SPEAKER VERIFICATION RATE STUDY USING THE TESPAR CODING METHOD

Eugen LUPU[*] **Zoltán FEHÉR**[*] **Petre G. POP*** ,
[*]Technical University of Cluj-Napoca
Faculty of Electronics &Telecommnications - Communications Dept.
3400 Cluj-Napoca ROMANIA
Tel: +40-64-406043 Email : EugenLupu@com.utcluj.ro

Abstract. The paper presents a study on the speaker verification rate, using the TESPAR (Time Encoding Signal Processing and Recognition) coding method, when the speech signal is sampled at different rates. The effect of filtering was studied, as well. The TESPAR method is a processing and recognition method in the time domain, proposed by [1]. The key problem is to define the TESPAR alphabet used for the TESPAR coding process. In this paper is proposed an approach to generate this alphabet using the Kohonen Neural Networks in a vector quantisation process. For the recognition process more parallel Multi Layer Perceptron (MLP) neural network were used. As inputs for training/test vectors of the MLP-NN the TESPAR-S matrices were employed.

INTRODUCTION

TESPAR coding is a method based on the approximations to the locations of the 2TW 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 ("Shanon 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 are indistinguishable 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.

Probably Shannon's sampling theorem has affected mostly the human understanding in the communication field :

"If a function f(t) contains no frequencies higher than W cps, it is completely determined by giving its ordinates at a series of points spaced 1/2Wseconds apart."

These model involves detecting the ordinates of a waveform at a series of points equally spaced 1/2W second apart. From this theorem, presented in figure 1, a variety of mainly linear transforms (e.g. Fourier, LPC, Wavelet or Walsh) has been developed for describing and classifying key features of the sampled data set [1].

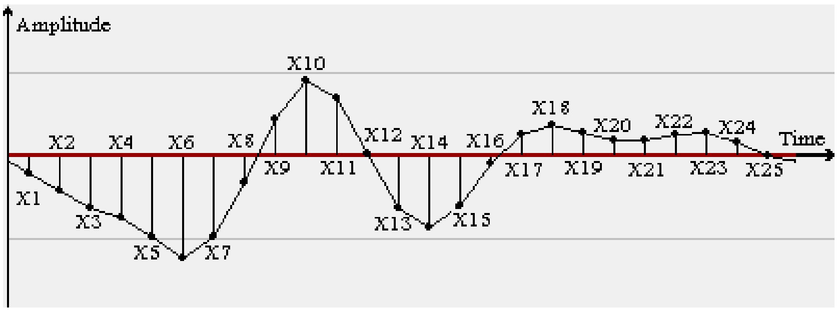

Figure 1 Regular sampling

This coding strategies involve the following requirements :

a) the use of amplitude descriptors
b) the use of regular sampling
c) an approximation domain dependent upon the numbers of bits per sample

TESPAR CODING

Another work [2] investigates the effects of *amplitude clipping* on the intelligibility of the speech waveform. For the so-called *infinite clipping* format, whereby *all amplitude information* was removed from the waveform results a binary transformation that preserved only the zero-crossing points of the original signal, (figure 2). If the speech waveform was differentiated, prior to infinite clipping the results show that the mean random-word intelligibility scores of 97.9% were achieved [1]. These observations show that the main proportion of the information of interest in the speech waveform (i.e. its intelligibility) is contained only in its zero-crossings. These results call into question the status of the three requirements a), b) and c) previously stated. The information kept by the infinitely-clipped samples represent the *duration* of the intervals between the zero-crossings of the waveform, figure.2. These *zero-based* durations are thus *signal-derived* and not generated at regular 1/2W second time intervals. For speech waveforms these samples will be *irregularly spaced*. These observations represents the background for the development of TESPAR (Time Encoded Signal Processing And Recognition.) method.

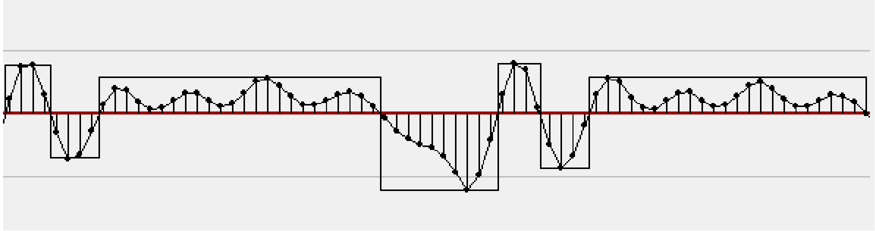

Figure 2 Infinite-clipping keeps only the zero-crossing information

There should be noticed that the infinite clipping data set represents an approximation to the original waveform, which preserves the intelligibility of band-limited speech. 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 (represented the zero crossing) 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.

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.) in the wave shape that appear 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, in order to generate the TESPAR symbol alphabet.

These two descriptors are:

- the duration between successive real zeros (in number of samples), which defines an e*poch*
- the shape 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 band-limitation of the signal imposes significant restrictions upon the maximum and minimum duration of any epoch, and also upon the maximum number of significant waveform extrema points that each epoch may contain.

The longest epoch may have a duration approximately equal to half the period of the lowest frequency component allowed by band-limiting; the shortest epoch may have a duration approximately equal to half the period of the highest frequency component allowed within the band of signal. Also, short epochs have no or few features, whilst long epochs may contain few or many features. For the simplest implementation, each epoch may be classified in terms of its *duration (D) - number of samples* and the *number of minima (S),* that it contains.

The TESPAR coding process is presented in Figure3, using an alphabet (symbol table) to map the duration/shape (D/S) attributes of each epoch to a single descriptor or symbol.

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. Another option is the A-matrix, which is a two dimensional NxN matrix that contains the number of times each pair of symbols appears in a lag of n symbols.

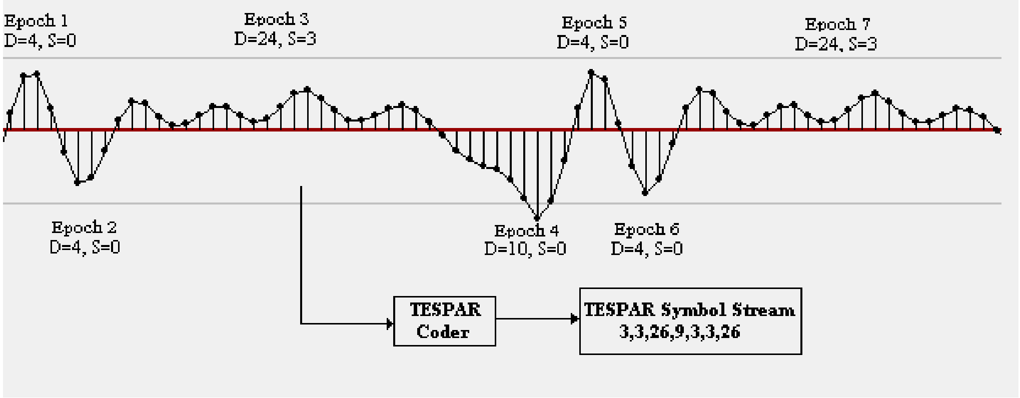

Figure 3 The TESPAR coding process

TESPAR ALPHABET DEFINITION

In order to define the TESPAR alphabets, 2 minutes of high quality speech record, sampled at 8, 11, 22KHz with 16 bits resolution, were employed. For the speech filtering a band pass filter (0.05-4 KHz) was used.

Then a Borland C++ Builder 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 a Kohonen neural network. The results for the TESPAR alphabet issued by the Kohonen Neural Network are shown in figure 4.

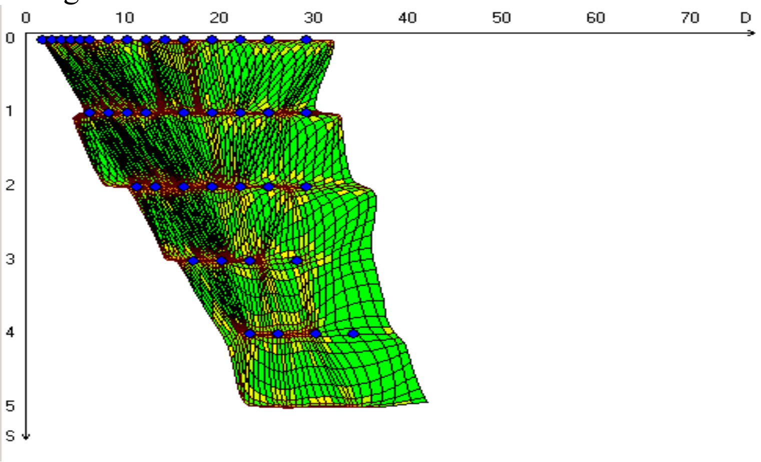

Figure 4 The Kohonen map provided for 11 kHz sampling rate, filtered speech

This vector quantisation process delivers the symbols-table of the TESPAR alphabet (see table 1), which is used to map a TESPAR symbols for each signal waveform epoch, in the TESPAR coding process. The TESPAR coding process provides a symbol string that may be ordered in a TESPAR matrix. These matrices are ideal to be used as fixed-sized training and interrogation vectors for the MLP neural-networks.

	s=0	s=1	s=2	s=3	s=4	s=5
d=1	1	-	-	-	-	-
d=2	1	-	-	-	-	-
d=3	2	2	-	-	-	-
d=4	3	3	-	-	-	-
d=5	4	5	5	-	-	-
d=6	6	5	5	-	-	-
d=7	7	5	5	5	-	-
d=8	7	5	5	5	-	-
d=9	8	9	9	9	9	-
d=10	8	9	10	10	10	-
d=11	11	12	10	10	10	10
d=12	11	12	10	10	10	10
d=13	13	14	10	10	10	10
d=14	13	14	15	15	15	15
....						
d=38	39	39	39	39	39	39

Table 1 TESPAR alphabet for 11 kHz sampling rate, filtered speech

SPEAKER VERIFICATION EXPERIMENTS

A significant advantage of representing the time-varying speech signal with TESPAR matrices is the fact that TESPAR matrices have the same size, regardless the length of the utterance.

TESPAR-S matrices (vectors) were employed for the performed experiments.

A database of 30 speakers (20 male + 10 female) was used; each speaker has uttered 5 times the same voiced sentence (in Romanian) : *"Aoleu, lâna are molii"* .

Every utterance was coded with the TESPAR alphabet and the TESPAR-S matrices were derived from each sentence.

Figure 5 shows two TESPAR-S matrices resulted from the same sentence uttered by a male speaker. The high degree of similarity between them can be easily noticed.

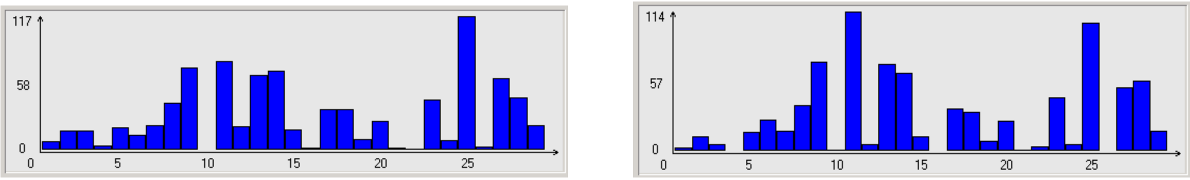

Figure 5 TESPAR-S matrices from the same speaker

In order to improve the recognition rate more parallel neural-networks with the same structure were used for the training and classification process.

In the experiments 15 parallel MLP neural networks were used for the verification task. For each neural-network the following architecture was employed : X - neurons for the input layer (where X is the TESPAR-S matrices length), 35 for the hidden layer and 21 for the output layer (20 for each enrolled speaker and 1 for impostors).

Three TESPAR-S matrices from every speaker or impostor enrolled in the experiment were employed for the training process and the others for the verification task, table 2.

Utterances for	Enrolled speakers	Impostors
Training	60	20
Testing	40	100

Table 2. Utterances used for the experiments

The experiment consists of 40 admissible attempts and 100 impostor attempts. All the speakers enrolled in the experiment employed a common password the voiced utterance presented above.

These experiments yield the results shown in table 3.

F_s/filter	**Nr. of symbols**	**F_A**	**F_R**
22KHz/ unfiltered speech	51	0.9%	2.77%
22KKHz/ filtered speech	45	0.49%	0%
11KHz/unfiltered speech	43	0.49%	2.77%
11KHz/ filtered speech	39	0.49%	0%
8KHz/ unfiltered speech	35	-	-
8KHz/ filtered speech	33	-	-

F_A –False Acceptance F_R- False Rejection

Table 3 Results of the experiments

CONCLUSIONS

The acceptance decision was taken by employing the majority rule, applied for the output of the 15 parallel MLP neural networks.

The results of the experiments prove the high capabilities of the TESPAR method in the verification tasks noticed also in [1]. The decrease of the sampling frequency lead to the decreasing of the dimension of the alphabet. The same effect can be noticed when the speech signal is filtered.

It can be estimated that the best results were provided for the sampling rate of 11KHz when the speech signal was filtered. When a frequency of 8KHz was used the training of the Neural-Networks was not achieved.

Future study is to be performed on the effect of filtering on the alphabet, for different band pass, lower than 4KHz. Also, the effects of other signal processing algorithms applied before the coding process are to be studied. The use of the TESPAR-A matrices will be experimented, as well.

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

[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] J. Holbeche, R. D. Hughes and R. A. King, "Time Encoded Speech (TES) descriptors as a symbol feature set for voice recognition systems", *IEE Int. Conf. Speech Input/Output; Techniques and Applications*, pp. 310-315, March 1986.