similarly IETL2 and IETU2 generate other end point pair (NB2 = 21 and NE2 = 36). The optimum end point pair (N1 = 20 and N2 = 37) has been generated by taking the average of the above two end point pairs.

Feature Extraction : Feature extraction is performed to extract some useful parameters of the speech signal and to provide data compression. Fig.5 shows the flow of steps implemented for the computation of features-cepstral coefficients.

Frame blocking provides the generation of overlapped segments in order to make variation quite smooth from frame to frame. Windowing minimizes discontinuities at beginning and end of each frame. Auto-correlation is performed on each windowed frame to be used for LP analysis. The Levinson's recursion algorithm [5, 6] is implemented to compute LPC coefficients, which is a faster and efficient method. The robust feature - cepstral coefficients, is computed using the LPC coefficients [3. 7]. The higher order and lower order cepstral coefficients are sensitive to noise and hence weighting them with raised co-sine function coefficients minimizes their effect. The weighted cepstral coefficients, thus derived for the spoken word, constitute the pattern of the spoken word.

Figure 6 illustrates the plot of hamming window, auto-correlation for voiced and unvoiced section of the speech, weighting function, LPC coefficients and cepstral coefficients/frame for the spoken word 'Left'. In Fig.6, the auto-correlation plot of voiced section of speech shows quasi-periodic nature while for unvoiced section it shows the random nature. The plot also shows frame to frame smooth variation of cepstral coefficients than the LPC coefficients.

Modes of Working : The SRIJAN works in two modes namely training and recognition.

Training Mode : In training mode, the system stores the patterns of all the words to be recognized along with their identities. The stored patterns are known as reference patterns or templates. The training procedure for speaker dependent / speaker independent and isolated / continuous word recognition systems are different.

Recognition Mode : In recognition mode, the spoken word pattern is compared with all reference patterns and the reference pattern that shows best match with the spoken word pattern is classified as the recognized word. The non-linear time normalization (between the spoken and reference pattern) is being performed during pattern matching using dynamic time warping (DTW) algorithm [2]. The DTW compresses or expands the patterns in time

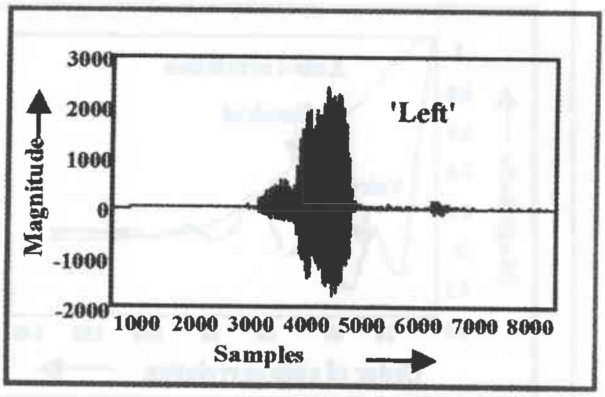

Fig. 3 Digital data for spoken word 'Left'

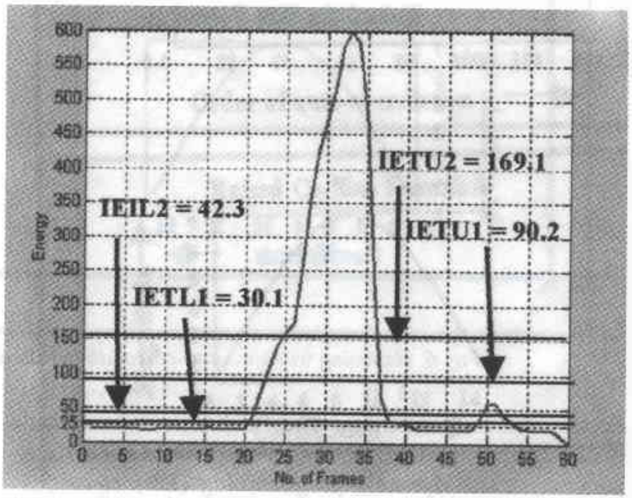

Fig. 4 Energy plot for the spoken word 'Left'

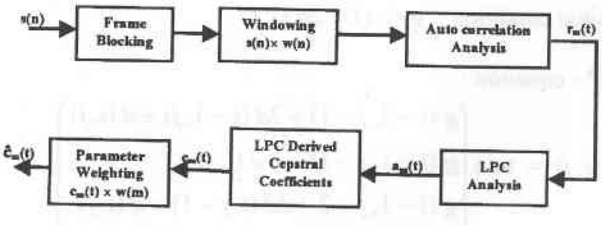
s(n)
Frame Blocking
Windowing s(n)× w(n)
Auto correlation Analysis
r_m(t)
LPC Analysis
a_m(t)
LPC Derived Cepstral Coefficients
c_m(t)
Parameter Weighting c_m(t) × w(m)
ĉ_m(t)

Fig. 5 Feature extraction process

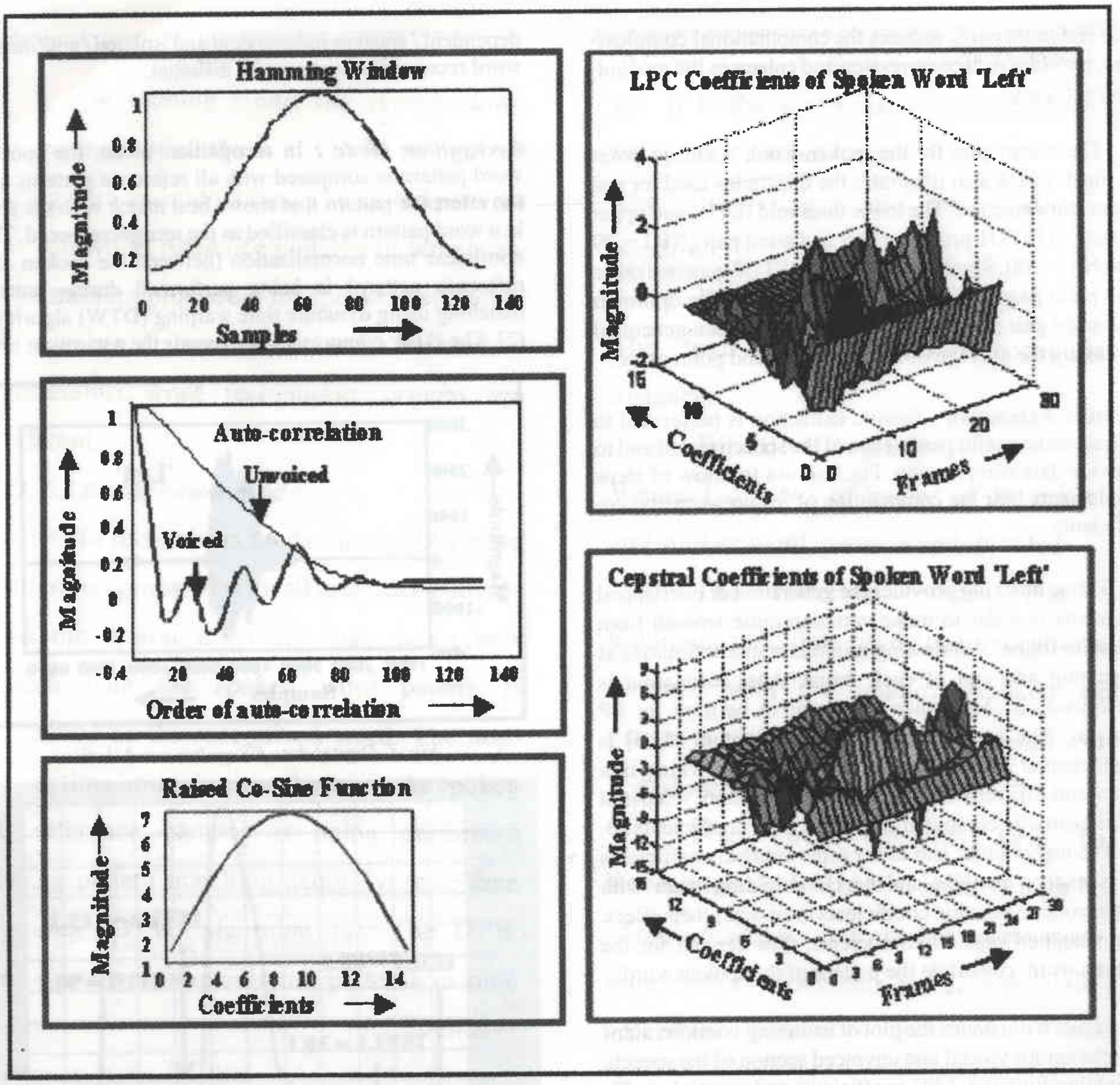

Fig. 6 Hamming window, auto-correlation, raised-cosine function, LPC and cepstral coefficients/frame

domain for best matching. The Euclidean distance between the spoken and reference pattern is used as decision parameter. The dynamic programming (DP) equations used are as follows :

Initial condition : $g\,(1,1) = 2d\,(1,1)$

DP - equation :

$$g\,(i,j) \;=\; \min \begin{pmatrix} g\,(i-2, j-1) + 2d\,(i-1, j) + d\,(i,j) \\ g\,(i-1, j-1) + 2\,d\,(i,j) \\ g\,(i-1, j-2 + 2d\,(i, j-1) + d\,(i,j) \end{pmatrix}$$

where d(i, j) is the euclidean distance and g(i, j) is the accumulated distance between two patterns at point i and j.

The restricting condition (adjustment window) :

$j - r < = i < = j + r$

where r is the window width

The total time normalized accumulated distance :

$D\,(A, B) = 1/N\,g\,(I, J)$

where I, J = No. of frames of the reference and spoken word patterns and N = I + J.

The reference pattern that has minimum total time normalized accumulated distance with the unknown / spoken word pattern is considered as the recognized word.

Robust Features of SRIJAN

Many new concepts have been incorporated in SRIJAN to achieve high level of recognition accuracy over a wide variety of acoustical environments and ease the training procedure. The implementation of several techniques namely, multiple reference patterns, averaged reference pattern, online updating, interactive online reference pattern updating, CMS etc. have enhanced the robustness of the system to a great extent.

Multiple Reference Pattern Technique

This technique uses multiple patterns of the same word as reference patterns. The use of two reference patterns has considerably increased the system accuracy but at the expense of memory and execution time.

Averaged Reference Pattern Technique

The averaging technique of SRIJAN generates a single reference pattern from the two patterns (of frame lengths N1, N2) of the same word. It contains the average of cepstral coefficients for overlapping number of frames and coefficients for non-overlapping frames is limited to a length (N1+N2)/2. This averaging technique can be used iteratively to generate a single reference pattern from more than two patterns of the same word. This technique greatly increases the system accuracy without need of extra memory and execution time.

The algorithm implemented to generate reference pattern R of SEG frames (for the two patterns A and B having N1 and N2 frames; N2> N1) is as follows :

```
a) Compute : SEG = (N1 + N2)/2
b) Find N = minimum (N1 and N2)
c) Do : for i = 0 to SEG-1

Do : for j = 0 to L-1
if (i < N) R [i] [j] = (A [i] [j] + B [i] [j] /2
else R [i] [j] = B [i][j])
```

where L = No. of cepstral coefficients/frame.

The pictorial representation of the above algorithm for the case, N2> N1, is shown in Fig.7.

Online Updating Technique

This technique enables online updating of the reference pattern with the pattern of the spoken word (that matches with the reference pattern) using the above averaging technique. This technique has increased the system accuracy by accommodating the surrounding variations. However, this unsupervised technique suffers from a drawback, i.e. when a spoken word matches with wrong word, it not only updates the wrong reference pattern but also continues to match with the wrong word. The block diagram of the SRS incorporating the online updating technique is shown in Fig.8.

The effect of online updating the reference pattern with the patterns of spoken word, 'Left', is shown in Fig.9. The reference pattern 'rf' and spoken word pattern 'sp' generates the reference pattern 'rf1' using averaging technique in first iteration of updating. Similarly using 'sp1', 'sp2', and 'sp3' respectively produces 'rf2', 'rf3' and 'rf4'.

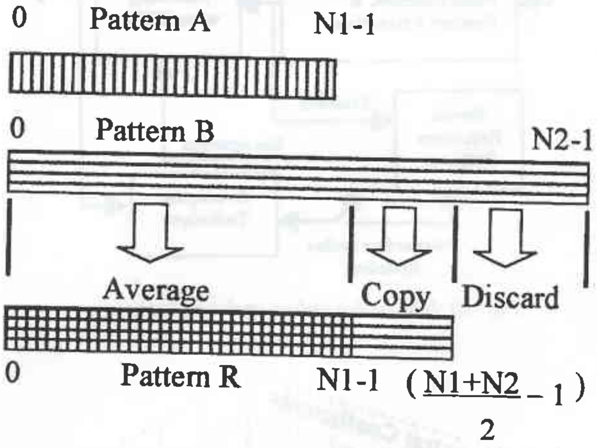

Fig. 7 Pictorial representation of the algorithm

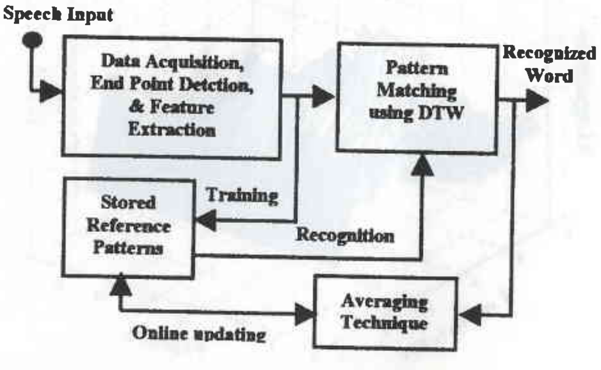
Speech Input
Data Acquisition, End Point Detction, & Feature Extraction
Pattern Matching using DTW
Recognized Word
Stored Reference Patterns
Training
Recognition
Averaging Technique
Online updating

Fig. 8 Online updating technique

Interactive Online Updating Technique

This technique of SRIJAN offers the same functionality as online updating technique except updating of the reference pattern is supervised and is done interactively to eliminate the drawback of the previous technique. The block diagram of this technique is shown in Fig.10. This technique has extremely increased the system recognition accuracy for a particular speaker and also has led to a multi-user adaptable system for different speakers. The system adapts to a new speaker just after two iterations of updating the existing reference patterns of old/previous speaker and thereby avoids the requirement of fresh training overhead.

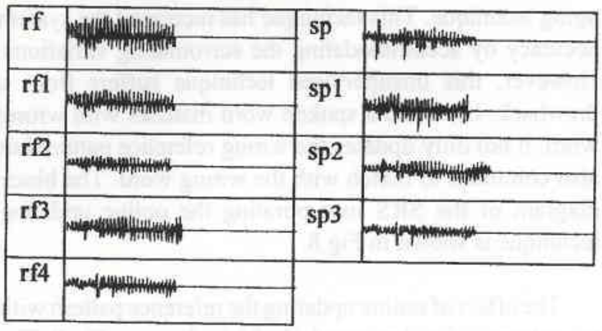

Fig. 9 Effect of online updating

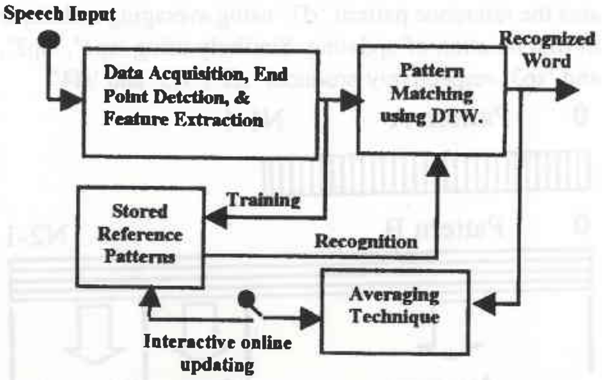
Speech Input
Data Acquisition, End Point Detction, & Feature Extraction
Pattern Matching using DTW.
Recognized Word
Training
Stored Reference Patterns
Recognition
Averaging Technique
Interactive online updating

Fig. 10 Interactive online updating technique

CMS Technique

The cepstral mean subtraction technique performs subtraction of mean of cepstral coefficients of a frame from its each cepstral coefficient. This frame wise cepstral normalization is expressed by the equation as follows :

$$z [n] = c [n] - 1/P \sum_{n=1}^{P} c [n]$$

where P is the total number of cepstral coefficients per frame, c[n] is cepstral coefficient and z[n] is the normalized cepstral coefficient. This frame wise normalization technique improves the system performance than cepstral mean normalization (CMN) technique [8], which performs normalization on pattern-by-pattern basis. The application of above averaging technique on normalized cepstral coefficients resulting from CMS technique has further enhanced the system accuracy. The plot of the cepstral coefficients and normalized cepstral coefficients is shown in Fig.11 for the spoken word 'Left'.

Results and Conclusions

The SRIJAN has been implemented on EZ-LAB board hosted on a PC. The use of cepstral coefficients, as feature vectors, has increased the system accuracy up to 91%. The optimization of end point detection procedure by changing

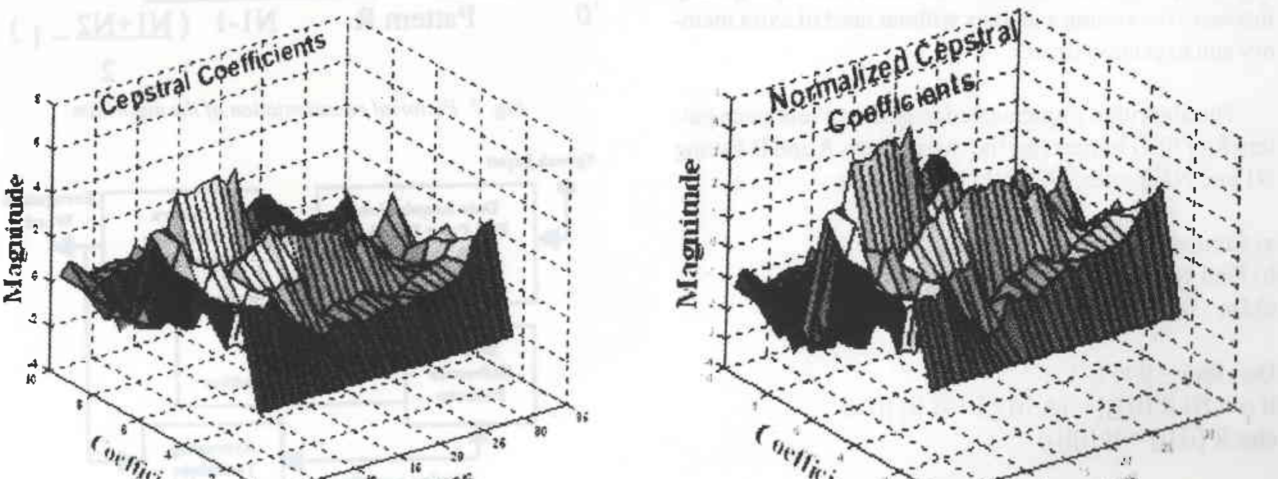

Fig. 11 Plot of cepstral and normalized cepstral coefficients for spoken word 'Left'

thresholds enables in the saving of computations and data compression.

The implementations of multiple reference patterns, averaged reference pattern, online and interactive online reference pattern updating techniques have enhanced the SRIJAN accuracy to a great extent. The use of normalized cepstral coefficients as feature vectors, resulting from CMS technique, has extremely increased the recognition accuracy. The application of averaging technique on normalized cepstral coefficients has further enhanced the SRIJAN recognition accuracy.

The implementations of these techniques have led to a robust system leading towards its use into cockpit application. The recognition accuracy scored by different technique is given in Table-1 for a set of 30 words/digits as listed in Table-2. Table-1 shows that the use of two reference patterns of a word offers better recognition accuracy than the averaging technique (using two patterns) at the expense of memory and execution time. The interactive online updating technique has not only increased the system accuracy up to 97% but also has led to adaptability of the system by multi-user with lesser training overhead. The CMS technique has enhanced the system accuracy up to 95% without need of updating the reference patterns as is done in online and interactive online updating techniques. It also does not require extra memory and execution time as required by the multiple reference patterns technique and therefore this technique has been proved superior to other techniques.

Table-1 : Accuracy Score Obtained by Different Techniques

Techniques	Feature used	Accuracy (up to)
Single pattern	Cepstral coefficients	91%
Multiple patterns (two)	Cepstral coefficients	96%
Averaged patterns (two)	Cepstral coefficients	94%
Online updating	Cepstral coefficients	92%
Interactive online updating	Cepstral coefficients	97%
Single pattern	Normalised Cepstral coefficients	93%
Averaged pattern (two)	Normalised Cepstral coefficients	95%

Table-2 : Selected Words and Digits

Sl. No.	Spoken Utterance	Sl. No.	Spoken Utterance
1	Process	16	Trim
2	Search	17	Low
3	Look	18	Yaw
4	Open	19	Change
5	Close	20	Measure
6	Single	21	Zero
7	Multi	22	One
8	Distance	23	Two
9	Remove	24	Three
10	Pitch	25	Four
11	Name	26	Five
12	Up	27	Six
13	Down	28	Seven
14	Left	29	Eight
15	Right	30	Nine

Acknowledgements

The authors take this opportunity to thank Shri MD Aravamudhan, Director, Aeronautical Development Establishment (ADE), Shri Suresh Kumar, Associate Director, ADE and Shri KG Rama Manohar, Scientist F, Head, CST Division, ADE for their inspiration and support during this research and development activity. The authors would also like to thank Mr Iqsaam-Ul-Mukhtar, Scientist 'B', all colleagues and friends who have helped and given their invaluable suggestions.

References

1. Rabiner, L.R. and Sambur, M.R., "An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Syst. Tech. Journal, Vol.54, February 1975, pp.297-315.

2. Hiroaki Sakoe and Seibi Chiba., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. Acoust. Speech, and Signal Process, Vol.ASSP-26, February 1978, pp.43-49.

3. Atal, B.S., "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic

Speaker Identification and Verification", Journal of Acoust. Soc. Amer., Vol.55, June 1974, pp.1304-1312.

4. Markel, J.D. and Gray, A.H., "Linear Prediction of Speech", Berlin : Springer-Verlag, 1976.

5. Mazel, D.S. and Hayes III, M.H., "Reflections on Levinson's Recursion", ICASSP-88, Vol.3, April 1988, pp.1632-1635.

6. Richard A. Haddad and Thomas W. Parsons., "Digital Signal Processing, Theory, Applications and Hardware", 1991, New York, Computer Science Press, Appendix-E, pp.582-587.

7. Picone, J. et.al., "Signal Modeling Techniques in Speech Recognition", proceedings of the 1993 IEEE Automatic Speech Recognition and Understanding Workshop, Vol.81, No.9, pp.1215-1247, September, 1993.

8. Suresh Balakrishnama., Final Paper on "Speech Recognition Using Mel Cepstrum, Delta Cepstrum and Delta-Delta Features", Submitted to Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, Mississippi 39762, to fulfill the requirements for ECE 89993 : Fundamentals of Speech Recognition, December, 5, 1998.

9. Martin, T., "Application of Limited Vocabulary Recognition Systems", in Rec. 1974 Symp. Speech Recognition, Dr Reddy, Ed. New York, Academic, 1975, pp.55-71.