БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

КИБЕРНЕТИКА И ИНФОРМАЦИОННИ ТЕХНОЛОГИИ · Том **4**, № 2
CYBERNETICS AND INFORMATION TECHNOLOGIES · Volume **4**, No 2

София . 2004 . Sofia

# A Robust Feature for Speech Detection*

*Atanas Ouzounov*

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

***Abstract:*** *A new robust feature intended for speech detection is proposed in this paper. This feature is the mean of the absolute values of the delta spectral autocorrelation function of the power spectrum and it is named the Mean-Delta (MD) feature. For noisy speech, the trajectory variations of the MD feature are compared with those ones of the other robust feature – the Energy-Entropy (EE) feature, developed in* [6]. *The selected noisy speech samples from two databases (the SpEAR database* [8] *and the BG-SRDat corpus* [9]) *are used in the experimental study. In all the tests the MD feature has demonstrated the same or better results than those of the EE.*

***Keywords:*** *speech detection, voice activity detection, spectral autocorrelation, speaker recognition.*

## 1. Introduction

The location of speech embedded in various non-speech events has many names, of which some are speech detection, endpoints detection, voice (speech) activity detection, and speech/non-speech segmentation.

The speech detection algorithms can be separated into two general categories. The first embraces the algorithms that analyze the time variations (trajectories) of selected parameters. These algorithms utilize a combination of a state automata and a set of thresholds (fixed or adaptive) in order to produce a speech or non-speech decision for particular segment based only on the trajectories characteristics [5, 7]. The second category is comprised of algorithms based on a pattern recognition

* This research is supported in part by the National Science Fund of the Bulgarian Ministry of Education and Science, Contract No И-1302/2003.

technique. In this case, during the training mode the reference models for two classes (i.e., speech and non-speech) are created based on selected speech features. In classification mode, each segment is associated to one of the classes based on some kind of similarity function [1, 11].

The most frequently used features for speech detection are different energy transformations of the signal, spectral flatness, spectral entropy, autocorrelation functions, etc. The more comprehensive description of the speech features used for speech detection is given in [2, 5, 7, 10].

The feature selection for speech detection tasks usually is composed of two stages. The first stage is a preliminary selection. It is based on a visual evaluation on the graphically represented parameters. This selection is a feasible task only in cases when the parameters possess reasonable graphical representation.

The latter stage is the final feature selection and a recognition scheme usually is applied. The developed speech detection algorithm is embedded as a component of a complete speech or speaker recognition system. The effectiveness of different speech detection features is estimated experimentally based on their indirect influence on the recognition performance [1, 11].

In the present study, a robust feature indented for trajectory-based speech detection algorithms is proposed. This feature is the mean of the absolute values of the delta spectral autocorrelation function of the power spectrum. For different noisy speech signals the trajectory's variations of the proposed feature are compared with these ones of the other robust feature – the energy-entropy based one, developed in [6]. This is done by visual evaluation on the graphically represented trajectories of both features.

## 2. The spectral autocorrelation function

There are two well-known definitions of the term "Spectral AutoCorrelation Function" (SACF). The first one is proposed in [4] where the magnitude Fourier spectrum is flattened by bank of lifters and the spectral autocorrelation is computed at the output of each lifter. This autocorrelation function is used to obtain an estimation of the pitch frequency for noisy speech. In the second definition, proposed in [3] SACF is defined as discrete quantities of the magnitude spectrum with spectral resolution as in the Fourier transform used to obtain the spectrum. In [3] the SACF was utilized for developing of new linear prediction algorithm based on spectral envelope of the speech signal.

In the present study, the second definition of the SACF will be used. An interesting feature of SACF, first demonstrated in [3], is the fact that this function is periodic for not only voiced speech but also for unvoiced one. However, for the unvoiced speech this periodicity is not so significant.

Let $x(i)$ be a discrete speech signal, where $i = 0,...,I-1$, $I$ is the number of samples and the spectrum $X(k)$ of $x(i)$ obtained by the discrete Fourier transform is

$$X(k) = \sum_{i=0}^{I-1} x(i) \exp(-j2\pi \frac{ki}{I}), \quad (1)$$

where $k$=0, ..., $K$/2, $K$ is the number of points in the discrete Fourier transform.

According to [3] the spectral autocorrelation $R_B(l)$ is defined with the magnitude spectrum $|X(k)|$ as

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

where $l = 0, ..., L$; $L$ is the number of correlation lags and $L = K/2$.

In equation (2), the magnitude spectrum above Nyquist frequency is used in autocorrelation computation. It is known that for real signals this spectrum is symmetric to this one below Nyquist frequency.

We studied the spectral autocorrelation functions based not only on the magnitude spectrum but also on the power spectrum. Some preliminary experiments are carried out and they revealed that the power spectrum is more suitable when the main goal is to obtain more sensitive to spectral changes autocorrelation function.

If $S(k)$ is the power spectrum of $x(i)$ and $S(k) = |X(k)|^2$, where $k = 0, ..., K/2$ then the spectral autocorrelation function $R_p(l)$ is defined with the power spectrum as

$$R_p(l) = \sum_{k=0}^{K/2-1-l} S(k)S(k+l), \tag{3}$$

where $l = 0, ..., L$, $L$ is the number of correlation lags and $L = K/2 - 1$.

The three examples of typical speech signals and their normalized SACFs up to lag $L = 100$ are shown in Fig. 1.

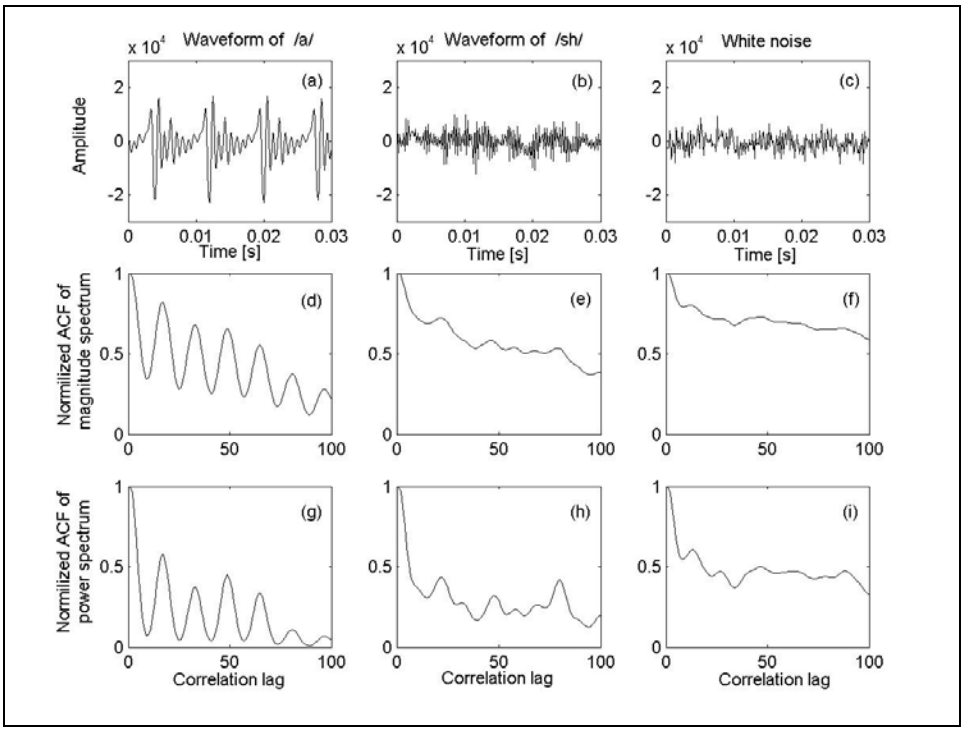


Fig.1. Examples of SACFs for: a) – voiced speech signal /*a*/; b) – unvoiced speech signal /*sh*/ and c) – white noise

We analyzed selected frames from three different signals. The first one is a part from a speech sound /*a*/, the second one is a part from a speech sound /*sh*/ and the third one is a white noise (generated by a program). The analysis parameters are sampling frequency of 8 kHz and frame length of 240 samples. The frame is Hamming-windowed and padded with zeros for 1024-point Fast Fourier Transform (FFT).

In each column in Fig.1 are shown the waveform of the signal and normalized SACFs ($R_B(0) = 1$, $R_p(0) = 1$) of the magnitude and power spectrums, respectively. It is clear that the spectral autocorrelation function of the voiced sound has a periodic property as is noted in [3]. The peaks in the autocorrelation functions in Fig.1 are more enhanced when functions are calculated by the power spectrums (see Fig.1 – (d) and (g), (e) and (h), (f) and (i)). It is interesting to note that for the unvoiced speech sound the power spectrum-based autocorrelation function possesses more clearly peaks than this one for white noise.

Hereafter the term "spectral autocorrelation function" in text will mean the autocorrelation function that is defined with the power spectrum.

## 3. The robust feature

For speech frames the spectral autocorrelation function is distinguished for more peaks with higher amplitudes than for non-speech ones (as seen in Fig. 1). We suppose that a single parameter, which values depend on the number of peaks and their amplitudes, can be utilized as feature for speech detection.

To obtain this parameter, in the study the spectral autocorrelation function is subjected to delta processing. The delta processing (i.e. the first-order orthogonal polynomial coefficient calculation) is applied in the autocorrelation domain, not in the time domain, as it is usually done [12].

We define the delta spectral autocorrelation function (DSACF) $\Delta R_p(n, l)$ for the $n$-th frame as

$$\Delta R_p(n,l) = \frac{\sum_{q=-Q}^{Q} qR_p(n,l+q)}{\sum_{q=-Q}^{Q} q^2}, \tag{4}$$

where $l = 0, ..., L$; $Q$ is typically between 2 and 5, i.e. regions from 5 to 11 lags are analyzed in the autocorrelation domain, and $n = 0, ..., N - 1$, $N$ is the number of frames. We suppose that $R_p(n, l) = 0$ for $n < 0$ and $l >$ L, i.e. the first and last few values of $\Delta R_p(n, l)$ will be influenced by these border conditions.

The examples of three typical speech signals (the same as in Fig.1), their normalized SACFs and the corresponded DSACFs (Q=3) up to lag $L = 100$ are shown in Fig. 2.

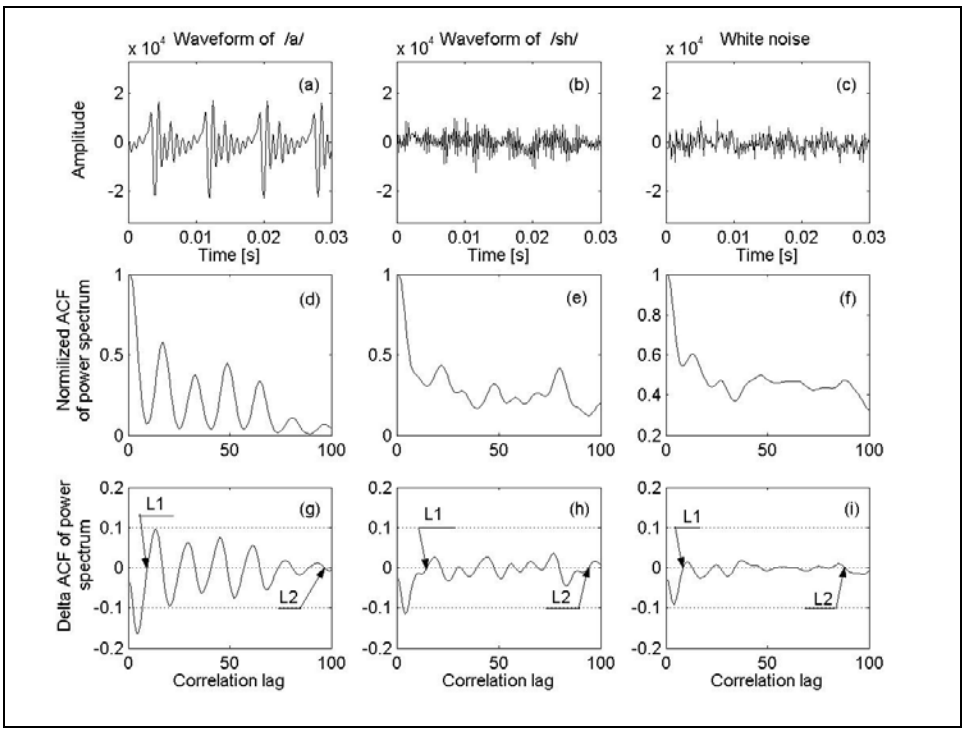


Fig. 2. Examples of SACFs and corresponded DSACFs for: a) – voiced speech signal */a/*, b) – unvoiced speech signal */sh/* and c) – white noise.

Based on the examples shown in the last row in Fig. 2 we propose as speech detection feature to be used the mean $m_d$ of the absolute values of the DSACF estimated for the part of correlation lags. For $n$-th frame $m_d(n)$, is computed accordingly to the formula

$$m_d(n) = \frac{1}{\Delta L} \sum_{l=L_1}^{L_2} \left| \Delta R_p(n,l) \right|, \tag{5}$$

where $\Delta R_p(n, l)$ is the DSACF in (4) for lag $l$, $L_1$ and $L_2$ are the boundary lags and $\Delta L = L_2 - L_1 + 1$. Hereafter $m_d(n)$ will be called Mean-Delta (MD) feature.

We decided to analyze the lags of the DSACF placed between its first ($L_1$) and last zero crossing ($L_2$) points (see the Fig. 2 (g), (h) and (i)). In this way, we avoid the border effects in the DSACF due to the delta feature calculation and the influence of the close to zero lag part of the SACF.

The described above ideas are included in the following algorithm for the MD feature calculation:

– divide the analyzed speech signal into sequence of frames with frame length of 40 ms and frame shift of 10 ms;

– for each frame apply Hamming window;

– for each frame compute the power spectrum of windowing speech signal via FFT – the FFT-points are $K$=1024;

– for each frame compute the non-normalized biased spectral autocorrelation function by equation (3) with $L = K/2 - 1$;

– for each frame compute the delta spectral autocorrelation function by equation (4) with $Q = 3$;

– perform a trajectory smoothing of delta spectral autocorrelation function (inter-frame processing). To produce the smooth trajectory of $\Delta R_p(n)$ we base on the idea proposed in [10] and used to obtain the so-called long-term spectral envelope. Here we apply the same idea (in fact this is a trajectory's smoothing algorithm) but on the trajectory of $\Delta R_p(n)$. The smoothed version of $\Delta R_p(n, l)$ is $\Delta R_p^S(n, l)$ and it is defined as [10]:

$$\Delta R_p^S(n,l) = \max\left\{\Delta R_p(n+j,l)\right\}_{j=-J}^{j=+J}, \tag{6}$$

where $n$ is the current frame, $l$ is the current lag and $J=2$;

– for each frame find first and last zero-crossing points $L_1$ and $L_2$ on the smoothed delta spectral autocorrelation function $\Delta R_p^S(n, l)$;

– for each frame compute $m_d(n)$ using $\Delta R_p^S(n, l)$ in (5) for lags between points $L_1$ and $L_2$;

– smooth $m_d(n)$ in time (inter-frame processing) with linear method (over five frames).

## 4. Energy-Entropy (EE) based feature

A new feature for speech detection obtained by combination of the energy and the spectral entropy is proposed in [6]. The experiments, which were carried out, revealed that this feature is more robust than the energy alone in the real-world environment as inside the driving car.

This EE feature is computed for every speech frame as follows (the frame index is omitted for simplicity):

– compute the energy $E$

$$E = \sum_{i=0}^{I-1} x(i)^2 ; \tag{7}$$

– estimate the probability density function $P(k)$ for the frequency component $k$ as

$$P(k) = \frac{\left|X(k)\right|^2}{\sum_{k=0}^{K}\left|X(k)\right|^2} , \tag{8}$$

where $k = 0, ..., K/2$;

– limit the speech spectrum frequency range (0-4000 Hz) to the range from 250 Hz to 3750 Hz and add some heuristic rules, namely if frequency component $k$ is below 250 Hz and above 3750 Hz then $X(k) = 0$; if $P(k) \geq 0{,}9$ then $P(k) = 0$;

– compute the negative entropy

$$H = \sum_{k=0}^{K/2} P(k)\log(P(k)) ; \tag{9}$$

– compute the EE feature as

$$M = (E - C_E)(H - C_H), \tag{10}$$

$$EE = \sqrt{(1 + |M|)},$$

where $C_E$ and $C_H$ are the average energy and entropy of the first 10 frames respectively.

In this algorithm a smoothing in time with linear method is additionally included – the same procedure as this one for the MD feature.

## 5. Experiments

We carried out series of experiments that can be separated into two groups. The aim of these experiments is to perform graphically and to evaluate visually the trajectories of analyzed features. These features are the currently proposed MD feature and the energy-entropy based one – EE feature, proposed in [6]. During the first group of experiments, we used selected noise-corrupted speech samples from the SpEAR database [8]. In the second group of experiments, we utilized noisy telephone speech samples from the BG-SRDat corpus [9].

In order to make a correct comparison between MD and EE features we have to compute them in the same frequency range. We selected the range accepted in [6], i.e. from 250 Hz to 3750 Hz. The obtained trajectories in all experiments are normalized in the range from zero to one to allow direct comparison between them.

### 5.1. Experiments No1

The SpEAR database was created for performance evaluation of different speech enhancement systems in adverse/noisy environments. The SpEAR database contains samples of noise-corrupted speech that have been recorded by acoustically combining clean speech and noise (with a clock-synchronous procedure for time-alignment).

We selected one record from Lombard section and one from noisy speech recordings section in the SpEAR database. All records have clear reference and corresponded noisy speech samples with different Signal-to-Noise Ratio (SNR). All selected wave files are with sampling frequency of 16 kHz at 16 bits, PCM format and mono mode [8].

The record from the Lombard section contains speech corrupted with factory noise that is recorded in a car production hall. The SNR of clean reference is 27.28 dB and for noisy recording SNR = – 9.96 dB;

The record from the noisy section contains speech corrupted with white noise. It is acquired by sampling high-quality analog noise generator. The SNR of clean reference is 40 dB (no noise) and for noisy recording SNR = 2.37 dB;

In Fig. 3 the factory noise speech example from the Lombard section in SpEAR database and the corresponded trajectories of the MD and EE features are shown . In the figure are included: (a) the clean speech sample, (b) noisy version of (a), (c) MD feature contour for noisy speech in (b) and (d) EE feature contour for noisy speech in (b).

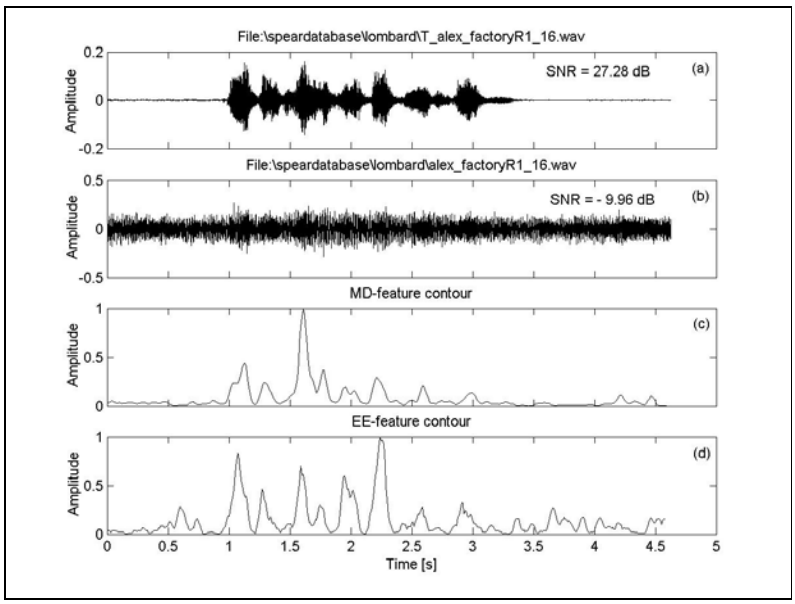


Fig. 3. Examples from the SpEAR database: a) clean speech sample, b) noisy version of a) with factory noise, c) MD feature contour for noisy speech in b) and d) EE feature contour for noisy speech in b)

In Fig. 4 are shown the white noise speech example from the noisy recording section in SpEAR database and the corresponded trajectories of the MD and EE features. In the figure are included: (a) the clean speech sample, (b) noisy version of (a), (c) MD feature contour for noisy speech in (b) and (d) EE feature contour for noisy speech in (b).

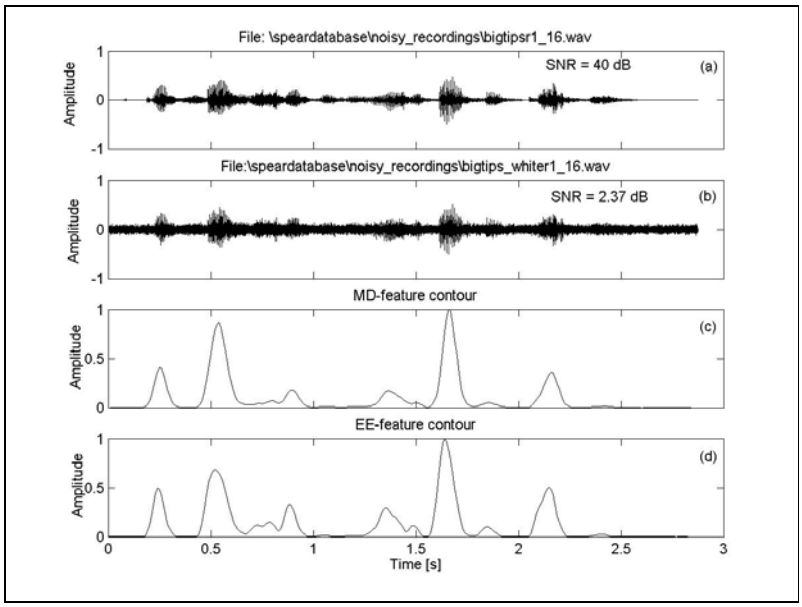


Fig. 4. Examples from the SpEAR database: a) clean speech sample, b) noisy version of a) with white noise, c) MD feature contour for noisy speech in b) and d) EE feature contour for noisy speech in b)

## 5.2. Experiments No 2

The BG-SRDat is a corpus in Bulgarian language recorded over noisy analog telephone channels and intended for speaker recognition. To achieve more realistic real-world conditions the speech data is collected by different types of telephone calls (internal-routing, local and long-distance) and various acoustical environments (noisy offices, halls and streets). The corpus does not provide the clean speech reference and its noisy version. It comprises only real-world noisy speech records. The speech data included in the BG-SRDat are sampled with frequency of 8 kHz at 16 bits, PCM format and mono mode [9].

We selected one record, which is typical of the BG-SRDat. It distinguishes for the presence of high-level pulse noise (due to the apparatus commutations in analog telephone line). In Fig. 5 are shown the noisy speech example from the BG-SRDat corpus in (a), its spectrogram in (b) and the corresponded trajectories of the MD feature in (c) and EE feature in (d).

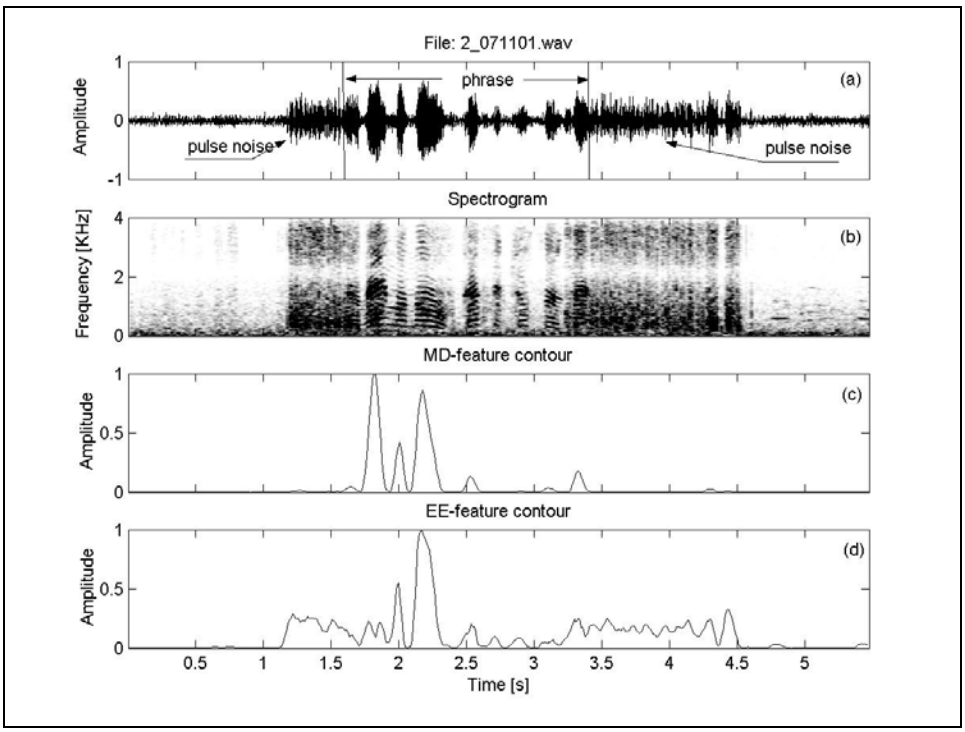


Fig. 5. An example from the BG-SRDat corpus: (a) noisy speech sample, (b) the spectrogram of (a), (c) MD feature contour for noisy speech in (a) and (d) EE feature contour for noisy speech in (a)

## 6. Discussion

In series of experiments with noisy speech data, we compared the trajectories variations of the two features mentioned above in the text. Our goal was to use these features in speech detection tasks where there is not any information about the position of noise

and speech fragments in the analyzed data. This limitation gets our experiments much closer to the real-world environment tasks. In this case, we can utilize the feature trajectory's characteristics to make a speech/non-speech decision. We decided to use the feature value (contour level) as measure for the presence of speech. This is a so-called "energy-type" approach for speech detection. It based on the assumption that the low levels in feature's contour correspond to the non-speech frames or frames with consonants and the high levels ones – mainly to the voiced or semi-voiced frames. Moreover, we can use a set of thresholds (fixed or adaptive) to separate the speech and non-speech events. Our visual analysis has to answer the question – which one of the presented here features is more suitable for using in the mentioned above energy-type approach?

The results from experiments No 1 with noisy samples from the SpEAR database are shown in Fig. 3 and Fig. 4 and they revealed interesting facts. The contours of two compared features are similar only in Fig. 4 where the added noise is white and SNR is not very low (2.37 dB). In this case, the energy-type approach can be used in speech detection tasks for both features. In the rest of experiments, the obtained contours of the EE feature demonstrate unsatisfactory results. Based only on the EE feature's values it is difficult to make reliable decision about the place of the speech and non-speech parts in the analyzed data, e.g., see Fig.3 (d), around time axis tick 0.5 s and after time axis tick 3.5 s. And vice versa, the trajectory of the MD feature at these time axis positions (see Fig. 3(c)) allows finding out the non-speech fragments in easier way.

The results from experiments No 2 are shown in Fig. 5. The main reason to include this speech data in the study was to observe the behaviour of both features for speech contaminated with a high-level pulse noise. It can be observed in Fig.5 (d), that the pulse noise has a strong effect on the EE contour. In the part of the EE feature trajectory (e.g., between time axis ticks 3.5 s and 4.5 s) where there is only a pulse noise the level is higher than this one in middle of the phrase (around time axis tick 2.5 s) where there is only a speech. The MD feature contour does not perform in the same way – it is more robust to the pulse noise as seen in Fig. 5 (c).

We carried out additional experiments (with data from both databases) where the EE and MD features are computed from the full-range power spectrum (from 0 Hz to the Nyquist frequency). For both features, the obtained results are worse than the presented once. However, the change for the worse is more significant for the EE feature than the MD one.

The using of spectral entropy as feature in speech detections tasks is based on the assumption that the spectrum is more organized in the speech regions than in the noise ones. The entropy is maximal when the signal is white noise and minimal when it is pure tone. It is known that the entropy-based speech detection algorithms are more suitable for speech signal in white or quasi-white noise and provide poor results for colored noise [14]. It is reasonable to combine the entropy and another speech parameter in order to compensate individual drawbacks as is done for the EE feature. However, this feature is obtained in [6] by an *ad hoc* procedure and in consequence of that in some cases the individual parameters drawbacks dominate in the final feature.

It is known that the SACF is an effective feature for voiced frames detection [3, 4]. In addition, as is shown in Fig. 1 and Fig. 2, for the unvoiced sounds the SACF possesses more clearly peaks than for the white noise. The delta processing on the SACF enhances the peaks and eliminates the influence of the mean of autocorrelation

function on the final MD feature. Additional restriction of the lags range ($L_1$ and $L_2$ in (5)) reduces the close to zero lag part (this part depends mainly on the spectral envelope shape) effect on the MD feature. As a result, we obtain a feature which is more robust than the EE feature especially in the cases where the speech signal is corrupted with pulse or colored noise.

## 7. Conclusions

In this study, we analyzed two features intended for speech detection algorithms. The first one proposed in [6] is formed by integration both of the spectral entropy and the signal energy (EE feature). The authors in [6] utilized this feature for endpoints detection in driving car noisy environment. Their algorithm uses the EE feature and a set of thresholds and durations to find the frames where the speech message starts and where it ends.

The second proposed feature is the mean of the absolute values of the delta spectral autocorrelation function of the power spectrum (named MD feature). We performed visual evaluation on the graphical representations of the EE and MD features for noisy speech samples selected from two sources – the SpEAR database and the BG-SRDat corpus.

Based on experimental results we made the following conclusions:

– the behaviour of both features depends on the type of noise. This dependence is more significant for the EE feature. In our study, the best results (for both features) are obtained when the speech is corrupted with white noise. The worst ones are for EE feature when the noise is pulse;

– the properties of both features also depend on the frequency range of the power spectrum used for their computation. For noisy speech it is not recommended (for both features) to use the spectral bins above 4 kHz, regardless of the fact that the MD feature is less affected by this part of spectrum;

– the EE feature is an integrated entropy-energy feature. The additional analysis revealed that the direct using of the speech energy as it is done in [6] is not a reliable way to create an integrated feature, especially in the cases when the speech is corrupted with pulse noise and the SNR is very low;

Our further work will include the development of the MD feature-based speech detection algorithm. We will evaluate this algorithm in the context of speaker recognition system, in order to estimate the efficiency of this new feature as a component of a complete system.

## References


1. Bendiksen, A., K. Steiglitz. Neural networks for voiced/unvoiced speech classification. – ICASSP, 1990, 521-524.
2. Ganapathiraju, A., L. Webster, J. Trimble, K. Bush, P. Kornman. Comparison of energy-based endpoint detectors for speech signal processing. – In: Proceedings of the IEEE Southeastcon, Tampa, Florida, USA, 1996, 500-503.
3. Hong, Kook Kim, Hwang Soo Lee. Use of spectral autocorrelation in spectral envelope linear prediction for speech recognition. – In: IEEE Transactions on SAP, Vol.7, 1999, No 5, 533-541.

4. L a h a t, M., R. J. N ie d e r j o h n, D. A. K r u b s a c k. A spectral autocorrelation method for measurement of the fundamental frequency of noise corrupted speech. – IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-35, 1987, 741-750.
5. J u n q u a, J.-C., B. M a k, B. R e a v e s. A robust algorithm for word boundary detection in the presence of noise. – IEEE Transactions on SAP, vol.2, July 1994, No 3, 406-412.
6. L i a n, J. Z h e n g H u a n g, C h u n g - h o Y a n g. A novel approach to robust speech endpoint detection in car environment. – ICASSP'2000, 1751-1754.
7. L i, Q., J. Z h e n g, A. T s a i, Q. Z h o u. 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.
8. The SpEAR Database.
**http://cslu.ece.ogi.edu/nsel/data/index.html**
9. O u z o u n o v, 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.
10. R a m ir e z, J, J. C. S e g u r a, C. B e ni t e z, A. de la T o r r e, A. R u b io. Efficient voice activity detection algorithms using long-term speech information. – Speech Communication, Vol. 42, April 2004, No 3-4, 271-287.
11. S h in, W., B. L e e, Y. L e e, J. L e e. Speech/non-speech classification using multiple features for robust endpoint detection. – ICASSP'2000, 1399-1402.
12. S o o n g, F., A. E. R o s e n b e r g. On the use of instantaneous and transitional spectral information in speaker recognition. – In: IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 36, 1988, No 6, 871-879.
13. Z h u, J., F. C h e n. The analysis and application of a new endpoint detection method based on distance of autocorrelated similarity. – In: Eurospeech'99, 105-108.
14. R e n e v e y, P., A. D r y g a j l o. Entropy based voice activity detection in very noisy conditions. – In: Eurospeech'01, 1883-1886.

## Един параметър за детекция на говор

*Атанас Узунов*

*Институт по информационни технологии, 1113 София*
*E-mail: atanas@iinf.bas.bg*

(Р е з ю м е)

В работата е предложен нов параметър, предназначен за използване в алгоритмите за откриване (детекция) на говор. Този параметър е средната стойност на модула на делта спектралната автокорелационна функция, получена от спектъра на мощността на анализирания сигнал и е наречен Mean-Delta (MD) параметър. В работата експериментално е сравнено изменението на траекторията на предложения параметър с тази на Entropy-Energy (EE) параметъра, основаващ се на спектралната ентропия. Реализирани са експерименти със зашумен говорен материал от две бази данни – SpEAR и BG-SRDat .

Получените резултати показват, че MD параметърът се влияе в значително по-малка степен от вида и нивото на шума в сравнение с EE параметъра.