Buletinul Ştiinţific al Universităţii "Politehnica" din Timişoara

Seria ELECTRONICĂ şi TELECOMUNICAŢII
TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Tom 49(63), Fascicola 1, 2004

Application for TESPAR coding study and speaker recognition experiments

Eugen LUPU[1] Vasile V. MOCA[1] Petre G. POP[1]

Abstract - TESPAR coding (Time Encoding Signal Processing and Recognition) represents an effectiveness alternative to the other common methods (Dynamic Time Warping, Vector Quantization, Hidden Markov Models, etc.) used for speech/speaker recognition. The important advantage of this method is the time processing of signal with a decrease of two orders of magnitude of the computational requirements.
This work presents an application for TESPAR coding study allowing speaker recognition experiments using parallel neural networks architecture.
Keywords: TESPAR coding, speaker recognition, parallel neural networks, archetypes, TESPAR-A matrix.

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 can not 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 [1][3].

II. TESPAR SPEECH CODING BACKGROUND

The key in the interpretation of the TESPAR coding possibilities consists in the complex zeros concept. T'e ban'-'mite' signa's genera'ed by '-t'-l information sources include complex zeros that are '-' all 'hy'i'ally de'e'-'bl'. Th' re'l ze'os 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.

Locating all complex zeros involves the numerical factorization of a $2TW^{th}$-order polynomial. A signal waveform of bandwidth W and duration T, contains 2TW zeros; usually 2TW exceeds several thousand. The numerical factorization of a $2TW^{th}$-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 [2] [4][5].

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

- the waveform is segmented between successive real zeros (this defines an epoch);
- this duration information is combined with simple approximations of the wave shape between these two locations.

In the simplest implementation of the TESPAR method [1][3][8], 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;
- the shape *(S),* the *number of minima* between two successive real zeros.

The TESPAR coding process is presented in fig. 1, using an alphabet (table1) to map the duration/shape (D/S) attributes of each epoch to a single descriptor or s mbol [6 8].

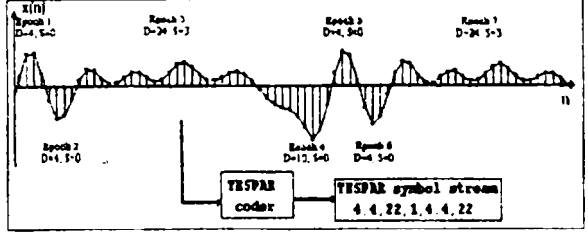

Fig 1 The TESPAR coding process

[1] Technical University of Cluj-Napoca, Faculty of Electronics & Telecommunications, Comm Dept., e-mail Eugenl upu@com.utcluj.ro

In order to define the TESPAR alphabet a high quality speech record is scanned by the developed application and for each epoch detects the descriptors: duration (D-samples) and shape (S-number of minima). These pairs of descriptors represent points in the DxS plan assigned to each epoch. They are the training data set for the vector quantization process made by the Linde-Buzo-Gray algorithm.

This vector quantization process delivers the symbols-table of the TESPAR alphabet, which is used to map a TESPAR s mbols for each si nal waveform e och, 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 applications using this method. In our experiments a 32 symbols alphabet was used, with symbols between 0-31, as it can be noticed in table 1 [7][8].

Table 1

S/D	0	1	2	3	4	5
-	-	-	-	-	-	-
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

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 [7][8]. Another option is the A-matrix, which is a two dimensional NxN matrix that contains the number of apparition of each pair of symbols at a "*lag*" distance of n symbols (fig. 3) [1][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.

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 band-limited signals using TESPAR:

- classifying using archetypes;
- classifying with neuronal networks [1][3][7].

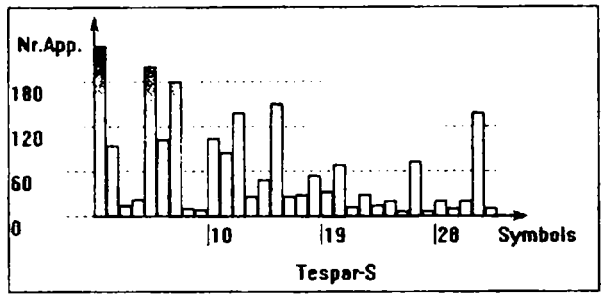

Fig 2. TESPAR S-matrix

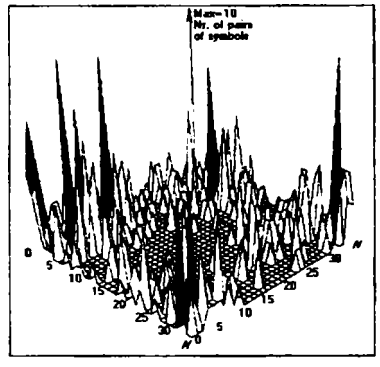

Fig 3. TESPAR A-matrix

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 stored 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. THE APPLICATION OVERVIEW

The application for TESPAR coding study and speaker recognition experiments was realized using Visual C++ 6.0 environment, which allows fast Windows applications developing with all interface facilities. The applications main functions are:

- wav file manipulation;
- TESPAR alphabet generation;

- TESPAR-S, TESPAR-A matrices and the archetypes generation;
- training of parallel MLP neural networks with multiple hidden layers;
- the classification task using archetypes and distances or neural networks;
- to perform speaker recognition experiments on-line or off-line for a large speech data base;
- to save the experiments results in MS-Excel compatible format.

"Fig.4." shows some working windows of the application for the speech files manipulation and processing.

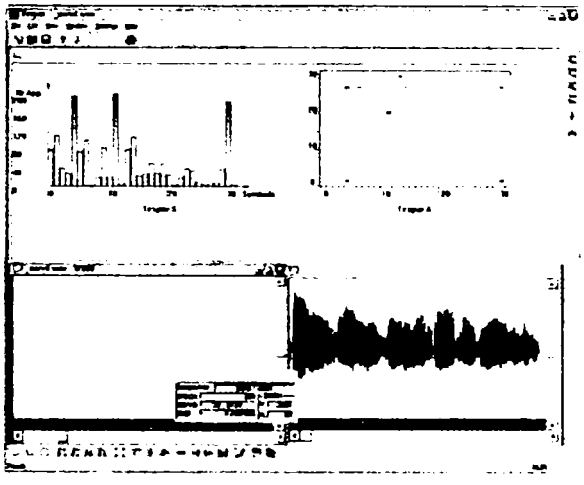

Fig.4 Working windows of the application

To achieve the above goals the environment offers the following facilities:

- record, play, view, load and save operation with wav files;
- wav file editing (cut, silence/speech manual/ automatic separation), speech signal filtering (low-pass) to smooth the wave;
- TESPAR alphabet generation, saving and loading. The facility of visualization step by step the alphabet generation process to get statistical on the epochs distribution in the SxD plan;
- TESPAR-S and TESPAR-A matrices and archetypes generation and visualization;
- the easy manipulation of the speech database, adding or removing speech records or using a variable number of records for training for each enrolled speaker;
- building, high speed automatic training, testing, visualization, saving and loading of the MLP neural networks (with many hidden layers);
- selectable method of classification: archetype distances or MLP classification;
- the system flexibility in fine tuning of the experiments conditions;
- experiments can be performed "on-line" using the record speech facility or by using a record from the speech data base;
- automatic classification of a large number of speech files and results saving;
- export of the results files in a common format;
- help guide to use the environment [7].

IV. SPEAKER RECOGNITION EXPERIMENTS AND RESULTS

The environment facilitates to perform different experiments using a speech data base or "on-line". In the classification process, the distance calculation between the archetypes and test matrices can be employed or parallel MLP neural networks. Some results of the experiments using the distance calculation between the archetypes and test matrices were been presented in [7]. In this paper the experiments focus on the use of MLP neural networks and the TESPAR-A matrices in the classification tasks because they offer better results. The experiments used our speech database of 50 speakers (40 male + 10 female); each speaker has uttered 5 times the same voiced sentence [6][7]. It must be specified that the speech data base utterances were recorded in a professional environment [6][7]. Every utterance was coded with the TESPAR alphabet and the TESPAR-A matrices were derived from each sentence. In order to test the possibilities of the recognition system, which employ MLP, the following conditions were employed:

- the implicit alphabets for the different sampling rates;
- the system is closed, with 33 enrolled speakers, while each of them had uttered 5 times the same voiced sentence; three of them were used for the training process;
- all the utterances were used for the test;
- 17 parallel MLP neural networks were used for the recognition task (with different hidden layers configurations).

The state of the neural networks can be seen in a window after each classification that uses MLP, fig.5. The results provided by the experiments for different sampling rate of the speech signal and maximal error allowed in the training process of the MLP are presented in the table 2 for *closed systems*.

For the *opened systems* the best results are presented in table 3. for FAR (False Acceptation Rate) and FRR (False Rejection Rate) and a result that is a good balance of the both.

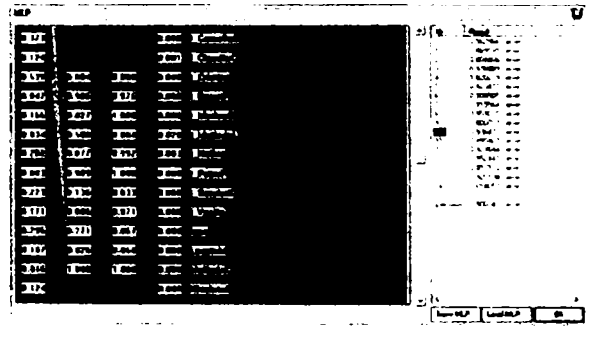

Fig.5 The internal state of the MLP and a taken decision

Table 2

F_s	Max Err allowed for training	Best results [%] (1 hidden layer with 10 neurons)	Best results [%] (2 hidden layers with 2X10 neurons)
22kHz	0.2	99.39%	100.00%
	0.1	99.39%	100.00%
	0.07	99.39%	100.00%
	0.05	99.39%	100.00%
	0.03	99.39%	100.00%
11kHz	0.2	99.39%	100.00%
	0.1	99.39%	99.39%
	0.07	99.39%	99.39%
	0.05	99.39%	99.39%
	0.03	99.39%	99.39%
8kHz	0.2	99.39%	99.39%
	0.1	99.39%	99.39%
	0.07	99.39%	99.39%
	0.05	99.39%	99.39%
	0.03	99.39%	99.39%

Table 3

F_s	Hidden Layer(s) neurons	FAR	FRR	Optimized
22kHz	10	3.03%	2.42%	FAR
	10	3.64%	1.82%	FRR
	10	3.64%	1.82%	FAR&FRR
	10 – 10	0.61%	7.27%	FAR
	10 – 10	1.82%	1.21%	FRR
	10 – 10	1.82%	1.21%	FAR&FRR
11kHz	10	0.00%	9.70%	FAR
	10	4.24%	0.61%	FRR
	10	4.24%	0.61%	FAR&FRR
	10 – 10	0.00%	8.48%	FAR
	10 – 10	1.21%	1.82%	FRR
	10 – 10	1.21%	1.82%	FAR&FRR
8kHz	10	1.82%	7.88%	FAR
	10	2.42%	1.82%	FRR
	10	2.42%	1.82%	FAR&FRR
	10 – 10	0.00%	10.91%	FAR
	10 – 10	1.82%	1.82%	FRR
	10 – 10	1.82%	1.82%	FAR&FRR

V. CONCLUSIONS

The results of the experiments prove the high capabilities of the TESPAR method in the classification tasks noticed also in [1][3].

For the *closed system* the results are above 99%, however an improved performance of 100% can be noticed for MLP with two hidden layers. The sampling frequency and the training rate do not seem to greatly affect the performances of the system.

If for the closed system the performances are very close, for the *opened system* an extra layer in the MLP decreases the FAR with 1-3 % and in some cases increases the FRR with about 1%, the overall result is a performance increase.

The effect of the sampling frequency on the system performance is not very important (3-4%); generally, the best results are obtained for 8 and 11 kHz sampling rate.

The environment allows much flexibility in performing the experiments, in:

- files manipulation;
- alphabet dimension determination;
- the MLP manual parameters setting in order to control the training process and to assure a fast convergence;
- maximum allowed error selection for the training;
- the results saving in MS-Excel format.

The application may be improved by using the effects of other signal processing algorithms, applied before the coding. In order to improve the recognition rate the employment of more parallel MLP neural networks with an increased number of layers can be used. However caution must be present when increasing the numbers of hidden neurons because the neural network looses the abstraction in favor of memory capacity. To validate the system more experiment are to be made using much amounts of utterances, stored on a long period of time and different speakers are advisable to test the system.

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] Lupu E., Moca V.V., Pop G.P., "Environment for Speaker Recognition Using Speech Coding", *Proceedings of the International Conference COMMUNICATIONS 2004*, June 03-05, 2004, Bucharest, pp. 199-204.

[8] Lupu, E., Pop, G. Petre *Prelucrarea numerică a semnalului vocal. Elemente de analiză şi recunoaştere*. Ed. Risoprint 2004