Journal of ELECTRICAL ENGINEERING, VOL. 55, NO. 5-6, 2004, 131–137

SPEECH SIGNAL DETECTION IN A NOISY ENVIRONMENT USING NEURAL NETWORKS AND CEPSTRAL MATRICES

Juraj Kačur [*] — **Gregor Rozinaj** [*] — † **Sergio Herrera-Garcia** [**]

In this article a new flexible speech detection method comprising two relatively modern approaches like artificial neural networks (ANN) and cepstral matrices is presented. Cepstral matrices obtained via linear prediction coefficients were chosen as the eligible speech features. This technique is known to provide reliable log spectrum estimation at a low cost. Furthermore, both spectral and time characteristics can be efficiently, which is an essential aim here. Several WSS noises and different SNR settings were tested. In the range of 3 to 13 dB the ANN approach remarkably outperformed the energy and zero crossing method and improved the accuracy of the other algorithm based on cepstral matrices as well.

K e y w o r d s: cepstral matrices, neural networks, MLP, speech detection, CLPC vectors, WSS noises

1 INTRODUCTION

Although speech detection algorithms usually do not exist as stand alone systems, they can be found in many different applications. These algorithms undertake particular tasks that are vital in many speech processing applications. Even though a lot of various detection algorithms are known, it is still worth designing new detection algorithms that are more efficient. Unfortunately, there is yet no universal detection algorithm working reliably in all possible noises and settings.

The task of voice activity detection consists in labeling the end points of words if uttered isolated or marking active segments of continuous speech signals. Apparently, finding word boundaries in continuous speech is much more difficult and these algorithms form a different group of systems that we are focusing on in this paper. Another problem is the speech itself containing high-energy vowels of various lengths as well as unvoiced consonants of low energy and higher frequency components; even intervals of silence create regular parts of speech signals. The situation dramatically deteriorates if there is some background noise present. Noises can have various characteristics that make their separation from speech sometime almost impossible. This results in several groups of detection algorithms, each for particular environment. Even articulation mistakes like lip clicking and exhalation are in some applications also regarded as noise.

The driving motivations behind this article are growing application areas like: speech compression and transmission, speech recognition, speech enhancement, medicine, *etc.*

2 CEPSTRAL MATRICES AND CLPC VECTORS

It is known that spectral envelopes and their low frequency variations over the time are important for speech recognition. Articles [2] and [3] describing the basic construction and properties of cepstral matrices show their effective usage in recognition systems. Based on those achievements we decided to use them in our detection system, since both recognition and detection are very similar in their speech analysis part. Decisive advantage of cepstral matrices is their simple static and dynamic representation of speech features. Static characteristics refer to the shape of signal spectrum whereas dynamic characteristics reflect transition processes of phonemes within words.

The classical construction of cepstral matrices usually involves the following steps. Speech signals are divided into consecutive segments that are usually overlapped by a certain part of their length. Each segment is then windowed and transformed into the frequency domain by the DFT. Real cepstra vectors are calculated applying both logarithm to the spectral magnitude and the inverse DFT. Strictly speaking the result would be only an aliased version of a real cepstrum. Eliminating this deficiency, the method of zero padding can be used. Finally, the DFT is employed to the successive blocks of real cepstra vectors in order to assess their time dynamics. Sometime a DCT is used instead of the DFT to evaluate these time variations, and thus avoiding complex arithmetic. Log magnitude spectra vectors are derived as follows in (1):

$$X(f,t) = \log\Big|\sum_{n=0}^{N-1} x(n+Kt)e^{\frac{-j2\pi nf}{N}}\Big|\,, \qquad 0 \le n \le N-1\,, \quad 0 \le f \le N-1\,. \tag{1}$$

Here N is the number of samples in the t-th speech segment and K is the number of samples separating neighbouring segments. Then elements of the l-th cepstral ma-

[*] Slovak University of Technology, Faculty of Electrical Engineering and Information Technology, Department of Telecommunications, Ilkovičova 3, Bratislava, Slovakia, e-mails: kacur@aladin.elf.stuba.sk, gregor@ktl.elf.stuba.sk

[**] CITEDI-IPN 2498 Roll Dr. #757, Otay Mesa, San Diego, California,92154, USA, e-mail: sherrera@citedi.mx

ISSN 1335-3632

trix can be derived according to (2).

$$c(n,m)_l = \frac{1}{MN}\sum_{t=l}^{M-1+l}\sum_{f=0}^{N-1} X(f,t)e^{\frac{j2\pi tm}{M}}e^{\frac{j2\pi fn}{N}}, \quad 0 \le n \le N-1, \quad 0 \le m \le M-1. \tag{2}$$

Here $c(n,m)_l$ is the element of the l-th cepstral matrix, M indicates the number of real cepstra vectors in the cepstral matrix, and N is the size of a single real cepstra vector. Coefficients $c(n,m)$ with lower indexes of n represent the log spectral envelope and those with higher indexes reflect excitation characteristics. M axis provides the time variation for existing cepstral coefficient. Those coefficients having lower indexes of m stand for global variations and those with higher indexes represent short-time changes. Thus if only global variations of the log spectral envelope are needed other elements in the matrix can be neglected.

In our work we used a modified cepstral matrix construction via the linear prediction coefficients (LPC). Great advantage of LP coefficients is their good ability to represent the spectral envelope, especially when their number is properly adjusted. Usually a good choice for speech is from 8 to 12 coefficients, but the number may vary according to the sampling frequency, location as well as the shapes of formant frequencies. An insufficient number of LP coefficients would not be able to express it accurately and their redundant number would reflect local spectral properties, too. It can be shown, that with an arbitrary accuracy the LP coefficients can approximate any zeros of a rational transfer function, except those lying on or outside the unit circle in the "Z" plane, according to equation (3):

$$1 - r^* z^{-1} = \frac{1}{\sum_{k=1}^{\infty} r^k z^{-k}} \approx \frac{1}{1+\sum_{k=1}^{M} r^k z^{-k}}, \quad |r| < 1. \tag{3}$$

Where r is the zero lying inside the unit circle and M is the number of expansion terms for a given precision. If we are interested only in the magnitude characteristic, which is the case here, then any zero lying outside the unit circle can be replaced by its reciprocal zero lying within the range of the unit circle. To obtain LP coefficients efficiently, the autocorrelation approach is used together with the Durbin method. LPC, referred here as a_i are estimated by neglecting excitation term $u(n)$ in equation (4), valid for the linear model of speech production.

$$x(n) = \sum_{i=1}^{P} x(n-i)a_i + Gu(n). \tag{4}$$

P represents the number of LPC, $u(n)$ is the unitary excitation of the speech model and G stands for the gain. It can be shown [1] that a correct estimation is only made if the excitation is a white noise, provided the number of LPC is properly chosen. Although LP coefficients have theoretical ability to represent spectral magnitude of any rational transfer function, practical results are not as brilliant. LPC modelling usually fails in the representation of zeros that is caused by methods commonly used for LPC calculation. MSE minimization of the prediction error leads to the emphasis of spectral peaks while shapes of spectral valleys are neglected, because of their small energy. Details about LPC computation can be found in *eg* [1, 6].

It is advantageous to transform these LPC to cepstral LP coefficients (CLPC). By doing so, we can easily express the log magnitude spectra of the speech, use various filtering methods to reduce the LPC deficiency in zeros modelling, etc. CLPC can be calculated by applying logarithm to the transfer function expressed by LPC (5). If all poles are inside the unit circle, which is true when using autocorrelation method, the Taylor's series expansion can be employed as shown in (5).

$$\ln\big(H(z)\big) = \ln\Big(\frac{G}{1+\sum_{i=1}^{P} a_i z^{-i}}\Big) = \ln\Big(\frac{G}{\sum_{i=0}^{P} a_i z^{-i}}\Big) = \sum_{k=0}^{\infty} c(k) z^{-k}. \tag{5}$$

Here $c(k)$ is the k-th CLPC. It can be shown that a recursive formula transforming LPC to CLPC can be derived from (5) in the form of (6)

$$c_0 = \log(G), \qquad c_1 = -a_1,$$
$$c(k) = \begin{cases} -a_k - \sum_{i=1}^{k-1} \frac{i}{k} c(i) a_{k-1}, & 2 \le k \le Q, \\ -\sum_{i=1}^{Q} \frac{k-i}{k} c(k-i) a_i, & Q < k. \end{cases} \tag{6}$$

Here a_k is the k-th LP coefficient, $c(k)$ is the k-th CLP coefficient and Q is the number of LP coefficients. The number of CLPC — Q' should be greater than Q so that a satisfactory spectral approximation is achieved and usually an inequality: $Q' \ge Q^* 3/2$ is met. The term c_0 is closely related to the gain and thus carries no information about the spectrum shape. Therefore, we will not include this term in the final CLPC vector. Actually, energy parameters if not processed separately, can even deteriorate the detection process in cases of small SNR. Finally, these vectors are used to form the two-dimensional cepstrum that spans proper time duration, over which variations are assessed by a DCT or DFT. In Fig. 1 two blocks of cepstral vectors and corresponding cepstral matrices for noise and noised speech are shown.

3 ARTIFICAL NEURAL NETWORKS

There are a great variety of artificial neural networks, which are classified according to different criteria such as topology, learning algorithms and paradigms etc. Here we adhere to only one class (multi-layer perceptron — MLP), which is widely used in signal processing applications. In Fig. 1 we can observe an obvious distinction between noised signal and WSS noise merely by looking

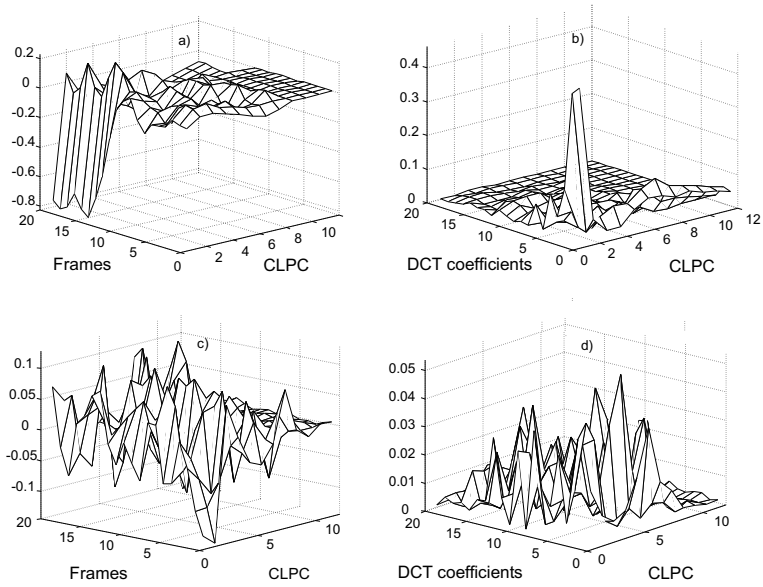

Fig. 1. a) The block of 20 CLPC vectors taken from a word, deteriorated by the white noise and SNR=3dB, b) An absolute value of the cepstral matrix constructed for the A case, c) The block of 20 CLPC vectroes taken from the white noice, d) An absolute value of the cepstral matrix constructed for the C case

at it. This natural and simple method "look and see" seems to work quite reliably even at low SNR and in various sorts of WSS noises. This so-called mapping is obviously rather tricky and mathematically not very traceable. Problems that are not well defined and mathematically vague are known to be successfully solved just by the ANN approach.

Methods of mapping can be well realized by the multilayer perceptron (MLP). It is stated by the universal approximation theorem saying: function $f(\cdot)$ can approximate any function $F(\cdot)$ whose inputs are bounded to $(0, 1)$ with optional accuracy if $f(\cdot)$ is defined as follows:

$$f(x1, x2, \ldots, xp) = \sum_{i=1}^{M} \alpha_i * \varphi\Big(\sum_{j=1}^{p} w_{ij} * x_j - \theta_i\Big). \quad (7)$$

Here α_i, θ_i, w_{ij} are constant real numbers and φ is bounded, continuous and monotonically raising function. This mathematical statement is directly applicable to the multi-layer perceptron with p inputs, one hidden layer of unspecified number of neurons and 1 neuron in the output layer, which act as a linear mixer. For better representation, a MLP realization of equation (7) is shown in Fig. 2. Training of MLP aims to set weights so that the average total square error for all input output training pairs is minimized. This approach would require storage of all necessary data for each neuron and training pair, which would be later used for new weights calculation. For real applications this would be unfeasible, so we usually resort to its approximation. Then weights are adjusted for each training vector pair that leads to the computation of instant error only [5], [7]. Unfortunately, despite relaxing the training conditions there is no analytic solution to this minimization problem. Furthermore, the backpropagation learning algorithm [7] used for MLP does not guarantee reaching the global minimum. This algorithm sets weights climbing up the negative gradient of the error space according (8):

$$\Delta w_{ij}(n) = -\eta \frac{\delta \varepsilon(n)}{\delta w_{ij}(n)}, \; \varepsilon(n) = \sum_{i=1}^{k} \big(x_i(n) - d_i(n)\big)^2. \quad (8)$$

Where η represents the learning parameter, the higher its value the faster the learning process is, but it is done on the expense of its stability. Variables $x_i(n)$ and $d_i(n)$ are the actual output of the i-th neuron in the output layer in the n-th step and its expected value respectively. This learning parameter is processed in different ways in the phase of training to avoid speed and convergence problems. Usually η changes with the course of the training process (its value declines in the time). Further, its value can differ in various layers (the deeper the layer the smaller its value). Another technique to tackle the abovementioned problems is called momentum technique, which aims to speed up the training process while maintaining good stability. It is done by introducing a momentum term that takes in account previous changes of the weight. If the current and previous changes correlate, the adjustment step is bigger otherwise the adjustment is smaller (9).

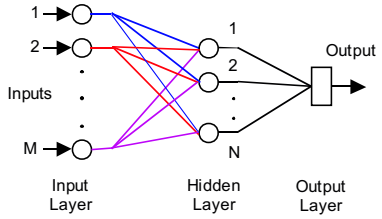

Fig. 2. The MLP containing 1 hidden kayer and realizing the function approximation given by (7).

$$\Delta w_{ij}(n) = \eta \delta_j(n) y_i(n) + \alpha \Delta w_{ij}(n-1)\,. \qquad (9)$$

Here $\delta_j(n)$ is the local gradient (the rule of weight adjustment defined for the backpropagation algorithm [5, 7]) for the j-th neuron in the n-th step, $y_i(n)$ is the input signal coming from the i-th neuron and α is the moment constant; $0 \leq \alpha < 1$, for convergence reason.

The set of input output training vector pairs X and D should be representative enough so that substantial features of the required mapping can be observed, learned and well generalized. Training stops when the testing set of the input output pairs (pairs that are not used for weights adjustments) reaches the minimal approximation error, while the training set still exhibits an error decline in the course of further training. This at the first glance abnormal behaviour can be explained as MLP tries to memorize the training data while loosing its ability to generalize; this phenomenon is known as overtraining. Thus it is always necessary to design actually 2 equal sets of input output vectors (training and testing). When adjusting weights after each training pair, it is usually advised to randomise the order in which the training vectors pairs are submitted to the network. It is done to prevent the MLP training to depend on certain input-output sequences.

4 DETECTION ALGORITHM

Design of a good detection algorithm involves many factors like: choosing proper speech features, setting up eligible measuring techniques and specifying reasonable decision criteria. In the following, each function is described in detail.

4.1 System Description

Imposing WSS restriction we based our detection system on evaluating negligible changes of spectral envelopes over the reasonable time duration. In WSS noises there should be negligible amount of variations of spectral envelopes that are of different kind. However, our algorithm must be flexible enough to allow some fluctuations either caused by estimation dispersions or by the present noise not strictly fulfilling WSS assumption. On the contrary, in a noised word some specific kinds of variations, which would reflect transitions between phonemes in the word, should be prevailing. A simple presentation of these important features can be easily done by cepstral matrices, as shown in paragraph 2. In the next stage there is an evaluation algorithm assessing sorts of changes and spectral shapes. Its aim is to provide us with some measure of certainty, whether it is noise-only block or that may be a part of word. In this stage we decided to use the MLP approach for its outstanding classification properties. Then the output of MLP is, so-called, criteria function. In the final, decision-taking step we use a simple smoothing of the criteria function by a linear FIR filter and the elementary threshold-based detection algorithm to determine the exact boundaries of tested utterances. This should further increase the overall robustness of the system by taking in account the behaviour of the criteria function over some time period and not only one cepstral matrix. In Fig. 3 the whole system is depicted using a block scheme.

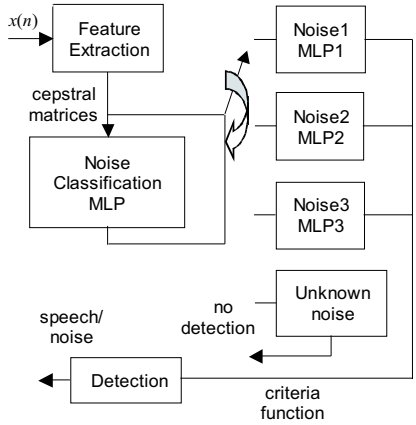

Fig. 3. The block diagram of voice activity detection system for WSS noises, using cepstral matrices and ANN.

4.2 Algorithm Settings

In order to give the maximum efficiency to the proposed algorithm it is important to set up its free parameters properly. Of course, it could be done optimally by exhausting experiments but that would be time unfeasible. Thus we rather utilized the general knowledge of speech and WSS noises.

Original, 22 kHz sampled signal was segmented into frames of the duration of 20 ms with 10 ms overlap of successive frames and windowed by Hamming window. This should well represent static intervals and keep the courses of feature vectors relatively smooth. The number of LP coefficients was set to 8 that provides a good estimation of spectral envelopes. These LPC were derived by the autocorrelation and Durbin methods and were further converted into 12 CLPC for more "natural" log spectra representation and straightforward handling. This number was a compromise between satisfactory log spectra approximation and the time and memory consumption. In order to be able to distinguish speech active intervals by spotting transition processes in words, we decided for a 210 ms length of cepstral matrices (20 CLPC vectors). Again this is a strict compromise between the accuracy of the detection and the reliability of the decision (the longer the matrix the more reliable decision can be made, but the less accuracy can be expected- exact position within the matrix is unknown).

As a proper class of ANN a fully connected MLP was chosen referring to the universal approximation theorem. However, it does not specify the number of neurons in the hidden layer and thus it is rather a matter of experiments. Although a reasonable upper limit can be determined following the total number of weights to be set and

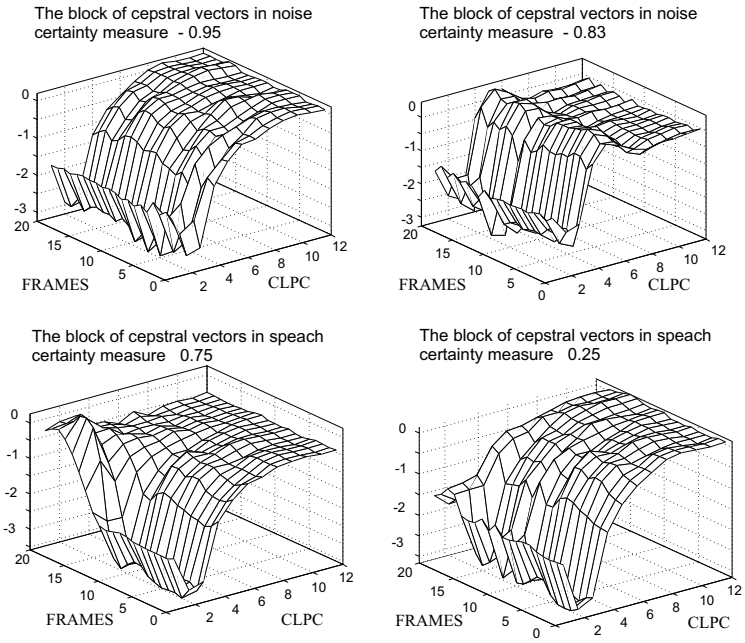

Fig. 4. The example of training matrices covering different environments and various certainty values, *ie* assessed believe of a given matrix being taken from word only.

the amount of available training data. It is clear (from the statistical point of view) that there must be multiple data for appropriate weight estimation. We used approximately 80 training pairs for each class of noise and each training vector (cepstral matrix) consists of 240 elements. Therefore it would not be very meaningful to use a network with more parameters to set like the total number of data; by experiments we set the number of hidden neurons to 10. Furthermore we decided to modify the activation function of the output layer neuron as given by (7) from identity to $tanh$. It is needed to bound the output somehow *ie* $(-1, 1)$, 1 for pure speech and -1 for noise only. However other (transition) states are vital and thus well trained too. This slight modification should not limit MLP's approximation ability; on the contrary it should increase its flexibility.

4.3 MLP Training

We found during the data preparation process that for satisfactory description of available WSS noises there were approximately 80 matrices needed for each. This provided a fair description of possible shapes and variations of cepstral matrices and thus a good generalization could be achieved. Data were carefully chosen for both testing and training sets covering a huge variety of different cepstral matrices. These matrices had to be selected and evaluated manually so that the scores of certainty were well balanced over the whole output interval $(-1, 1)$. In Fig. 4 there are some blocks of cepstral vectors each of which has obviously a different measure of certainty of describing a word interval and thus they are labelled distinctively. It should be noted that these are not the cepstral matrices yet, because DCT is still missing. During the data preparation process it was found that blocks of cepstral vectors are more easily recognized by human than cepstral matrices. It is documented in Fig. 1, where both cases are depicted next to each other. Shapes of cepstral matrices seem more like randomly generated even when being constructed over smooth signals. This is because of DCT's good decorrelation properties for most common signals where it can act as Hotelling transform [8]. On the contrary, cepstral matrices have achieved better approximation results as well as faster training times than those blocks of cepstral vectors. It can be caused by suppressing linear dependences of samples which may "hide" their less significant non-linear relations which are, however, more important for the classification itself. In fact we used only absolute values of those matrices. By doing so, we intended to highlight the amount of proper changes of cepstral vectors in the time and suppress the useless information about the exact location of theses changes within the matrix (we made the decision over the matrix as a whole not only part of it). Moreover, this enabled us to significantly reduce the number of training matrices, because we did not have to submit all the possible positions of those fluctuations. Instead, we pursued only their different shapes, which are important for the detection process. That is why, blocks of CLPC vectors were first manually selected and evaluated and subsequently converted to absolute value cepstral matrices that were finally used in the training process. This data preparation process is apparently most bent to human errors and solely depends on the designer experience. Training itself was accomplished using a simple backpropagation algorithm with random order of samples in each training epoch and was stopped just when the overtraining had been detected. Original weight initialisation was done with random uniformly distributed numbers in the interval $(-1, 1)$. Of course more networks were designed (dif-

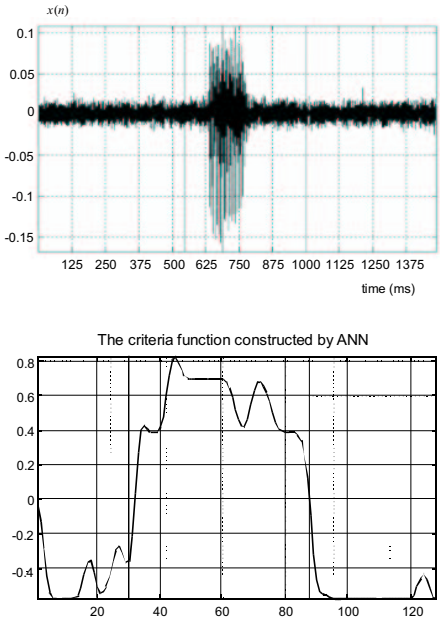

Fig. 5. The course of a noised Slovak word (upper graph) and accompanying criteria function (lower graph). SNR=10 dB. The solid vertical lines mark word boundaries.

ferent initialisations) and that one with the least total error was finally selected. During the course of training it was shown that it is effective to build more networks each for a particular noise or a group of noises. By having done so, results were much better, training was faster, data preparation much simpler and above all the structure became much more flexible. However, it was done at the expense of designing a noise classifier, which proved not to be difficult to build. Our classifier was a fully-connected MLP with 4 neurons in the hidden layer and was able to classify 4 different environments. Some spectrally close noises were put together without any loss of accuracy. It was able to detect unspecified noise too where the detection could not be done.

5 EXPERIMENTS AND RESULTS

All experiments were executed on a set of 36 Slovak specially chosen words, where each word existed in several utterances (male, female). All most important Slovak phonemes were alternated at the beginning and at the end of words and thus this should have provided as objective evaluation of our system as possible. We tested these words in the white noise and three different WSS colour noises. However, two of them were very similar in the cepstral representation and so they were grouped together. The words were artificially noised in the following SNR values: 3, 7, 10, 13 and 20 dB where the average SNR value of "clean" recordings was about 30 dB. Each word was located in the separate wav file, which had at least 400 ms interval of silence or noise before and after the utterance. To get better notion of what the criteria functions look like, there are depicted courses of a noised Slovak word with accompanying criteria function in Fig. 5, both with marked end pints.

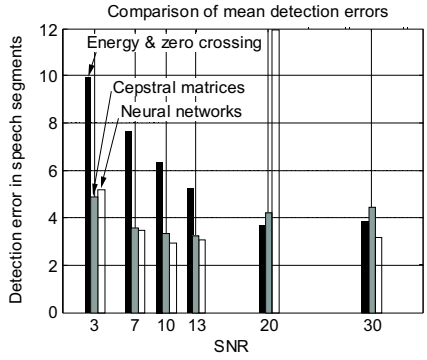

Fig. 6. Comparison of a mean detection errors for different detection methods as a function of SNR.

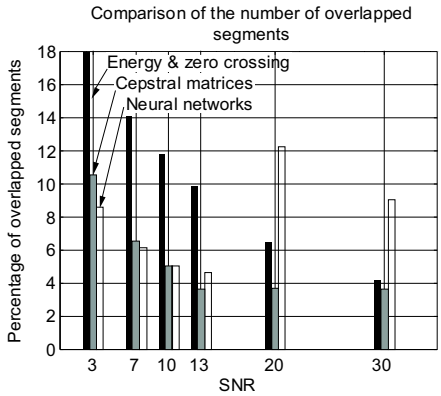

Fig. 7. Comparison of mismatched frames (overlaped values of criteria functions) constructed for different algorithms.

Under those conditions we performed several tests for accuracy, detection reliability, noise classification errors and we compared them with other approaches. As the reference methods we chose the basic energy and zero crossing method and the method, which tests global variations in cepstral matrices (uses the same speech features) [9]. Perhaps, the most important attribute we tested is the mean detection error for all algorithms, which is depicted as a function of SNR in Fig. 6. Another important parameter of those algorithms is their ability to mark speech segments without using other information in account like previous and following frames, etc. This means we counted mismatched signal segments, *ie* noisy segments above the threshold and speech segments lying below it. The higher the mismatch (% of overlapped values of the criteria function) the more sophisticated detection algorithm must be employed to achieve the same detection error. This feature is presented in Fig. 7. Finally we tested the noise classifier for all supported noises as well as for unknown ones. Table 1 gives the worse classification error for each group of noise as a function of SNR.

5 CONCLUSION AND COMMENTS

- Essential attributes (the detection error and the ratio of misclassified samples) for both detection algorithms

based on cepstral matrices, shown in Figs. 6 and 7 as a function of SNR do not exhibit constantly declining courses as the energy methods do and as it would be natural. They appear to have their global minimum between 10–17 dB. This rather abnormal behaviour can be explained by cepstral matrices being energy independent and by their high sensitivity to articulation mistakes (lip clicking, exhalation, ...), which are of lower energy but still causing eligible spectral changes near utterances.

Table 1. Classification errors of the MLP noise classifier

Noises	SNR 3 dB	SNR 7 dB	SNR 10 dB	SNR 13 dB	SNR 20 dB
White like noise	0.46 %	0.46 %	0.46 %	1.39 %	50.2 %
Hair like noises	0.46 %	0.46 %	0.46 %	0.465 %	1.39 %
TV like noises	0.46 %	0.46 %	0.46 %	0.46 %	0.93 %
Unknown	5.11 %	6.04 %	6.51 %	6.51 %	7.9 %

- The main, but the only drawback as seen in Figs. 6 and 7 is the necessity to submit not only adequate variety of cepstral matrices of noised words and noises respectively, but it is inevitable to cover different SNR levels too, if they are expected. Training sets used in our tests were assembled from samples at 10 dB of SNR only, which caused a higher detection error in both extreme cases. On the other hand we can observe notable robustness of the ANN approach in the broad interval ranging from 3 to 13 dB.
- We may cover some kinds of non- stationary noises as the ANN approach can be regarded as a sort of recognition methods, which are capable to operate in non- WSS noises [4].
- It is important to stress that the final success is a subject of the designer's experience in the phases of training and data preparation.
- Proper cooperation of ANN and cepstral matrices exhibits outstanding properties even in low SNR values and provides a flexible structure that can be optimally adjusted to the current environment.
- Speech is a strongly dynamic process and as such it would be more appropriate to use recurrent ANN [5]. However, it would make the process of training, data preparation, etc. extremely difficult. This can be easily avoided by using cepstral matrices that contain vital time dynamics. Thus MLP with cepstral matrices proves to be a very effective tool for speech processing applications.

References

[1] DELLER, J. R.—PROAKIS, J. G.—HANSEN, J. H.: Discrete-Time Processing of Speech Signals, Macmillan Publishing Company, Englewood Cliffs, 1993.

[2] HSIO-FEN PAI—HSIAO-CHUAN WANG: A Study of Two-Dimensional Cepstrum Approach for Speech Recognition, Computer Speech and Language (1992) 6.

[3] MILNER, B. P.—VASEGHI, S. V.: Speech Modelling Using Cepstral-Time Feature Matrices and Hidden Markov Models, IEEE, 1994.

[4] RABINER, L.—BIING-HWANG JUAN: Fundamentals of Speech Recognition, Prentice Hall PTR, 1993.

[5] ROBINSON, T.—HOCHBERG, M.—RENALS, S.: The Use of Recurrent Neural Networks in Continuous Speech Recognition, Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, U.K., 1995, available at: http://svr-www.eng.cam.ac.uk/~ajr/rnn4csr94/rnn4csr94.html.

[6] JÁN, J.: Numerical Filtration, Analysis and Reconstructions of Signals (Číslicová filtrace, analýza a restaurace signálů), Vysoké učení technické v Brně, 1997. (in Czech)

[7] ORAVEC, M.—POLEC, J.—MARCHEVSKÝ, S.: Neural Networks in Numerical Signal Processing (Neurónové siete pre číslicové spracovanie signálov), FABER, Bratislava, 1998.

[8] ROZINAJ, G.–POLEC, J.–KOTULIAKOVÁ, J.–PODHRADSKÝ, P.—MARČEK, A.—MARCHEVSKÝ, S.: Numerical Signal Processing, II (Číslicové spracovanie signálov, II), FABER, Bratislava, 1997.

[9] KAČUR, J.—ROZINAJ, G.—HERRERA, S.: Word Boundary Detection Based on Global Variatiation Measuring CLPC Vectors Using Cepstral Matrices, International Conference on Telecommunications ICT'2000, Acapulco, Mexico, May 22-25, 2000,.

Received 22 January 2004

Juraj Kačur, born in Bratislava in 1976, obtained Msc degree (Ing) from the Faculty of Electrical Engineering and Information Technology of the Slovak University of Technology (FEI STU) Bratislava, in informatics-telecommunication. He was with the Slovak Academy of Sciences, Department of Speech Analysis and Synthesis where he participated in several projects. Since March 2000 he has been a PhD student at the Department of Telecommunication at FEI STU, Bratislava, where since 2001 he has been assistant professor. The field of his research activities includes: digital speech processing, high order statistic, wavelet transform, ANN and HMM.

Gregor Rozinaj received MSc and PhD in telecommunications from the Slovak University of Technology, Bratislava, Slovakia in 1981 and 1990, respectively. He has been a lecturer at the Department of Telecommunications of the Slovak University of Technology since 1981. In 1992–1994 he worked on speech recognition at Alcatel Research Center in Stuttgart, Germany. From 1994 to 1996 was with the University of Stuttgart, Germany working on automatic ship control. Since 1997 he has been Head of the DSP group at the Department of Telecommunications of the Slovak University of Technology, Bratislava. Since 1998 he has been Associate Professor at the same department. He is an author of 3 US and European patents on digital speech recognition and 1 Czechoslovak patent on fast algorithms for DSP. Dr. Rozinaj is a member of IEE and IEEE Communication Society.

Sergio Herrera-Garcia (1948–2003). Graduated from the Telecommunication Department of the Slovak University of Technology (MSc 1978, PhD 1995)and was responsible for research projects and involved in signal processing, image processing and digital filtering. He published several papers in these fields. From 1988 he was with the Mexican Research Institutes UNAM, IEE, CICESE, CINVESTAV. From 1999, he was with the CITEDI-IPN Research Center (www.citedi.mx) located at the US border in Tijuana, Baja California, Mexico, part of the National Polytechnic Institute (www.ipn.mx). His research interests included image compression and wireless communications.