SPEECH RECOGNITION SYSTEM USING SPEECH CODING ON EVM TMS320C541 BOARD

Eugen Lupu Petre G. Pop
Technical University of Cluj-Napoca
Communications Dept.
26-28 Baritiu str.
ROMANIA
Eugen.Lupu@com.utcluj.ro

Mircea Patras
Frosys Inc. of Cluj-Napoca
ROMANIA
mircea.patras@FROSYS.ro

Abstract - The contribution presents a speech recognition application developed on the EVM C541 board using the CCS (Code Composer Studio). The application represents the implementation of the TESPAR (Time Encoding Signal Processing and Recognition) coding method on a DSP support. The TESPAR alphabet for the coding process was obtained formerly. The information contained in the utterances is extracted by TESPAR coder and provide the TESPAR-A matrices. For the classification decision, the distances among the TESPAR-A test matrix and the TESPAR-A reference matrices are computed. The results of the experiments prove the high capabilities of the TESPAR method in the classification tasks noticed also in [1][3].

I. INTRODUCTION

TESPAR coding is a method based on the approximations to the locations of the 2TW (where W is the signal bandwidth and T the signal length) real and complex zeros, derived from an analysis of a band-limited signal under examination. Numerical descriptors of the signal waveform may be obtained via the classical 2TW samples ("Shannon numbers") derived from the analysis. The key features of the TESPAR coding in the speech-processing field are the following:

- the capability to separate and classify many signals that cannot be separated in the frequency domain

- an ability to code the time varying speech waveforms into optimum configurations for processing with Neural Networks

- the ability to deploy economically, parallel architectures for productive data fusion [3].

II. TESPAR SPEECH CODING BACKGROUND

The key in the interpretation of the TESPAR coding possibilities consists in the complex zeros concept. The band-limited signals generated by natural information sources include complex zeros that are not physically detectable. The real zeros of a function (representing the zero crossing of the function) and some complex zeros can be detected by visual inspection, but the detection of all zeros (real and complex) is not a trivial problem. To locate all complex zeros involves the numerical factorization of a 2TWth-order polynomial. A signal waveform of bandwidth W and duration T, contains 2TW zeros; usually 2TW exceeds several thousand. The numerical factorization of a 2TWth-order polynomial is computationally infeasible for real time. This fact had represented a serious impediment in the exploitation of this model. The key to exceed this deterrent and use the formal zeros-based mathematical analysis is to introduce an approximation in the complex zeros location [5].

Instead of detecting all zeros of the function the following procedure may be used:

- The waveform is segmented between successive real zeros and
- This duration information is combined with simple approximations of the wave shape between these two locations.

These approximations detect only the complex zeros that can be identified directly from the waveform.

In this transformation of signals, from time-domain in the zero-domain:

- The real zeros, in the time-domain, are identical to the locations of the real zeros in the zero-domain, and
- The complex zeros occur in conjugate pairs and these are associated with features (minima, maxima, points of inflexion etc.) that appear in the wave shape between the real zeros [4].

In this way examining the features of the wave, shape between its successive real zeros may identify an important subset of complex zeros.

In the simplest implementation of the TESPAR method [1], two descriptors are associated with every segment or epoch of the waveform.

These two descriptors are:

- The *duration (D), in number of samples,* between successive real zeros, which defines an *epoch*
- The shape *(S),* the *number of minima* between two successive real zeros.

In this simple TESPAR model implementation, not all complex zeros can be identified from the wave shape, so the approximation is limited to those zeros that can be so identified.

The TESPAR coding process is presented in fig. 1, using an alphabet (symbol table) to map the duration/shape (D/S) attributes of each epoch to a single descriptor or symbol [6].

The TESPAR symbols string may be converted into a variety of fixed-dimension matrices. For example, the S-matrix is a single dimension 1xN (N- number of symbols of the alphabet) vector, which contains the histogram of symbols that appear in the data stream (Nr. App), fig. 2. Another option is the A-matrix, which is a two dimensional NxN matrix that contains the number of times each pair of symbols appears at a "*lag*" distance of n symbols (fig. 3) [7]. The "*lag*" parameter provide the information on the short-term evolution of the analyzed waveform if its value is less than 10 or on the long-term evolution if its value is higher than 10. This bidimensional matrix assures a greater discriminatory power.

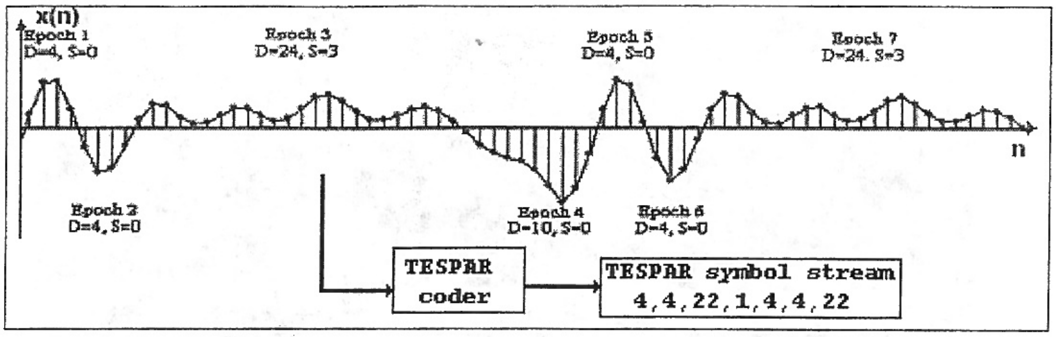


Fig. 1: The TESPAR coding process

The discriminatory power may be improved by using a matrix with three dimensions. There is also mentioned in the literature of the domain a new hybrid TESPAR DZ matrix. The main advantage of processing signals using the TESPAR method over traditional methods based on frequency descriptors is that *TESPAR matrices are fixed length structures*. These matrices are ideal to be used as fixed-sized training and interrogation vectors for the MLP neural-networks.

There are two main methods of classifying using TESPAR:

- classifying using archetypes
- classifying with neuronal networks.

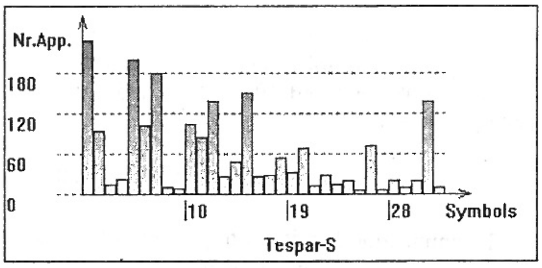


Fig. 2 : TESPAR S-matrix

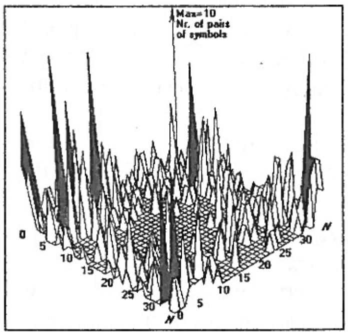


Fig. 3. TESPAR A-matrix

This paper deals with the first method that was implemented on the system. An archetype is obtained by averaging several matrices obtained from different versions of the same utterance. Such archetypes tend to outline the basic mutual characteristics and dim the particular cases that might appear in different utterances of the same word, for example.

The created archetype may be loaded in the database and then used. In the classification process, a new matrix might be created and then compared to the archetype. Many different forms of correlation can be used to achieve the classification. A threshold is required to establish whether the archetype and the new matrix are sufficiently alike; the archetype with the highest ratings is chosen after it has been compared to a threshold.

III. TESPAR ALPHABET DEFINITION

In order to define the TESPAR alphabets, two minutes of high quality speech record, sampled at 8, 11, 22 kHz with 16 bits resolution, were employed [6]. Then a (Visual C++) developed application scanned the record and for each epoch detected the descriptors: duration (D-samples) and shape (S-minima) (D/S). These pairs of descriptors represent points in the DxS plan assigned to each epoch. They are the training data set for the vector quantisation process made by the Linde-Buzo-Gray algorithm.

In fig.4. the epochs distribution is presented for the previosly mentioned two minutes of speech record, sampled at 8 kHz. One can notice that the epochs are concentrated round the origin, in the rectangle delimited area, where more than 98% of the epochs may be found and it is recommended to split only the centroid that provides the maximum distortion. The studies made on this approximation for different sampling rates are presented in table 1.

This vector quantisation process delivers the symbols-table of the TESPAR alphabet (see table1), which is used to map a TESPAR symbols for each signal waveform epoch, in the TESPAR coding process. The experiments proved that an alphabet with 29-32 symbols is sufficient for an acceptable approximation of signals in the classification application using this method. In our experiments a 32 symbols alphabet was used, with symbols between 0-31, as it can be noticed in table 2.

TABLE 1
D AND S LIMITS USED FOR DIFFERENT SAMPLING RATES

Sampling Frequency	S limit	D limit
22 kHz	5	70
11 kHz	5	40
8 kHz	5	30

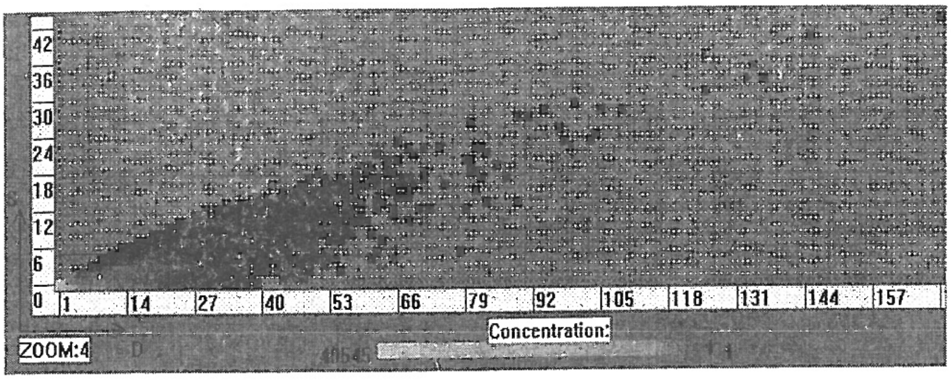


Fig. 4: Epochs distribution in the SxD plan (Fs = 8 kHz)

TABLE 2
TESPAR IMPLICIT ALPHABET FOR 8 KHZ SAMPLING RATE.

S/D	0	1	2	3	4	5
1	0	-	-	-	-	-
2	6	-	-	-	-	-
3	14	10	-	-	-	-
4	4	10	-	-	-	-
5	30	10	10	-	-	-
6	11	10	25	-	-	-
7	11	9	25	25	-	-
8	17	9	25	25	-	-
9	1	5	25	25	21	-
10	1	5	12	21	21	-
11	13	19	12	21	21	21
12	13	19	27	21	21	26
13	16	15	27	21	21	26
14	16	15	18	21	26	26
…………………………………						
30	20	20	20	20	3	3

IV. SPEECH RECOGNITION SYSTEM OVERVIEW

Fig.5 shows the block diagram of the application, this being shared between PC and the DSP board. In order to run the application we have to load the VoiceR program on the DSP board and to run the program A-Matrix Tools (on PC) if some reference matrices are to be loaded by the DSP program.

The applications on the DSP board are built using the CCS® environment that allows the fast application development using its own resources: C compiler, linker, debugger, simulator, RTDX and DSP/BIOS components [8].

A CCS environment-working window can be observed in fig. 6. The main facilities of the VoiceR application can be remark in the flow diagram, fig. 7.

The A-Matrix Tools program allows the TESPAR A-matrices transfer between the host PC and the EVM C541 board and offer facilities to extend the VoiceR program operation. The tasks of this program are the following:

- TESPAR A-matrices collections transfer between host PC and EVM board
- TESPAR A-matrix transfer between host EVM board and host PC
- save matrix/matrices collections to host PC hard disk in text files
- load matrix/matrices collections from host PC hard disk to DSP board
- 3D TESPAR-A diagrams visualization
- City-block distance computation between two TESPAR-A matrices
- Archetype generation for TESPAR-A matrices collections.

For the polling communication between the host PC and the DSP board the following ports are employed; for data the port 800h (PC) and 10H (DSP) and for control 808h (PC) and 14h (DSP) [7][8].

V. EXPERIMENTS AND RESULTS

The application facilitates to perform different "on-line" speech recognition experiments. In the classification process, the distance calculation between the A-matrices archetypes and the test matrices or parallel MLP neural networks may be employed. In this paper, the experiments focus on the use of distance calculation between the A-matrices archetypes and test matrices in the classification task.

Two types of experiments were been made, one using the ten digits as utterances and the other using different commands (left, right, up…). In the first experiment, seven speakers were enrolled for the system training and 10 speakers for the test. Each of the enrolled speakers uttered three times every digit for the training and ten times for the recognition. The results of this experiment are presented in fig.8. In this case an average recognition rate of 92% was obtained that we find to be good in the condition of using for test also utterances of not enrolled speakers.

For the other type of experiments, we used 10 commands words. In the "Test2", experiment the training was made by using the three utterances from every speaker to build its own archetype for every command. For the "Test3" experiment the archetype were been built by using 3 utterances from two enrolled speakers.

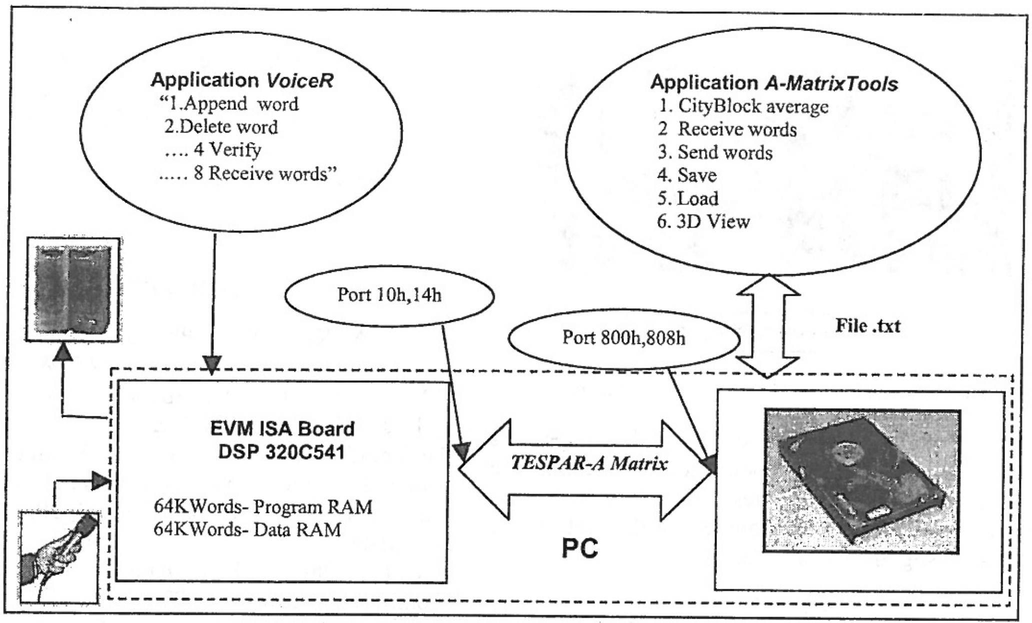
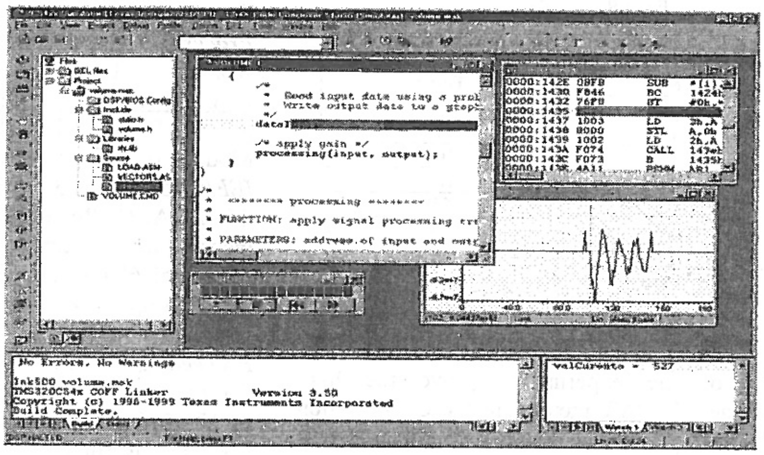


Fig.5. The block diagram of the application



```
    {
        /*
         *  Read input data using a prob
         *  Write output data to a stopP
         */
        dataI

        /* apply gain */
        processing(input, output);
    }
}

/*
 *  ******* processing *******
 *
 *  FUNCTION: apply signal processing tra
 *
 *  PARAMETERS: address of input and outp
```

```
0000:142E 08F8    SUB   *(1)
0000:1430 F846    BC    1424h
0000:1432 76F8    ST    #0h,*
0000:1435
0000:1437 1003    LD    3h,A
0000:1438 8000    STL   A,0h
0000:1439 1002    LD    2h,A
0000:143A F074    CALL  147eh
0000:143C F073    B     1435h
0000:143E 4A11    PSHM  AR1
```

```
No Errors, No Warnings

lnk500 volume.mak
TMS320C54x COFF Linker              Version 3.50
Copyright (c) 1996-1999 Texas Instruments Incorporated
Build Complete.
```

```
valCurenta =   527
```

Fig.6. CCS environment working window

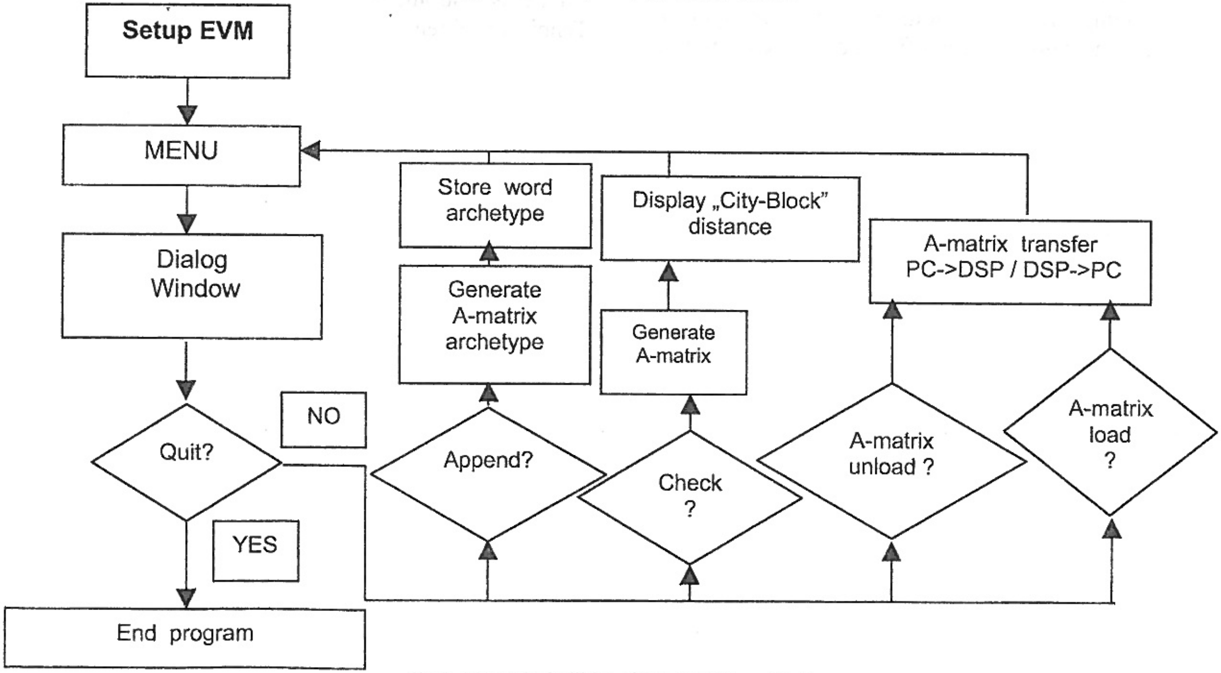


Fig.7. The main facilities of the VoiceR application

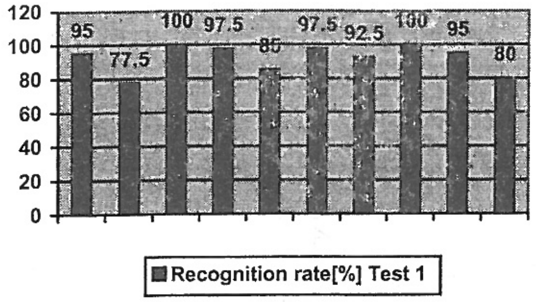


Fig.8. Digits recognition rate for the "Test 1" experiment

In every experiment to test the system all the speakers has uttered 10 times every command. The results are presented in table 3. For the "test2" experiment, a 98% average recognition rate was provided by the system and for the "test 3" slightly lower 97%.

TABLE 3
RECOGNITION RATE FOR THE "TEST 2" AND "TEST 3" EXPERIMENTS

Word	Recognition Rate [%] Test 2	Recognition Rate [%] Test 3
up	100	100
down	95	100
left	100	100
right	100	100
enter	95	95
cancel	100	95
abort	100	85
ok	100	100
back	90	95
forward	100	100

The results of the experiments prove the high capabilities of the TESPAR method in the classification tasks, noticed also in [1][3]. The results generally are better than 90% and the DSP resources are not very highly used. In order to improve the recognition rate the employment of MLP neural networks for classification will be employed.

To validate the system more experiment are to be made using much amounts of utterances and different speakers ar advisable to test the system. In addition, the effects of other signal processing algorithms applied before the coding process are to be studied.

VI. REFERENCES

[1] R. A. King, T. C. Phipps. "Shannon, TESPAR and Approximation Strategies", *ICSPAT 98,* Vol. 2, pp. 1204-1212, Toronto, Canada, September 1998.

[2] J. C. R. Licklidder, I. Pollack, "Effects of Differentiation, Integration, and Infinite Peak Clipping Upon The Intelligibility Of Speech*", Journal Of The Acoustical Society Of America*, vol. 20, no. 1, pp. 42-51, Jan. 1948.

[3] T.C Phipps, R.A. King. "A Low-Power, Low-Complexity, Low-Cost TESPAR-based Architecture for the Real-time Classification of Speech and other Band-limited Signals" *International Conference on Signal Processing Applications and Technology (ICSPAT) at DSP World,* Dallas, Texas, October 2000, www.dspworld.com/icspat/spchrec.htm.

[4] H. B. Voelcker, "Toward A Unified Theory of Modulation Part 1: Phase-Envelope Relationships*", Proc. IEEE*, vol. 54, no. 3, pp 340-353, March 1966.

[5] A. A. G. Requicha "The zeros of entire functions, theory and engineering applications" *Proceedings of the IEEE*, vol. 68 no. 3, pp. 308-328, March 1980.

[6] E. Lupu Z. Feher P.G. Pop "On the speaker verification using the TESPAR coding method", *IEEE Proceedings of International Symposium on "Signals, Circuits and Systems*", Iaşi, Romania, 10-11 July 2003, pp.173-176, ISBN 0-7803-7979-9

[7] Texas Instruments - TMS320C54x DSP Reference Set

[8] Texas Instruments – TMS320C54x Evaluation Module Tehnical Reference