BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **5**, No 2

Sofia • 2005

An Experimental Comparative Study of Three Robust Features for Speech Detection[1]

Atanas Ouzounov

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

Abstract: *The results from an experimental comparative study of three robust features intended for trajectory-based speech detection are presented in the paper. These features are the Mean-Delta (MD) feature [6], the Spectral Entropy (SE) [3] and the Spectral Entropy with Normalized frame Spectrum (SENS) [7]. Two experiments with noisy speech samples from two databases (the SpEAR database [2] and the BG-SRDat corpus [5]) are carried out. In the first experiment, the trajectory's variations of the features are compared by visual evaluation on their graphical representations. In the second one, the noise influence on the features trajectories is estimated by computing of the Euclidean distances between z-normalized trajectories of clean speech examples and their noisy versions. Based on experimental results two main conclusions are made: in comparison with other two features the MD feature trajectories are significantly less influenced by different type of noises; the SENS and especially the MD feature are more suitable for trajectory-based speech detection than the SE.*

Keywords: *speech detection, voice activity detection, spectral entropy.*

1. Introduction

When the speech or speaker recognition systems operate in noisy environment, it is often necessary to determine the speech and non-speech fragments in the analyzed signal. The speech segments provide data for speech or speaker model estimation, while the noise parameters estimated in the non-speech segments are used to compensate the influence of the noise on the recognition performance.

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

The finding of speech fragments in a given signal has many names, of which some are speech detection, endpoints detection, voice activity detection, and speech/ non-speech segmentation [4].

The algorithms for automatic speech detection can be divided into two general categories. The first one includes the algorithms that 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. The second category is comprised of algorithms based on a pattern recognition technique. In these algorithms, reference models for two classes (i.e., speech and non-speech) are created during the training phase based on selected speech features. In the classification phase, each segment is associated with one of the classes based on a selected similarity measure [3, 4, 9].

The selection of features intended for speech detection is usually 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 is usually 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 [3, 4, 7].

In last few years, the often-used features for speech detection in noisy environment are based on the spectral entropy characteristics [1, 3, 7]. In this case, the main assumptions are, firstly, the signal spectrum is more “organized” in the speech regions than in the noise ones and secondly the Shannon’s entropy can be used as an appropriate measure of signal organization [7].

In the paper, we study experimentally two different kinds of features intended for trajectory-based speech detection – one feature based on spectral autocorrelation and two others based on spectral entropy. The spectral autocorrelation-based feature is the mean-delta feature [6] while the spectral entropy-based features are the spectral entropy [3] and the spectral entropy with normalized frame spectrum [7].

Two experiments are carried out with different noisy speech examples. In the first experiment, the trajectory’s variations of the features are compared by visual evaluation on their graphical representations. In the second one, the noise influence on the features trajectories is estimated by computing of the Euclidean distances between z-normalized trajectories of clean speech examples and their noisy versions.

2. Robust features

2.1. The Mean-Delta feature

The Mean-Delta (MD) feature is proposed in [6] and it is defined as the mean of the absolute values of the delta spectral autocorrelation function of the power spectrum of speech signal. Let $x(i)$ is a discrete speech signal, where $i = 0, ..., I - 1$, I is the number of samples and the spectrum $X(k)$ of $x(i)$ is obtained by the Discrete Fourier Transform (DFT), where $k = 0, ..., K/2$, K is the number of points in the DFT.

The spectral autocorrelation function $R_p(l)$ is defined with the power spectrum as [6]

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

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

The Delta Spectral AutoCorrelation Function (DSACF) is the first-order derivative of the spectral autocorrelation function obtained by a polynomial approximation in a manner similar to the delta cepstrum evaluation [6]. For particular frame it is computed using only frame's spectral autocorrelation lags (intra-frame processing).

For the n-th frame the DSACF $\Delta R_p(n, l)$ is computed as

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

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; $n = 0, ..., N - 1$, and N is the number of frames.

For n-th frame the MD feature $m_d(n)$ is computed as follows:

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

where $\Delta R_p(n, l)$ is the DSACF in (2) for lag l, L_1 and L_2 are the boundary lags and $\Delta L = L_2 - L_1 + 1$.

In this study, two minor changes are made to the basic MD feature estimation algorithm proposed in [6]. Firstly, the trajectory smoothing by a local maximal value is not applied and secondly, instead of $m_d(n)$ in (3), the square root of it is used as MD feature. These minor changes are done in order to compensate partly the extra trajectory smoothing for some low-level speech sounds as was observed in some preliminary experiments. For more details about the MD feature, see [6].

2.2. The spectral entropy

The Spectral Entropy (SE) for the n-th frame is estimated in the following steps [3]. First, the probability density function $P(|X(n, k)|^2)$ for the spectrum $|X(n, k)|^2$ is computed as

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

where $k = 0, ..., K/2$ and $n = 0, ..., N - 1$. The heuristic rule is added, namely if $P(|X(n, k)|^2) \geq 0.9$ then $P(|X(n, k)|^2) = 0$. After this constrain is applied, the spectral entropy $H_c(n)$ for n-th frame is computed as follows:

(5) $$H_c(n) = -\sum_{k=0}^{K/2} P\left(|X(n,k)|^2\right) \log_2\left(P\left(|X(n,k)|^2\right)\right).$$

The negative SE $H_c^-(n)$ is defined as $H_c^-(n) = -H_c(n)$. It is more convenient in the trajectory-based speech detection algorithms to be used the negative SE, especially when this entropy will be combined or will be compared with the energy-based features.

2.3. The spectral entropy with normalized frame spectrum

It is known that the entropy curve of the speech regions with colored noise is very similar to the entropy curve of the non-speech regions [7]. To make the speech detection with entropy feature under colored noise conditions more reliable, in [7] is proposed to divide the spectrum of each frame by the average spectrum computed over all frames of the analyzed speech data (i.e. to normalize the frame spectrum). If is the magnitude spectrum for the *n*-th speech frame, where *n* = 0, ..., *N* – 1; *k* = 0, ..., *K*/2 and *K* is the number of points in the DFT and *N* is the number of frames, so the normalized spectrum $\left|\tilde{X}(n,k)\right|$ is computed as follows:

$$\left|\tilde{X}(n,k)\right| = \frac{\left|X(n,k)\right|}{\frac{1}{N}\sum_{n=0}^{N-1}\left|X(n,k)\right|}. \tag{6}$$

The probability density function $P(\left|\tilde{X}(n,k)\right|^2)$ for the spectrum $\left|\tilde{X}(n,k)\right|$ is estimated by normalizing the frequency components

$$P\left(\left|\tilde{X}(n,k)\right|^2\right) = \frac{\left|\tilde{X}(n,k)\right|^2}{\sum_{k=0}^{K/2}\left|\tilde{X}(n,k)\right|^2}, \tag{7}$$

and the Spectral Entropy with Normalized frame Spectrum (SENS) $H_w(n)$ for *n*-th frame is computed as

$$H_w(n) = -\sum_{k=0}^{K/2} P\left(\left|\tilde{X}(n,k)\right|^2\right).\log_2\left(P\left(\left|\tilde{X}(n,k)\right|^2\right)\right). \tag{8}$$

The negative SENS $H_w^-(n)$ is defined as $H_w^-(n) = -H_w(n)$.

3. Experiments

We carried out series of experiments that can be divided into two groups. The aim of first group of experiments is to display graphically the trajectories of the analyzed features and evaluate visually how suitable they are for the trajectory-based speech detection. The second group of experiments is intended to provide a preliminary and rough estimation of the noise influence on the feature trajectories.

During the experiments, we used selected noise-corrupted speech samples from two speech databases – the SpEAR database [2] and the BG-SRDat corpus [5].

In order to make a correct comparison between different features we have to compute all of them in the same frequency range. We selected the range accepted in [3], i.e. from 250 Hz to 3750 Hz. In all experiments, the obtained trajectories are normalized in order to allow direct comparison between them. The frame length is 30 ms, the frame shift is 10 ms and the FFT-points are 1024. The mean spectrum in the denominator of (6) is estimated over entire analyzed speech phrase. All contours are smoothed by 3-points moving-average filter.

Hereafter in the text, the attribute “negative” will be omitted in all names of entropy measures for more convenience.

As a rough measure of the noise influence on the feature trajectory, we decided to use the similarity between feature trajectories of the clean speech record and its noisy version. This similarity, more exactly the shape similarity, can be estimated by the Euclidean distance between both normalized trajectories. The trajectory normalization is recommended in [8] in order to obtain the distance between trajectories that is invariant to the trajectories’ scaling and shifting. In the study, the *z*-normalization trajectory technique is applied [8]. In this case, the normalized trajectory has a zero mean and a unit standard deviation. It is possible to apply this simple technique because the analyzed trajectories are with equal lengths and there is not local time shifting along them.

The normalized trajectory $T_N(n)$, $n = 1, ..., N$, where N is the number of trajectory’s frames, is estimated as

(9)
$$T_N(n) = \frac{T(n) - m_T}{\sigma_T},$$

where m_T, σ_T are the mean and standard deviation, respectively, computed over entire trajectory $T(n)$.

The average spectrum in the denominator of (6) is computed over all frames in the analyzed phrase. For some signals, this average spectrum can be small for certain frequencies. This fact leads to significant variations in normalized spectrum and further in the entropy estimation. To overcome this effect the authors in [7] recommend adding a white noise with small amplitude to the signal before the spectrum computation. During the experiments, only for the SENS calculation, white Gaussian noise is added to the analyzed signal. In this case, the achieved Signal-to-Noise Ratio (SNR) is 20 dB. This additional noise smoothes the SENS trajectory as was found in some preliminary experiments.

3.1. Experiment No 1

We selected three examples from “Lombard Speech” section and two others from “Noisy Speech Recordings” section in the SpEAR database. All examples have clean speech reference and corresponded noisy versions (time-aligned) with different SNR. All selected wave files are with sampling frequency of 16 kHz at 16 bits per sample, PCM format and mono mode [2].

The examples from the “Lombard Speech” section are:

- factory noise example – it contains speech corrupted with factory noise recorded in a car production hall. For the clean reference SNR = 27.28 dB and for its noisy version SNR = –9.96 dB;
- car noise example – it contains speech corrupted with noise recorded inside a driving car (Volvo 340). For the clean reference SNR = 27.00 dB and for its noisy version SNR = –14.58 dB;
- pink noise example – it contains speech corrupted with pink noise; the noise is acquired by sampling the signal from a high-quality analog noise generator. For the clean reference SNR = 21.23 dB and for its noisy version SNR = –10.33 dB.

The examples from the “Noisy Speech Recordings” section are:

- white noise example – it contains speech corrupted with white noise. The noise is acquired by sampling the signal from a high-quality analog noise generator;

for the clean reference SNR = 40 dB (no noise) and for its noisy version SNR = 2.37 dB;

• bursting noise example – it contains speech corrupted with bursting noise; the noise is computer generated using a white Gaussian random number generator; for the clean reference SNR = 40 dB (no noise) and for its noisy version SNR = 0.16 dB.

In Figs. 1-5 are shown the noisy speech examples from SpEAR database and the corresponded z-normalized trajectories of the SE, SENS and MD feature. The factory noise, the car noise and the pink noise examples are shown in Figs. 1-3, respectively. The white noise and bursting noise examples are shown in Figs. 4 and 5.

The BG-SRDat is a corpus in Bulgarian language collected over analog telephone lines and intended for speaker recognition. The speech data included in the BG-SRDat are sampled at 8 kHz with accuracy 16 bits, PCM format and mono mode [5].

We selected one speech data file, which is typical of the BG-SRDat. In this file, there are some segments with high-level pulse noise and some others with low-level harmonic noise probably due to the crosstalk. The BG-SRDat corpus comprises only real-world noisy speech records and it does not provide clean speech examples and their noisy versions as SpEAR database [2]. In order to obtain a clean reference for selected noisy speech data file a wave editing and a noise reduction technique are applied. This additional processing is done here only for illustration purposes only.

In Fig. 6 are shown the noisy speech example, its clean version, the corresponded *z*-normalized trajectories of the analyzed features and the results from manual speech detection task.

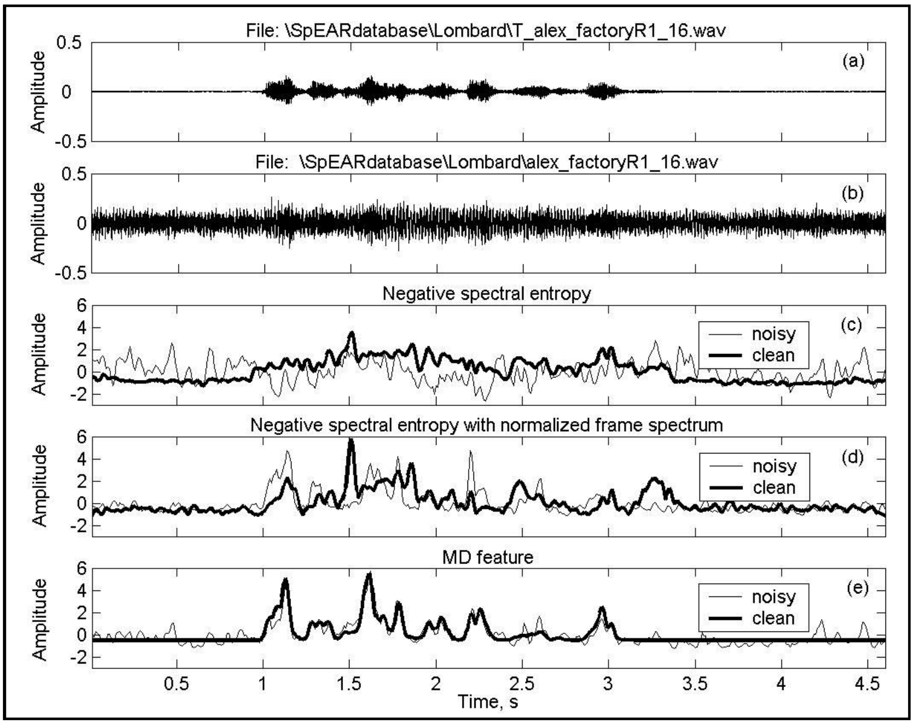

Fig. 1. Examples from the SpEAR database: (a) – clean speech sample; (b) – noisy version of (a) with factory noise; (c) – SE trajectories for speech samples in (a) and (b); (d) – SENS trajectories for speech samples in (a) and (b); (e) – MD feature trajectories for speech samples in (a) and (b)

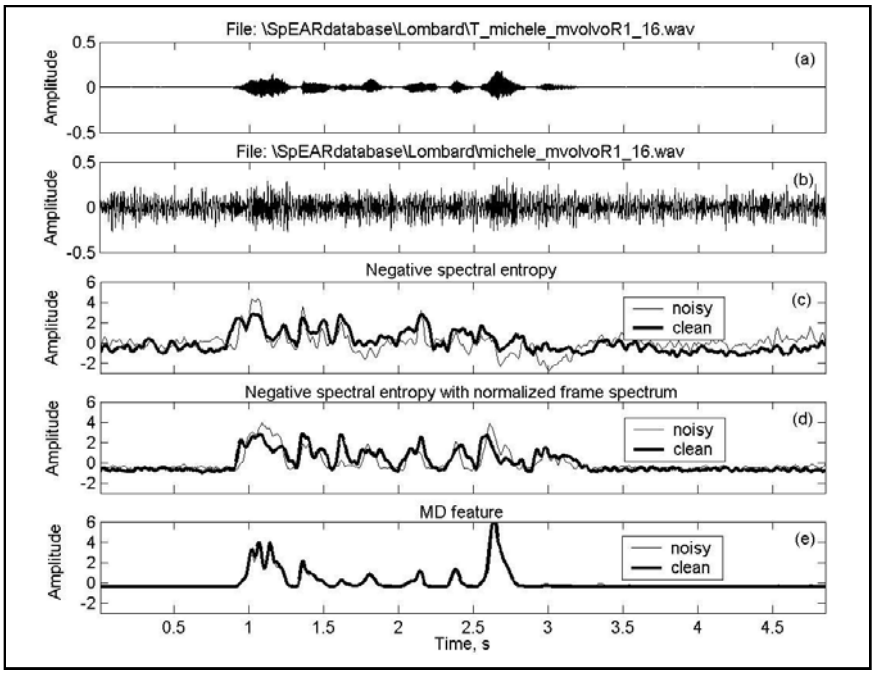

Fig. 2. Examples from the SpEAR database: (a) – clean speech sample; (b) – noisy version of (a) with car noise; (c) – SE trajectories for speech samples in (a) and (b); (d) – SENS trajectories for speech samples in (a) and (b); (e) – MD feature trajectories for speech samples in (a) and (b)

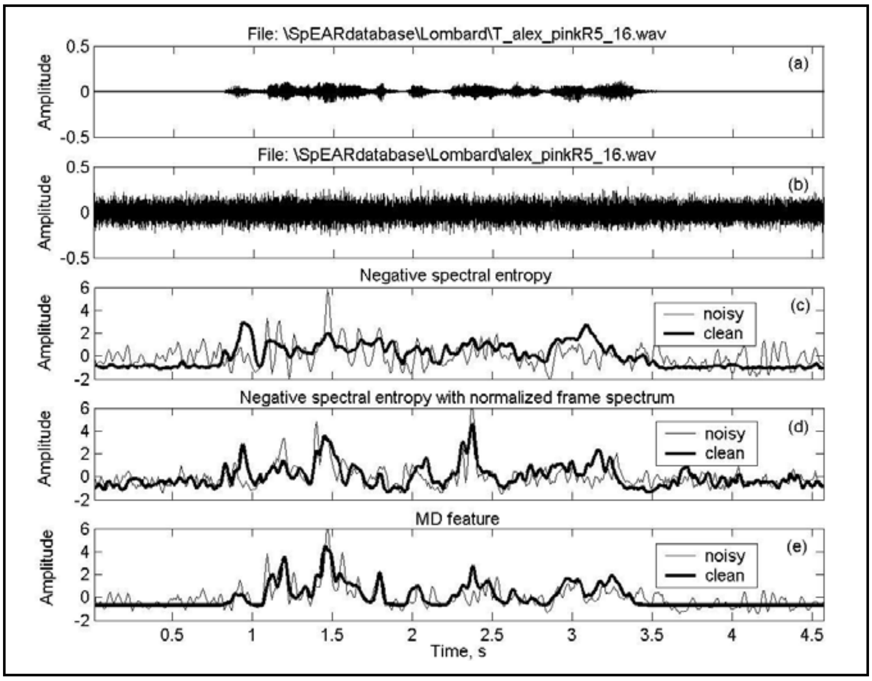

Fig. 3. Examples from the SpEAR database: (a) – clean speech sample; (b) – noisy version of (a) with pink noise; (c) – SE trajectories for speech samples in (a) and (b); (d) – SENS trajectories for speech samples in (a) and (b); (e) – MD feature trajectories for speech samples in (a) and (b)

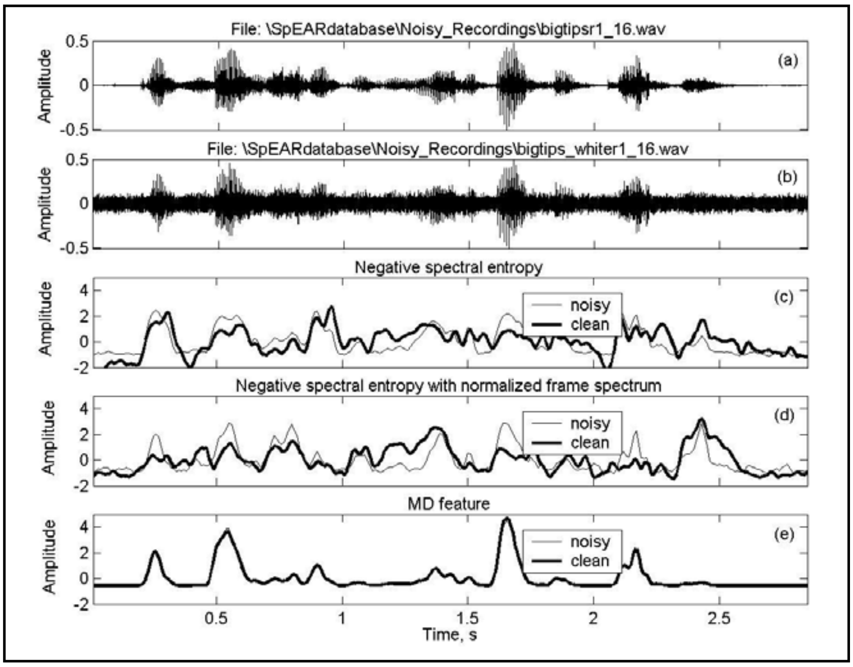

Fig. 4. Examples from the SpEAR database: (a) – clean speech sample; (b) – noisy version of (a) with white noise; (c) – SE trajectories for speech samples in (a) and (b); (d) – SENS trajectories for speech samples in (a) and (b); (e) – MD feature trajectories for speech samples in (a) and (b)

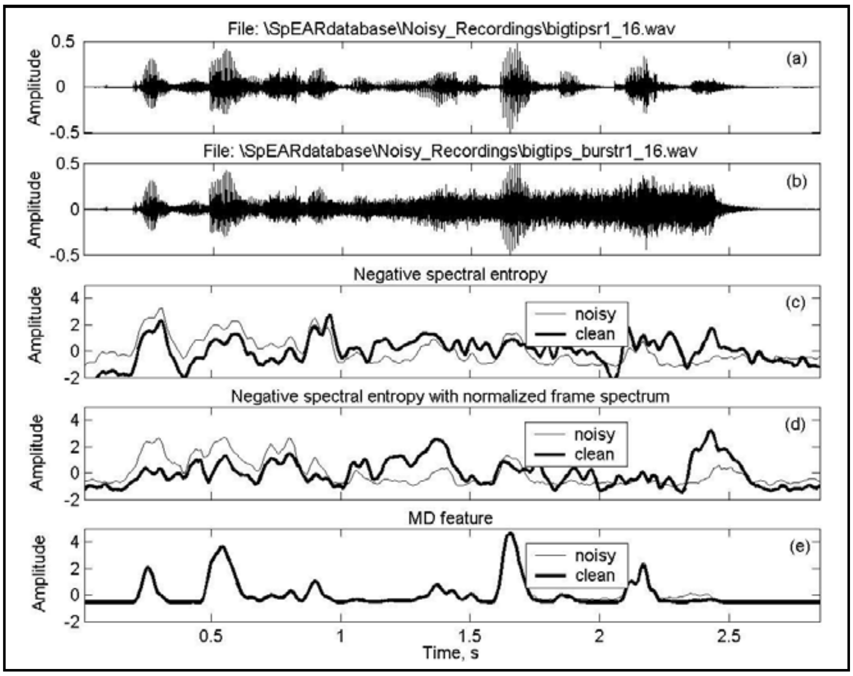

Fig. 5. Examples from the SpEAR database: (a) – clean speech sample; (b) – noisy version of (a) with bursting noise; (c) – SE trajectories for speech samples in (a) and (b); (d) – SENS trajectories for speech samples in (a) and (b); (e) – MD feature trajectories for speech samples in (a) and (b)

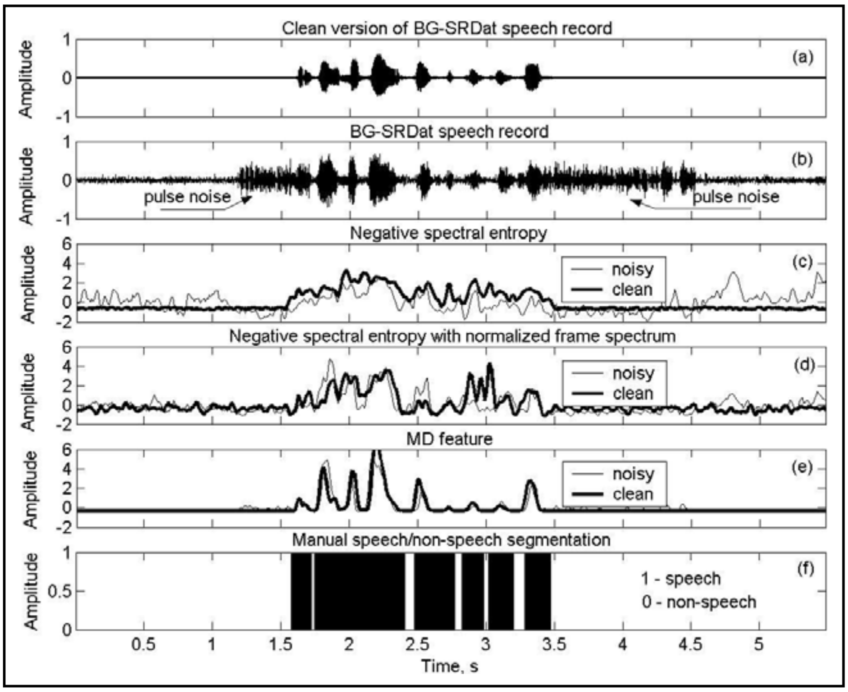

Fig. 6. An example from the BG-SRDat corpus: (a) – clean speech sample; (b) – noisy speech sample; (c) – SE trajectories for speech samples in (a) and (b); (d) – SENS trajectories for speech samples in (a) and (b); (e) – MD feature trajectories for speech samples in (a) and (b), (f) – manual speech detection results

3.2. Experiment No 2

The Euclidean distances between z-normalized trajectories of the clean speech examples and their noise versions for different features are shown in Table 1. The different features are noted in the table as follows: MD – Mean-Delta feature, SE – spectral entropy and SENS – spectral entropy with normalized frame spectrum.

Table 1. Euclidean distance between *z*-normalized trajectories of the clean speech examples and corresponded noisy versions

No	Features	SpEAR database examples					BG-SRDat example (Fig.6)
		Factory noise (Fig.1)	Car noise (Fig.2)	Pink noise (Fig.3)	White noise (Fig.4)	Bursting noise (Fig.5)	
1	MD	0.4972	0.1247	0.7161	0.0528	0.1603	0.5584
2	SE	1.3701	0.9509	1.1543	0.7976	1.0123	1.1701
3	SENS	1.0703	0.6926	0.8755	0.8953	1.1162	0.9043

4. Discussion

In the experiments with noisy speech data, we compared the trajectories of the MD feature with the trajectories of two spectral entropy-based features. We decided to use the frame feature value (trajectory level) as measure for the presence of speech.

This is a so-called “energy-type” approach for speech detection. It is based on the assumption that the low levels in feature’s trajectory correspond to the non-speech frames or frames with consonants and the high levels ones – mainly to the voiced or semi-voiced frames.

As can be seen in Figs. 1-4 – subplot (c), it is very difficult to make reliable decision (based only on the SE trajectory level) about the positions of the speech and non-speech parts in the analyzed data. On the contrary, the trajectories of the SENS and especially of the MD feature allow easily finding the speech and non-speech fragments – see Figs. 1-4 – subplots (d) and (e). The results shown in Fig. 5 are obtained with bursting noise. The varying noise amplitude partly complicates the utilizing of the SE and SENS contours for trajectory-based speech detection. Again, the MD feature performs itself very well.

In the MD feature trajectory can be noticed some segments with an extra smoothing, especially for fricative sounds. This effect can be clearly seen in Fig. 1 (e) (between time axis ticks 3.2 s and 3.4 s); in Fig. 2 (e) (between time axis ticks 2.9 s and 3.3 s) and in Fig. 4 (e) (between time axis ticks 2.3 s and 2.6 s).

The results obtained for speech example from BG-SRDat corpus are shown in Fig. 6. Again, the SE provides the worst result (see Fig. 6 (c) – between time axis ticks 4.7 s and 5.5 s – in this file position there are segments with low level harmonic noise probably due to the crosstalk), while the SENS and especially the MD feature allow easily finding the speech and non-speech fragments based only on trajectory levels.

The results shown in Table 1 reveal interesting fact – the MD trajectory shape is influenced by the different noises significantly less than the trajectories of the entropy features (the Euclidean distances for the MD feature are always the minimal). We suppose that this fact can facilitate the MD feature-based speech detection in the non-stationary noise environment (e. g. bursting noise).

5. Conclusions and future work

In the paper, three features intended for trajectory-based speech detection are analyzed. Two experiments are carried out with noisy speech samples selected from two databases. During the first experiment, the visual evaluation on features trajectories is done in order to estimate how suitable they are for “energy type” speech detection. A rough measure of the noise influence on the feature trajectory is computed during the second experiment. This measure is based on the Euclidean distance between z-normalized trajectories of the clean speech records and their noisy versions.

Based on experimental results the following conclusions are made:

- the behaviour of features’ trajectories depends on the type of noise – this dependence is more significant for the spectral entropy-based features;
- in comparison with other two features the MD feature trajectories are significantly less influenced by different type of noises – see Table 1;
- in most cases the SE feature is not suitable for trajectory-based speech detection;
- the SENS is promising feature but it is more influenced by non-stationary noises (as bursting noise) than the MD feature;
- in the MD feature trajectory can be noticed an extra smoothing, especially for the fricative speech sounds.

Our further work will include the development of an integrated feature-based speech detection algorithm (e.g., a combination of the MD feature and SENS). 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. Basu, S. A Linked-HMM model for robust voicing and speech detection. – ICASSP 2003, Vol.**1**, I-816-I-819.
2. 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**
3. Liang-sheng Huang, Chung-ho Yang. A novel approach to robust speech endpoint detection in car environment. – ICASSP'2000, 1751-1754.
4. 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.
5. 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.
6. Ouzounov, A. Robust feature for speech detection. – Cybernetics and Information Technologies, Vol. 4, 2004, No 2, 3-14.
7. Renevey, Ph., A. Drygajlo. Entropy based voice activity detection in very noisy conditions. – EUROSPEECH'01, 2001, 1883-1886.
8. Goldin, D. Q., P. C. Kanellakis. On similarity queries for time-series data: Constraint specification and implementation. – CP'95, 137-153.
9. Shin, W., B. Lee, Y. Lee, J. Lee. Speech/Non-speech classification using multiple features for robust endpoint detection. – ICASSP'2000, 1399-1402.