BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **6**, No 3

Sofia • 2006

Robust Features and Neural Network for Noisy Speech Detection

Atanas Ouzounov

Institute of Information Technologies, 1113 Sofia
E-mail: atanas@iinf.bas.bg

Abstract: *In this paper, the effectiveness of three features in speech detection tasks is experimentally studied. The first feature is obtained by processing of the spectral autocorrelation function, while the second one is based on the multi-band spectral entropy. The well-known mel-cepstrum is utilized as a third feature. A multi-layer perceptron based speech detector is developed and speech detection tasks with noisy data are carried out for each feature. The performance analysis of the speech detection results is done using the ROC curves and measures. The experimental results revealed that the feature obtained by processing of the spectral autocorrelation function is more suitable for noisy speech detection than the other two features.*

Keywords: *Speech Detection, Spectral Entropy, Neural Network, ROC Curve.*

1. Introduction

The finding of speech fragments in a given signal has many names, of which some are speech detection, endpoints detection, voice activity detection, speech activity detection, and speech/non-speech segmentation [6].

The speech detector is one of the key components in speaker recognition systems designed to operate in noisy real-world environments. The recognition error in such systems is due to many causes, one of which is the inaccurate speech fragments detection. The speech fragments usually provide data for speaker model estimation. The non-speech ones are discarded or are used for noise parameters estimation with the purpose of reducing the noise influence on the recognition performance.

The existing Speech Detection (SD) algorithms can be divided into three major groups: energy-type SD, statistics-based SD and pattern recognition-based SD. In each group, different features are utilized combined with different decision rules.

The algorithms belonging to the energy-type SD usually analyze the time variations (trajectories) of selected parameters and utilize a set of thresholds and finite-state automata in order to produce a speech/non-speech decision for a particular segment [6].

The statistical-based SD includes algorithms that examine statistical properties of speech and noise signals. It is assumed in [17] that the spectral components of speech and noise signals have Gaussian distribution. The speech detection in that case is based on the likelihood ratio test. Other works focus on the different properties of the higher-order statistics of speech and noise signals in order to separate them in noisy environments [10].

The third group comprises algorithms based on pattern recognition techniques. Recently, different classification approaches are proposed in attempts to find the optimal classification rule in speech detection. Some of these approaches are tree-based modeling [16], support vector machines [2] and neural network classifiers [5, 15].

During the last few years, the frequently used features for speech detection in noisy environments are based on the spectral entropy characteristics [7, 13]. In this case, the main assumptions are, firstly, the signal spectrum is more "organized" in the speech rather then in the noise regions and secondly the Shannon's entropy can be used as an appropriate measure of signal organization [13].

In this paper, the effectiveness of three features in speech detection tasks is experimentally studied. The first feature is the Mean-Delta (MD) feature, which is obtained by processing of the spectral autocorrelation function [12]. The second one is based on the multi-band spectral entropy characteristics [9]. The well-known Mel-Frequency Cepstral Coefficients (MFCC) is utilized as a third feature [19].

In order to examine the features, a Multi-Layer Perceptron (MLP) based speech detector is developed. For each of the mentioned above features are carried out speech detection tasks with noisy data. The effectiveness of the features is estimated by comparison of the detection results using the Receiver Operating Characteristics (ROC) graphs technique.

The present work is focused on the raw speech/non-speech classification without speech enhancement and hangover mechanisms. Its aim is to identify which feature is more effective in real-world noisy environments.

2. The robust features

2.1. The Mean-Delta feature

The MD feature is proposed in [12] and it is defined as the mean absolute value of the delta spectral autocorrelation function of the power spectrum of speech signal. Let $x(i)$ is a discrete signal, where $i = 0, \ldots, I - 1$, I is the number of samples and the spectrum $X(k)$ of $x(i)$ is obtained by the Discrete Fourier Transform (DFT), where $k = 0, \ldots, K/2$, K is the number of points in the DFT.

The biased spectral autocorrelation function $R_p(l)$ is defined with the power spectrum $|X(k)|^2$as [12]

$$R_{\mathrm{p}}(l) = \sum_{k=0}^{K/2-1-l} |X(k)|^2 |X(k+l)|^2 , \tag{1}$$

Where $l = 0, \ldots, L$, L is the number of correlation lags and $L = K/2–1$.

In order to remove the tilt in the spectral autocorrelation function and enhance its peaks, in [12] is proposed a parameter obtained in a way similar to the delta cepstrum evaluation [20]. It is named as Delta Spectral AutoCorrelation Function (DSACF). This parameter is computed as an orthogonal polynomial fit of the first-order derivative (in correlation domain) of the spectral autocorrelation function.

For a particular frame, the DSACF is computed utilizing only the frame's spectral autocorrelation lags. For the nth frame, the DSACF $\Delta R_{\mathrm{p}}(n, l)$ is

(2)
$$\Delta R_{\mathrm{p}}(n,l) = \frac{\sum_{q=-Q}^{Q} q R_{\mathrm{p}}(n, l+q)}{\sum_{q=-Q}^{Q} q^2},$$

where $l = 0, \ldots, L$; $n = 0, \ldots, N-1$, N is the number of frames. The parameter Q defines the window width around the lag l and it influences over the accuracy of the approximation. For the purpose of this study, it is chosen to be between 10 and 15 lags (based on preliminary experiments). For n-th frame the MD feature $m_{\mathrm{d}}(n)$ is computed as follows

(3)
$$m_{\mathrm{d}}(n) = \frac{1}{\Delta L} \sum_{l=L_1}^{L_2} \left| \Delta R_{\mathrm{p}}(n,l) \right|,$$

where $\Delta R_{\mathrm{p}}(n, l)$ is the DSACF in (2) for lag l, L_1 and L_2 are the boundary lags and $\Delta L = L_2 - L_1 + 1$. For more details about the MD feature, see [12].

Typically, this feature is designed for energy-type SD with trajectory analysis. In order to be used as a frame feature vector in pattern recognition tasks some changes in its definition have been made. Instead of mean value estimation in (3), here is proposed to find the maximal values of $\Delta R_{\mathrm{p}}(n)$ in different non-overlapping ranges of lags. The MD feature vector for nth frame is formed as $\{m_{\mathrm{d}}(1), \ldots, m_d(J)\}$. Its components are defined as follows (for simplicity, the frame index is omitted)

(4)
$$m_{\mathrm{d}}(j) = \max \left\{ \Delta R_{\mathrm{p}}(l) \right\}_{l=L_j}^{l=L_{j+1}},$$

where $j = 1, \ldots, J$, J is the number of ranges and $\{L_1, L_2\}, \ldots \{L_j, L_{j+1}\}, \ldots \{L_{2J-1}, L_{2J}\}$ are pairs of boundary lags for each range.

The proposed algorithm for the MD feature vector estimation is summarized as follows (for each frame):

- apply Hamming window to the analyzed signal;
- compute the power spectrum of the windowed signal via FFT;
- compute the non-normalized biased spectral autocorrelation function by equation (1) with lags $L=K/4$;
- compute the delta spectral autocorrelation function by equation (2);
- take the absolute value of the delta spectral autocorrelation function;
- divide the number of lags L into J non-overlapping lags ranges of equal size;
- find the maximal values of $\Delta R_{\mathrm{p}}(l)$ in the lags ranges $\{L_1, L_2\}, \ldots \{L_j, L_{j+1}\}, \ldots \{L_{2J-1}, L_{2J}\}$ according to (4);

- take the logarithm of the maximal values and obtain the MD feature vector in the form $\{\lg(m_d(1)), \ldots, \lg(m_d(J))\}$.

The last step in the MD feature vector estimation is the mean normalization. It is done by dividing the MD feature vector for each frame by the average MD feature vector computed over all frames. If the speech data consists of different speech records (files), the mean normalization should be applied for each file separately.

2.2. Multi-band spectral entropy

The spectral entropy for the n-th frame is estimated in the following steps [9]. First, the Probability Mass Function (PMF) $P(|X(n, k)|^2)$ for the full-band power spectrum $|X(n, k)|^2$ is computed as

$$P(|X(n,k)|^2) = \frac{|X(n,k)|^2}{\sum_{k=0}^{K/2} |X(n,k)|^2}, \tag{5}$$

where $k = 0, \ldots, K/2$, K is the number of DFT-points and $n = 0, \ldots, N-1$, N is the number of frames.

Second, the spectral entropy $H(n)$ for n-th frame is computed as follows

$$H(n) = -\sum_{k=0}^{K/2} P(|X(n,k)|^2).\lg_2(P(|X(n,k)|^2)). \tag{6}$$

The entropy in (6) is named as full-band spectral entropy [9]. To capture a local variation in the spectrum, the idea of multi-band spectral entropy is introduced in [9]. The core of this idea is to divide the full-band PMF into sub-bands and then the spectral entropy to be computed for each sub-band using full-band PMF. In this case, one entropy value is obtained for each sub-band.

According to [9] the Multi-Band Spectral Entropy (MBSE) feature vector for the nth frame is formed as $\{H_{\mathrm{MBSE}}(n, 1), \ldots, H_{\mathrm{MBSE}}(n, G)\}$ and its components are computed as

$$H_{\mathrm{MBSE}}(n,g) = -\sum_{k=B_g}^{B_{g+1}} P(|X(n,k)|^2).\lg_2(P(|X(n,k)|^2)), \tag{7}$$

where $P(|X(n, k)|^2)$ is the full-band PMF in (5); $g = 1, \ldots, G$, G is the number of sub-bands and $\{B_1, B_2\}$; … $\{B_g, B_{g+1}\}$; … $\{B_{2G-1}, B_{2G}\}$ are pairs of boundary spectral bins for each sub-band.

The MBSE feature is utilized in [9] as additional feature in automatic robust speech recognition tasks. So far, it is not used as feature for robust speech detection.

3. The neural network

A feedforward multi-layer perceptron structure with one hidden layer is used in the experiments. The selected activation functions of the neurons are – hyperbolic tangent function (in hidden layer) and sigmoidal function (in output layer). It is known that the standard Back Propagation (BP) learning algorithm for MLP suffers from slow convergence speed and local minima problem. In the past, some fast learning algorithms have been proposed for the MLP training, such as Rprop, Quickprop, the Levenberg-Marquadt, etc. [8]. In the present work, the Rprop algorithm with most typical parameters settings, according to recommendation in [14] is applied.

4. Experiments

In the experiments are utilized speech samples selected from two databases – updated version of the BG-SRDat corpus [11] and the SpEAR database [1].

The BG-SRDat is a corpus in Bulgarian language collected over noisy analog telephone channels and designed for speaker recognition. The speech data included in the BG-SRDat are sampled with frequency of 8 kHz at 16 bits, PCM format and mono mode [11].

The SpEAR database contains samples of noise-corrupted speech that have been recorded by acoustically combining clean speech and noise. All WAV files in the database are with sampling frequency of 16 kHz at 16 bits, PCM format and mono mode [1]. The speech files selected from SpEAR database are down sampled at 8 kHz. The analyzed frequency range is up to 4000 Hz. No additional filtering is applied. The analysis parameters are frame length – 30 ms, the frame shift – 10 ms, and the FFT-points – 512.

The number of sub-bands in the entropy estimation is $G=15$. The number of lags regions is the same, i.e., $J=15$ and $Q=15$ in (2).

A Mel-scale triangular filter bank with 24 filters is used to generate the MFCC feature. The MFCC vector size is 15 – 14 static coefficients and the zeroth cepstral coefficient. In addition, cepstral mean subtraction is applied (for each file separately) to obtain the MFCC feature.

The selected speech data is divided into four sets, named as S1, S2, S3 and S4. The purpose of each set is, as follows: S1 – for train and test, S2 – for train and test, S3 – for test and S4 – for validation. The sets S1, S2 and S4 are selected from the BG-SRDat, while S3 is from the SpEAR database. Each set comprises speech data from different speakers with different linguistic contents. The sets parameters (as number of speakers and total number of frames) are S1 – 3 speakers/5103 frames; S2 – 3 speakers/5828 frames; S3 – 2 speakers/1906 frames and S4 – 2 speakers/8918 frames.

To create the targets sequence for neural network training, all WAV files are manually segmented into speech/non-speech frames. For SpEAR database examples the segmentation is done on the clean files, but in the experiments are used their noisy versions. Unfortunately, the BG-SRDat corpus does not possess speech records in the form ‘clean speech reference – its noisy version’ and the segmentation in that case is done on the original noisy files.

The speech examples from SpEAR database are factory noise example (Signal-to-Noise Ratio (SNR) = –9.96 dB); car noise example (SNR = –14.58 dB); pink noise example (SNR = –10.33 dB) and F-16 noise example (SNR = –1.05 dB).

The MLP with a structure 15-20-1 is selected. The network has 20 neurons in one hidden layer and a single output neuron. The input vector size is set to 15. The used target levels are – minimal 0.1 and maximal 0.9 and the network is trained in batch mode. In testing mode, the output neuron level is thresholded at 0.5. No attempts are made in the experiments to estimate the optimal number of neurons and layers.

Here the combination of the MLP and a single feature is considered as a single classifier. In the experiments, three combinations are made and they are noted as MLP-MD, MLP-MBSE and MLP-MFCC, i.e. the classifiers MLP-MD, MLP-MBSE and MLP-MFCC are under analysis in the work. The effectiveness of the features is estimated indirectly by performance comparison of the classifiers.

The four train/test sets are prepared, noted as S1-S2, S1-S3, S2-S1 and S2-S3. These sets are utilized by classifiers in a multiple runs scheme [4]. In the experiments, 10 runs are performed (typically, runs are not more then 20 [4]).

The MLP training is stopped, when based on the validation test with S4 a global minimum in the output mean square error is found or this error is not changed significantly up to 300th epoch.

5. The performance comparison

It is known that the recognition error (as a single quantity) is not a reliable estimation of the recognition performance. This is true especially for the neural network classifiers where the learning algorithms include random number based procedures (e.g., in weights initialization, in training data selection, etc.) [4].

To compare the performance of the mentioned above classifiers the ROC technique is applied [3]. This technique is popular in the examining of the two-class classification problems. The ROC curve allows finding classifier that outperforms another using only the visual evaluation. The ROC space for two-class problem is two-dimensional plot where on *Y*-axis is True Positive Rate (TPR) and on *X*-axis is False Positive Rate (FPR). In our case (i.e. speech detection), the TPR is computed as ratio of the correctly classified speech frames to the speech frames. The FPR is computed as ratio of the incorrectly classified non-speech frames to the non-speech frames. Both rates are normalized between 0 and 1.

In order to compare the performance of different classifiers, their ROC curves are drawn on a common two-dimensional plane. Each curve is obtained by the results from a single run, i.e., it is an instance curve. In order to give an idea of the variance range, the instance curves are shown instead of their averaged version.

Additionally three ROC-measures are computed – recall (i.e. TPR), precision – (ratio of the frames correctly detected as speech to the frames detected as speech), and F-measure (harmonic mean of precision and recall) [3, 18]. Their values are shown in Table 1 and are obtained in the case when the MLP output is thresholded at 0.5. In the table, the maximal values for each measure (and for each train/test set) are in the shaded cells.

In the figures bellow are depicted the instance's ROC curves provided from the pair of classifiers {MLP-MD, MLP-MBSE} and {MLP-MD, MLP-MFCC}. The curves in Figs.1-4 are obtained when {MLP-MD, MLP-MBSE} utilize the train/test set S1-S2, S1-S3, S2-S1 and S2-S3, respectively. In Figs.5-8 are shown the ROC curves for classifiers {MLP-MD, MLP-MFCC} and the train/test set S1-S2, S1-S3, S2-S1 and S2-S3, respectively.

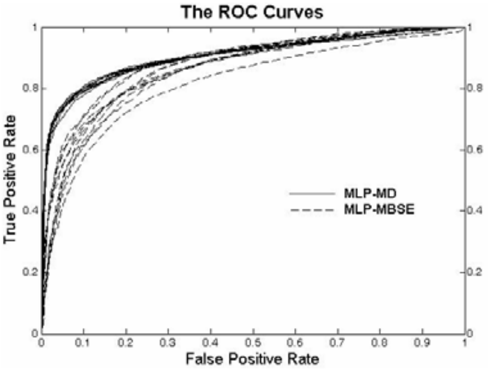

Fig. 1. Instance's ROC curves of 10 runs for MLP-MD and MLP-MBSE classifiers and train/test set S1-S2

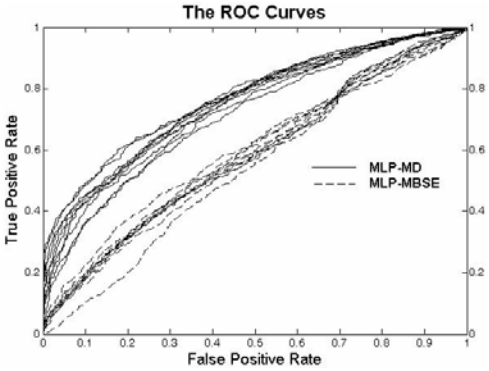

Fig. 2. Instance's ROC curves of 10 runs for MLP-MD and MLP-MBSE classifiers and train/test set S1-S3

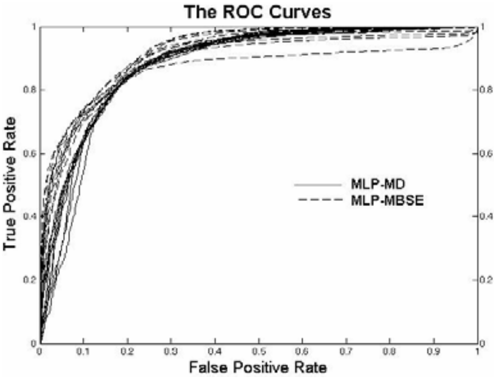

Fig. 3. Instance's ROC curves of 10 runs for MLP-MD and MLP-MBSE classifiers and train/test set S2-S1

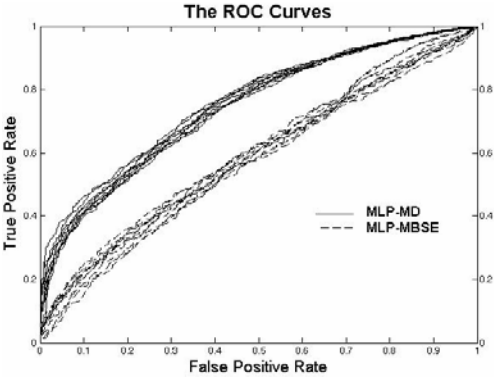

Fig. 4. Instance's ROC curves of 10 runs for MLP-MD and MLP-MBSE classifiers and train/test set S2-S3

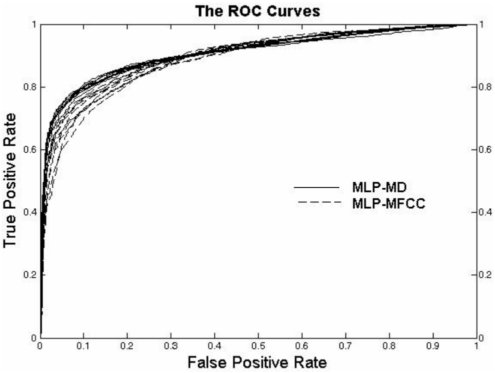

Fig. 5. Instance's ROC curves of 10 runs for MLP-MD and MLP-MFCC classifiers and train/test set S1-S2

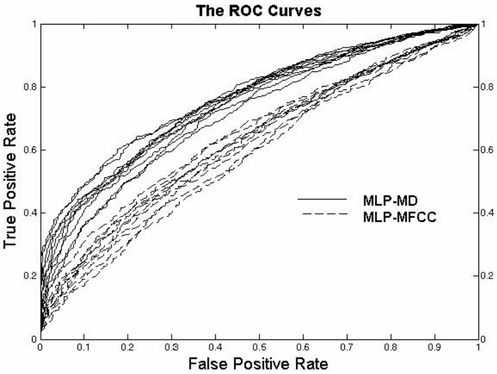

Fig. 6. Instance's ROC curves of 10 runs for MLP-MD and MLP-MFCC classifiers and train/test set S1-S3

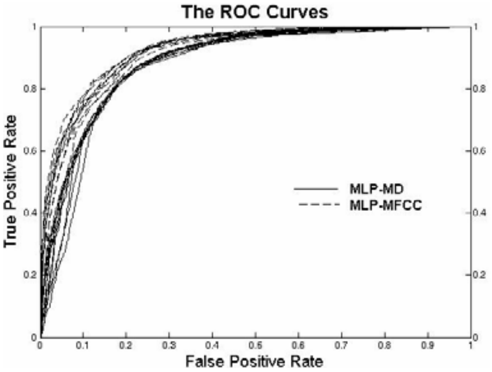

Fig. 7. Instance’s ROC curves of 10 runs for MLP-MD and MLP-MFCC classifiers and train/test set S2-S1

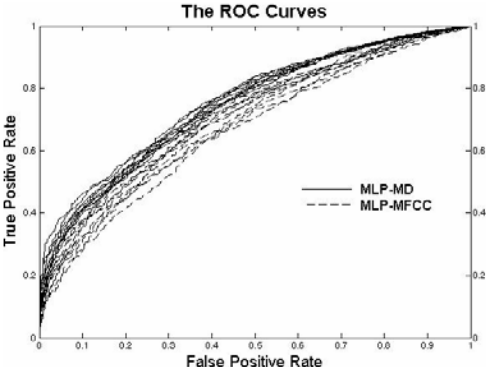

Fig. 8. Instance’s ROC curves of 10 runs for MLP-MD and MLP-MFCC classifiers and train/test set S2-S3

Table 1. ROC-measures

CLASSIFIERS	MEASURES	TRAIN/TEST SETS			
		S1-S2	S1-S3	S2-S1	S2-S3
MLP-MFCC	Recall	**0.854678**	0.870870	**0.893168**	0.943988
	Precision	0.868651	0.560420	0.877602	0.566393
	F-measure	0.861608	0.681976	**0.885317**	0.707991
MLP-MBSE	Recall	0.824922	0.514663	0.755782	0.411926
	Precision	0.835540	**0.586303**	**0.898380**	**0.622636**
	F-measure	0.830197	0.548152	0.820934	0.495823
MLP-MD	Recall	0.838098	**0.963050**	0.835130	**0.977419**
	Precision	**0.902047**	0.564358	0.876093	0.559886
	F-measure	**0.868898**	**0.711670**	0.855121	**0.711951**

6. Discussion and conclusions

It is clearly seen in Fig.1 that for almost all values of FPR the MLP-MD classifier outperforms the MLP-MBSE one. Moreover, it is evident in Fig.1 that the variance in the instance’s ROC curves for MLP-MD classifier is substantially lower than these ones for the MLP-MBSE classifier.

The curves that are depicted in Figs.2 and 4 reveal interesting fact about the ability of the MLP-MBSE classifier to work in noisy conditions. In the ROC plane, the diagonal line (between points (0, 0) and (1, 1)) represents the strategy of the random class selection [3]. If the classifier provides ROC curves close to this diagonal line then it would be ineffective in recognition tasks. As can be seen in Figs.2 and 4 that is the case with MLP-MBSE classifier for train/test sets S1-S3 and S2-S3. Again, the MLP-MD classifier outperforms the MLP-MBSE one.

Similar conclusions can be made about MLP-MD and MLP-MFCC classifiers based on curves shown in the Figs.5, 6 and 8, i.e. the MLP-MD yields better results.

Based only on the curves shown in Figs.3 and 7 it is not possible to say which classifier is the better one. In this case, the experiments yield unclear results and it is advisable to do a further research.

As seen in Table 1 the best performance in terms of the F-measure is achieved by the MLP-MD classifier – it is better in three of the four train/test sets.

Based on the obtained experimental results the following conclusions are made:

• in all test the MLP-MD classifier performs equal or better than MLP-MBSE one;

• for some train/test sets, the MLP-MBSE classifier is ineffective – its decision is close to the random class selection.

• in three of the four train/test sets, the MLP-MD classifier achieves the best performance in terms of the F-measure.

These results are a confirmation of our assumption that MD feature is more suitable for noisy speech detection than the multi-band spectral entropy. In some cases, the MD feature provides better results than the well-known MFCC. The major disadvantage of the MD feature is the increased computational costs. They are mainly due to the amount of calculation needed for the spectral autocorrelation function processing and the mean normalization.

We consider that the speech detector based on the joint work of the MD feature and the neural network is a good choice for one of the key components in the speaker recognition system designed to operate in real-world noisy environments.

References

1. Center for Spoken Language Understanding, Speech Enhancement and Assessment Resource (SpEAR) Database. Oregon Graduate Institute of Science and Technology. **http://cslu.ece.ogi.edu/nsel/data/SpEAR_database.html**
2. Enqing, D. et al. Applying Support Vector Machines to Voice Activity Detection. – In Proc. of the ICSP, 2002, 1124-1127.
3. Fawcett, T. An Introduction to ROC Analysis. – Pattern Recognition Letters, Vol. **27**, 2006, No 8, 861-874.
4. Flexer, A. Statistical Evaluation on Neural Network Experiments: Minimum Requirements and Current Practice. – In: Proc. of the 13th European Meeting on CSR, 1996, 1005-1008.
5. Kim, H.-I., S.-K. Park. Voice Activity Detection based on Radial Basis Function. – IEICE Transactions on Communication, Vol. **E88-B**, 2005, No 4, 1653-1657.
6. Li, Q., J. Zheng, A. Tsai, Q. Zhou. Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition. – IEEE Transaction on SAP, Vol. **10**, March 2002, No 3, 146-157.
7. Liang-Sheng, Huang, Chung-Ho Yang. A Novel Approach to Robust Speech Endpoint Detection in Car Environment. – In: Proc. of the ICASSP, 2000, 1751-1754.
8. LeCun, Y., Leon Bottou, Genevieve B. Orr, Klaus-Robert Muller. Efficient Backprop. Neural Networks, Tricks of the Trade, Lecture Notes in Computer Science LNCS 1524, Springer-Verlag, 1998.
9. Misra, H., S. Ikbal, S. Sivadas, H. Bourlard. Multi-resolution Spectral Entropy Feature for Robust ASR. – In: Proc. of the ICASSP, 2005, 253-256.
10. NemerW., R. Goubran, S. Mahmoud. The Fourth-Order Cumulant of Speech Signals with Application to Voice Activity Detection. – EUROSPEECH, 1999, 2391-2394.
11. Ouzounov, A. BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels. – Cybernetics and Information Technologies, Vol. **3**, 2003, No 2, 101-109.
12. Ouzounov, A. A Robust Feature for Speech Detection. – Cybernetics and Information Technologies, Vol. **4**, 2004, No 2, 3-14.
13. Renevey, P., A. Drygajlo. Entropy Based Voice Activity Detection in Very Noisy Conditions. – EUROSPEECH, 2001, 1883-1886.

14. Riedmiller, M., H. Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP algorithm. – In: Proc. of the ICNN, 1993, 586-591.
15. Shao, C., M. Bouchard. Efficient Classification of Noisy Speech using Neural Networks. – In: Proc. of the 7th ISSPA, 2003, 357-360.
16. Shin, W., B. Lee, Y. Lee, J. Lee. Speech/Non-Speech Classification using Multiple Features for Robust Endpoint Detection. – In: Proc. of the ICASSP, 2000, 1399-1402.
17. Sohn, J., W. Sung. A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation. – In: Proc. of the ICASSP, 1998, 365-368.
18. Ishizuka, K., H. Kato. A Feature for VAD Derived from Speech Analysis with Exponential Autoregressive Model. – In: Proc. of the ICASSP, Vol. **1**, 2006,789-792.
19. Toh, A., R. Togneri, S. Nordholm. Investigation of Robust Features for Speech Recognition in Hostile Environment. – In: Proc. of the Asia-Pacific Conference on Communications, 2005, 956-960.
20. Soong, F., A. Rosenberg. On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition. – IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. **36**, 1988, No 6, 871-879.