Telephone Speech Endpoint Detection by Using Robust Features

Atanas Ouzounov
atanas@iinf.bas.bg

Signal Processing & Pattern Recognition Department
IICT-BAS, Sofia, Bulgaria

SEMINAR PRESENTATION
November 11, 2014

INTRODUCTION

This presentation summarizes the research carried out by the author mainly described in the following publications:

14. **Ouzounov A.:** Telephone Speech Endpoint Detection using Mean-Delta Feature, *Cybernetics and Information Technologies,* vol.14, No.2, pp.127-139 (2014).

15. **Ouzounov A.:** Noisy Speech Endpoint Detection Using Robust Feature, Springer International Publishing Switzerland 2014, V. Cantoni et al. (Eds.): BIOMET 2014, LNCS 8897, pp. 105–117 (2014).

INTRODUCTION

- **What is Endpoint Detection (ED)**
  - Location of the beginning and the ending points of speech utterance.

- **Importance of ED**
  - The ED is the key component in speech and speaker recognition systems designed to operate in noisy real-world environments.

- **Current ED Algorithms**
  - based on the features' contour analysis [3, 6, 7, 20];
  - based on the pattern recognition techniques [17, 19, 22, 23].

- **Current ED features**
  - different energy transformations of the speech signal [3, 4];
  - autocorrelation functions [24];
  - spectral entropy [5, 6, 20, 30];
  - others features – bispectrum [9], wavelets [18], etc.

INTRODUCTION

- **The Presentation Goal**

  To summarize the research in the field of the contour-based telephone speech endpoint detection, which include:

  - development of new robust features for ED: the FFT magnitude spectrum-based Mean-Delta feature [14] and the Group Delay Mean-Delta feature [15].
  - evaluation of the endpoint detection accuracy of the proposed features and two additional ones – the modified Teager energy [4] and the energy-entropy feature [5];
  - estimation of the effect of the analyzed ED features in the Dynamic Time Warping fixed-text speaker verification task with short noisy telephone phrases in Bulgarian language [14, 15].

- **The Mean-Delta (MD) feature [12, 14]**

The MD feature is proposed by the author in [12] and is defined as the mean absolute value of the Delta Spectral AutoCorrelation Function (DSACF) of the speech spectrum. For a particular frame, the DSACF was computed utilizing only the frame's spectral autocorrelation lags and it is obtained in a way similar to the delta cepstrum evaluation - an orthogonal polynomial fit of the first-order derivative (in correlation domain). For the *n*th frame, the DSACF $\Delta R_P(n,l)$ is

$$\Delta R_P(n,l) = \frac{\sum_{q=-Q}^{Q} qR_P(n,l+q)}{\sum_{q=-Q}^{Q} q^2} \qquad (1)$$

where $l=0,\ldots,L$ is the number of correlation lags; $n=0,\ldots,N-1$, N is the number of frames; Q is the delta window and $R_P(.)$ is the biased Spectral AutoCorrelation Function (SACF) defined with the power [12] or with the magnitude spectrum [14]. For n^{th} frame the MD feature $m_d(n)$ is computed as follows

$$m_d(n) = \left[\sum_{l=0}^{L} \left| \Delta R_P^S(n,l) \right| \right]^{0.5} \qquad (2)$$

where $\Delta R_P^S(n,l)$ is the contour smoothed DSACF for lag l. This smoothing is obtained by the Long-Term Spectral Envelope (LTSE) algorithm, applied not on the spectral envelop as is in [16] but on the DSACF. For more details, see [14].

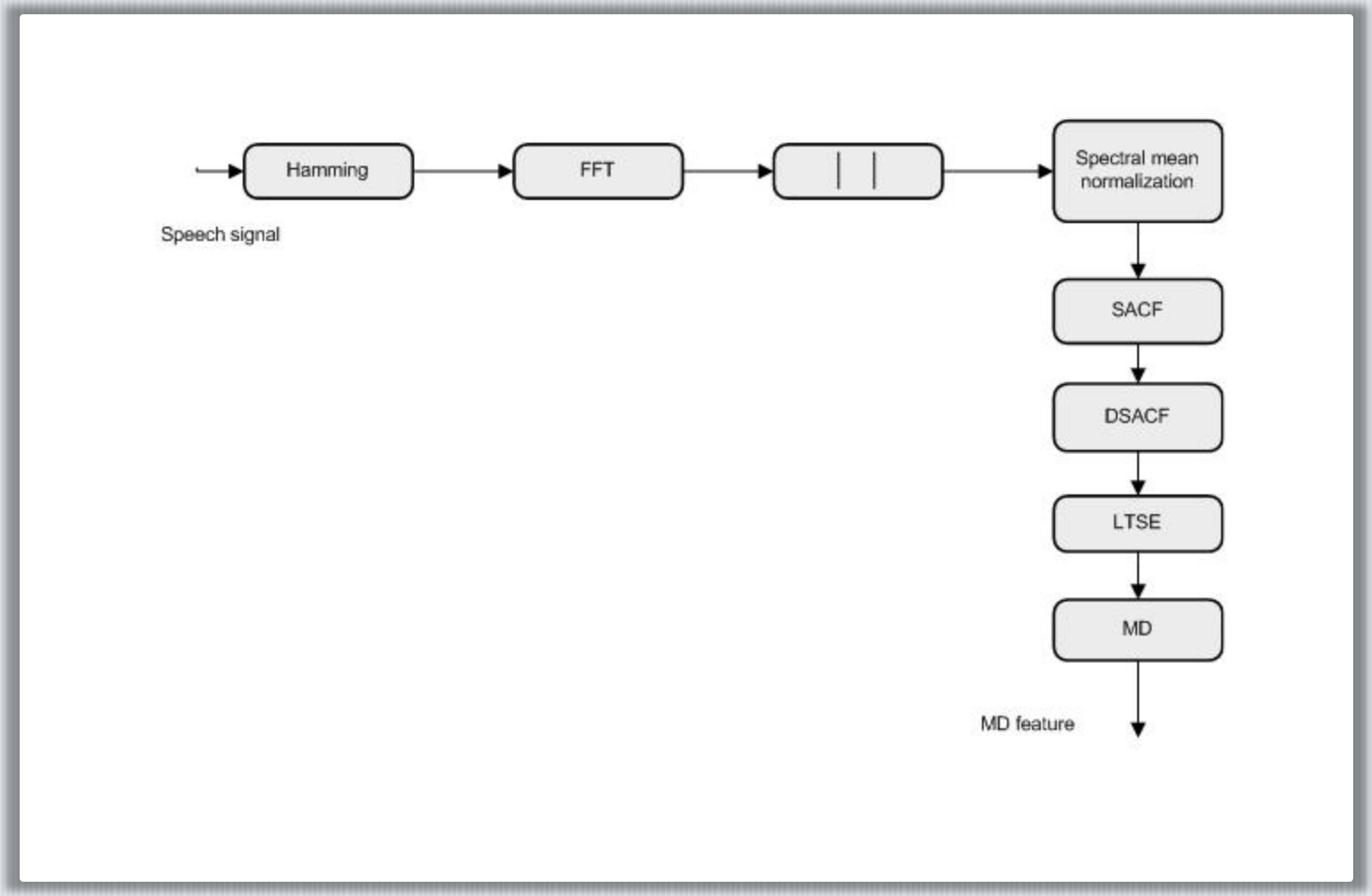

Fig.1. Block diagram of the MD algorithm.

The examples of three typical speech signals, their normalized SACFs and the corresponding DSACFs [12]

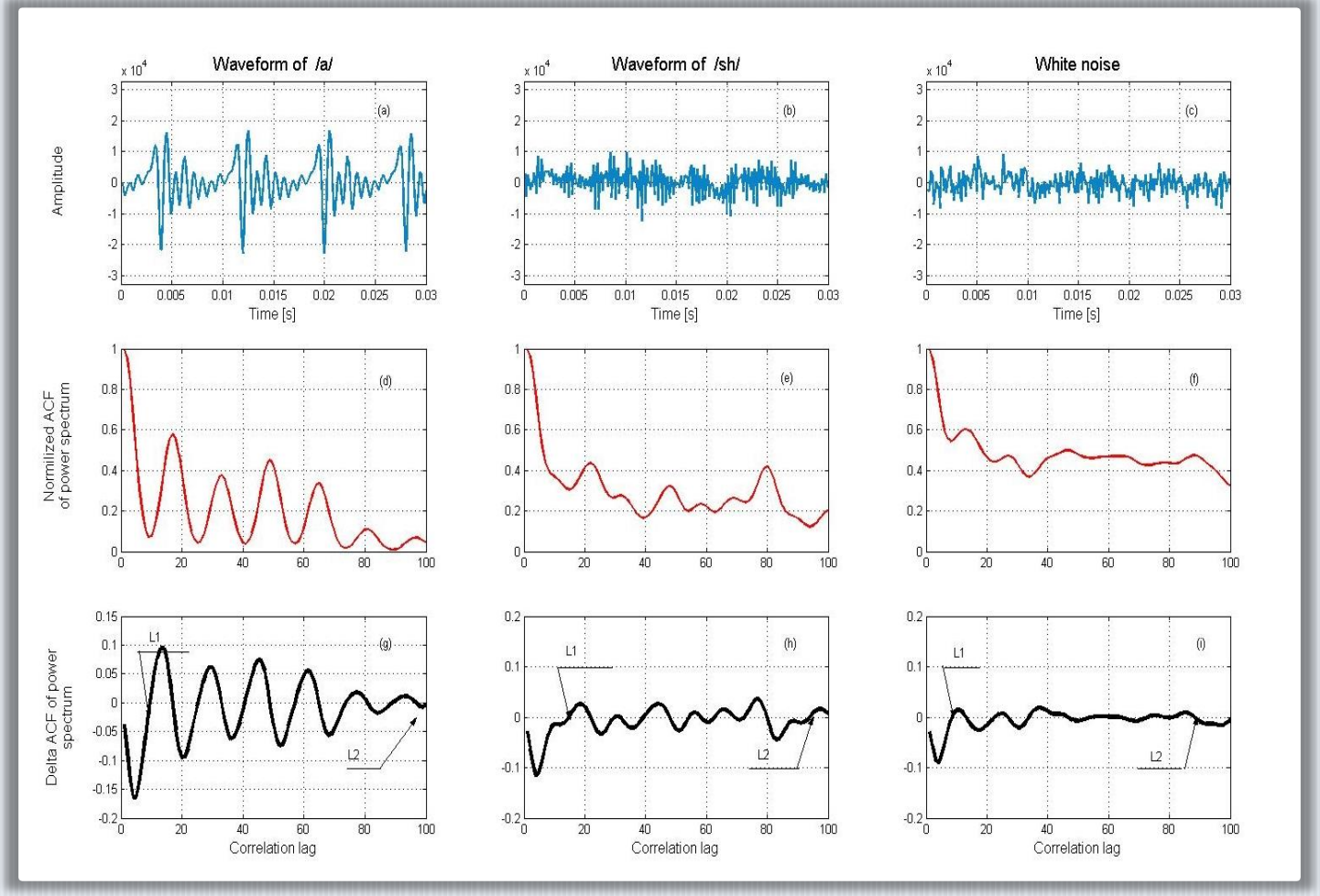

Fig.2 . Examples of SACFs and corresponding DSACFs: (a) voiced speech signal /*a*/; (b) unvoiced speech signal /*sh*/; and (c) white noise [12].

- **The Modified frame Teager Energy (MTE) feature [4]**

The algorithm for MTE feature calculation includes for each frame the following steps:

- calculate the power spectrum;
- weight each sample in the power spectrum with the square of the frequency;
- take the square root of the sum of the weighted power spectrum.

The MTE feature $E_t(n)$ for n^{th} frame is

$$E_t(n) = \left[\sum_{k=0}^{K/2} (k\Delta f)^2 \left| X(n,k) \right|^2 \right]^{0.5} \qquad (3)$$

where Δf is the frequency resolution, $\left| X(n,k) \right|^2$ is the FFT power spectrum and K is the FFT size;

- **The Energy Entropy (EE) feature [5]**

The EE feature is obtained by combination of the energy and the spectral entropy and for n^{th} frame is defined as

$$EE(n) = \sqrt{(1 + |E(n)H(n)|)} \tag{4}$$

where $E(n)$ is the frame energy and $H(n)$ is the frame spectral entropy.

$$E(n) = \sum_{i=0}^{I-1} x^2(n,i) \tag{5}$$

I - number of samples in the frame. The PDF $P(n,k)$ for the frequency k is

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

and $H(n)$ is

$$H(n) = -\sum_{k=0}^{K/2} P(n,k)\log(P(n,k)) \tag{7}$$

- **The GDMD feature**

The Group Delay Mean Delta (GDMD) feature is proposed by the author in [15]. This feature utilized the Mean Delta approach proposed in [12] but the spectral autocorrelation function is defined based on the Modified Group Delay Spectrum, not on the magnitude spectrum [14]. The aim of this is to obtain peak-enhanced delta spectral autocorrelation function and thereafter more effective Mean Delta feature.

The Modified Group Delay Spectrum (MGDS) $\tau_m(k)$ is proposed in [25, 26] and is defined as:

$$\tau_m(k) = sign.\left|\frac{X_R(k)Y_R(k)+Y_I(k)X_I(k)}{S(k)^{2\gamma}}\right|^{\alpha} \qquad (8)$$

where:

- *sign* is given by the sign of the term in the absolute value brackets;
- x(n) is the given speech frame;
- X(.) and Y(.) are the Fourier transforms of the sequences x(n) and nx(n);
- $k = 0, \cdots, K/2$; K is the FFT size;
- S(.) is the cepstrally smoothed spectrum of |X(.)| using low-order cepstral lifter l_w.
- α, γ and l_w are adjusted according to the particular requirements [25, 26].

The GDMD feature is computed in the same way as the MD one, but instead the FFT magnitude spectrum the MGDS is used – see Figs.1 and 4. In this case the biased spectral autocorrelation function $R_m(l)$ defined with the MGDS $\tau_m(k)$ is

$$R_m(l) = \sum_{k=0}^{K/2-l} \tau_m(k)\tau_m(k+l) \tag{9}$$

For the n^{th} frame the DSACF $\Delta R_m(n,l)$ according to (1) is

$$\Delta R_m(n,l) = \frac{\sum_{q=-Q}^{Q} qR_m(n,l+q)}{\sum_{q=-Q}^{Q} q^2} \tag{10}$$

For n^{th} frame the GDMD feature $m_{gd}(n)$ is computed as follows

$$m_{gd}(n) = \left[\sum_{l=0}^{L} \left| \Delta R_m^S(n,l) \right| \right]^{0.5} \tag{11}$$

where $\Delta R_m^S(n,l)$ is the smoothed DSACF for lag l obtained by the LTSE algorithm [16], applied on the DSACF. For more details, see [15].

Various spectra and corresponding parameters for a part of speech sound *'e'*

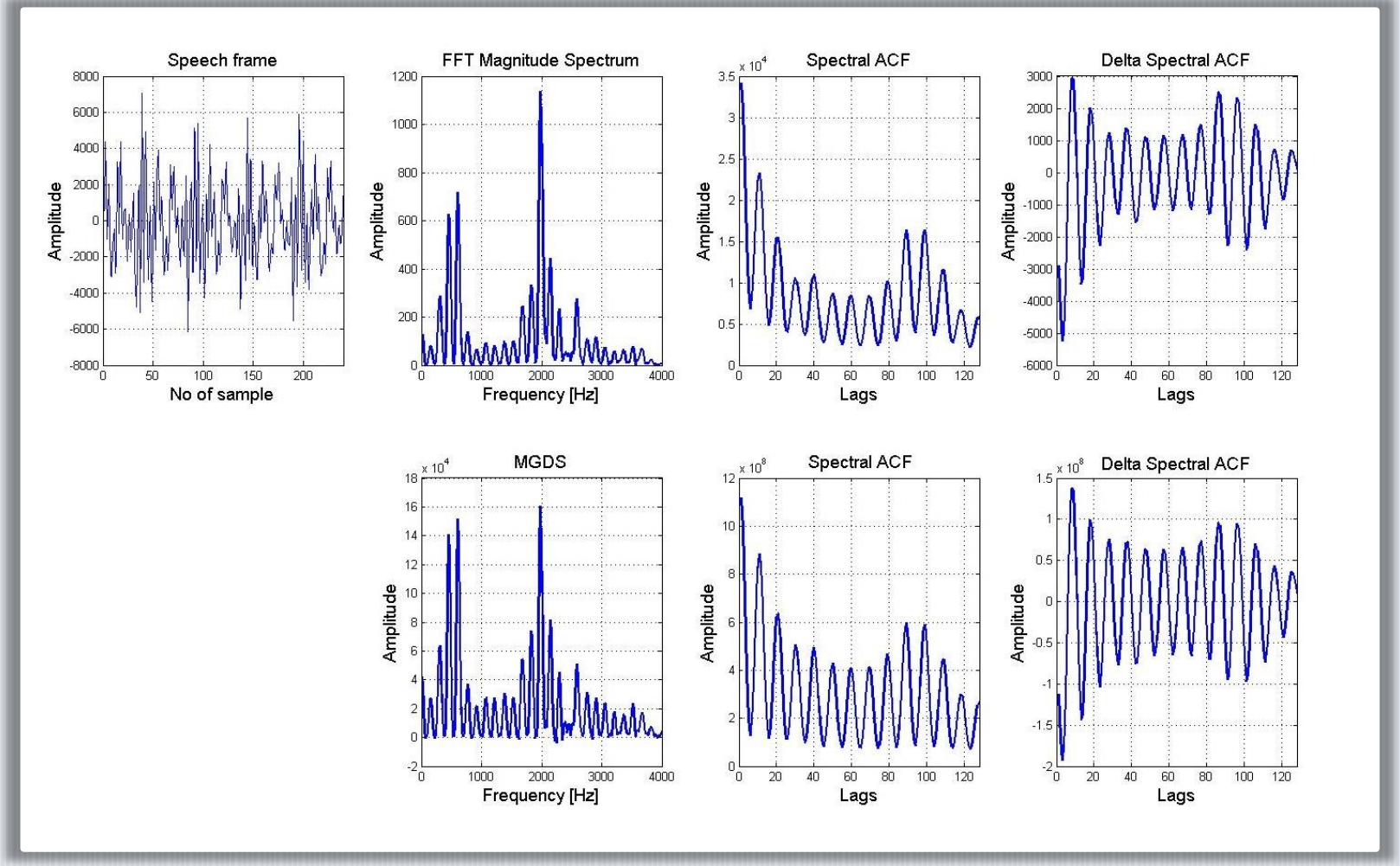

Fig.3. Various spectra and corresponding parameters for a part of speech sound *'e'*. The MGDS parameters are: α=0.6, γ=0.4, l_w = 32; [15].

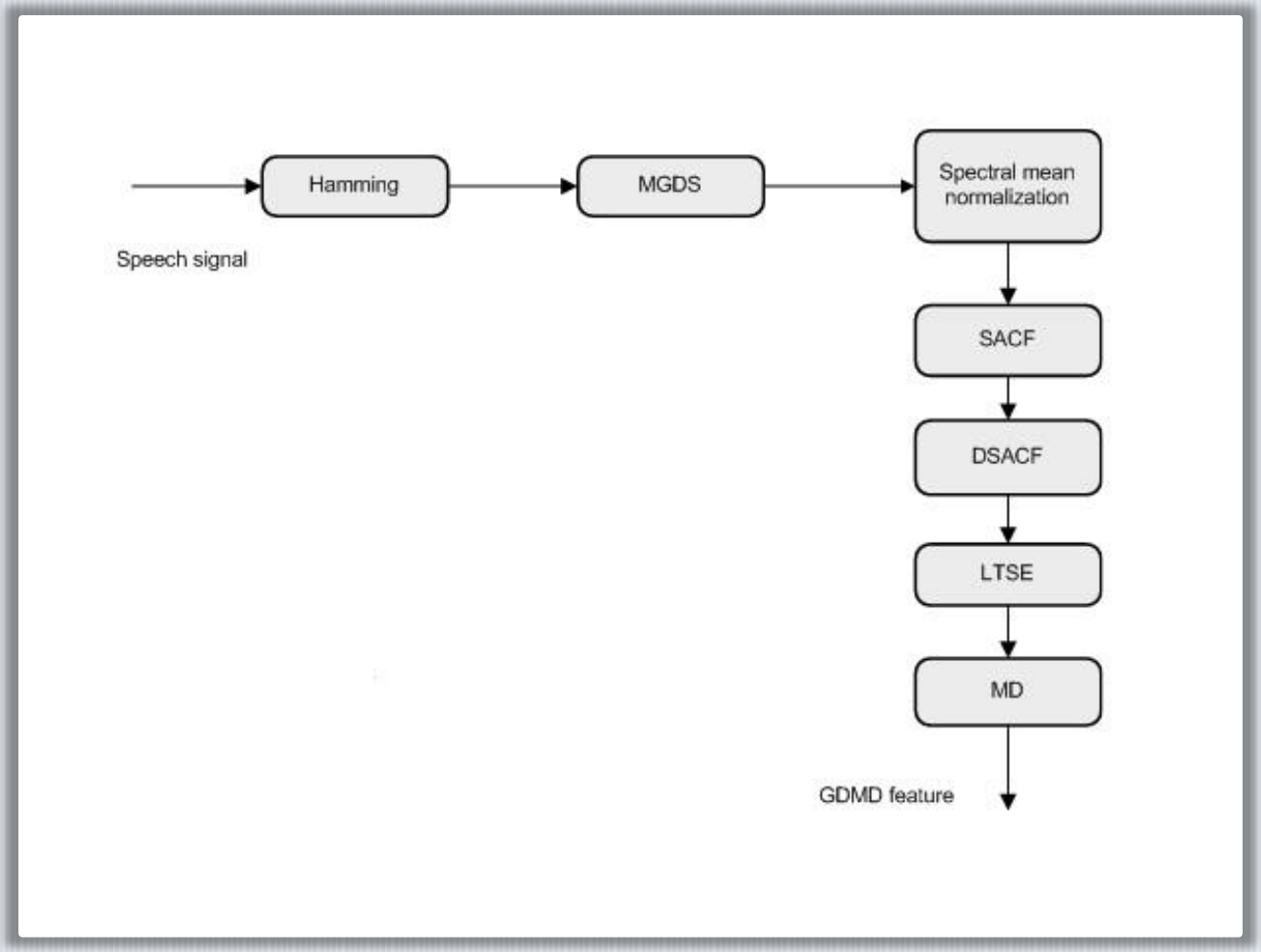

Fig.4. Block diagram of the GDMD algorithm

DETECTION ALGORITHM

- **The detection algorithm [14]**
  - based on the single parameter time contour;
  - designed for endpoint detection of a single word or a short utterance (few seconds);
  - uses two fixed thresholds and eight-state automaton;

The block diagram of the two thresholds setting algorithm is shown in Fig.5. For more details, see [14].

DETECTION ALGORITHM

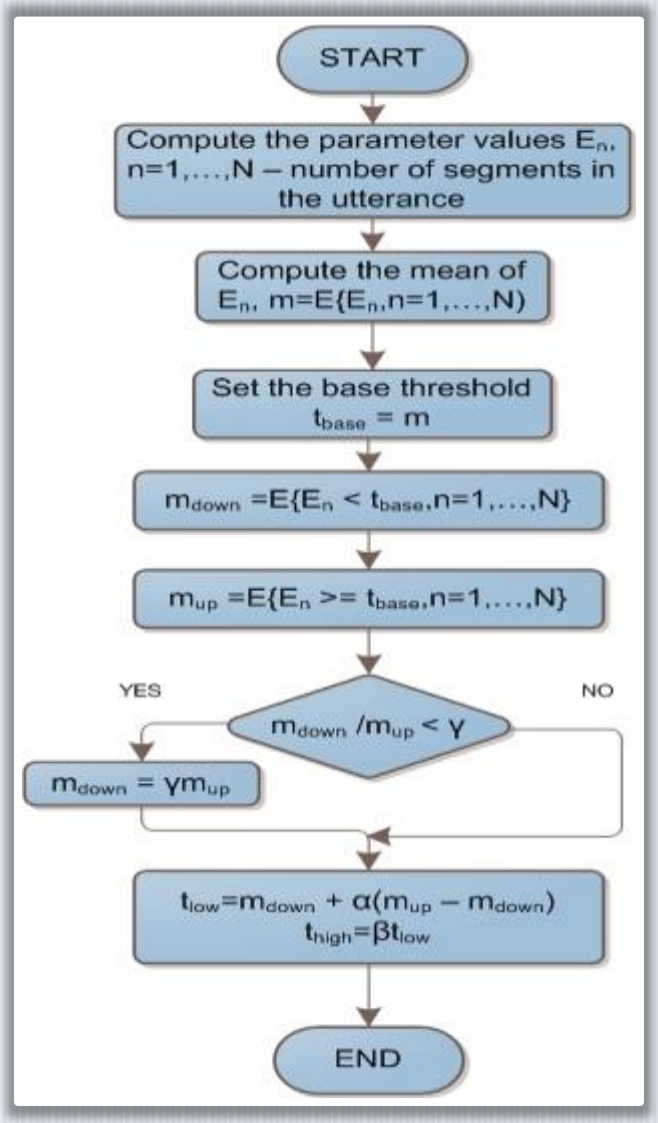

Fig.5. Two thresholds setting algorithm

EXAMPLES

The features' time-contours and thresholds for a noisy example from the NOIZEUS corpus [28, 29]

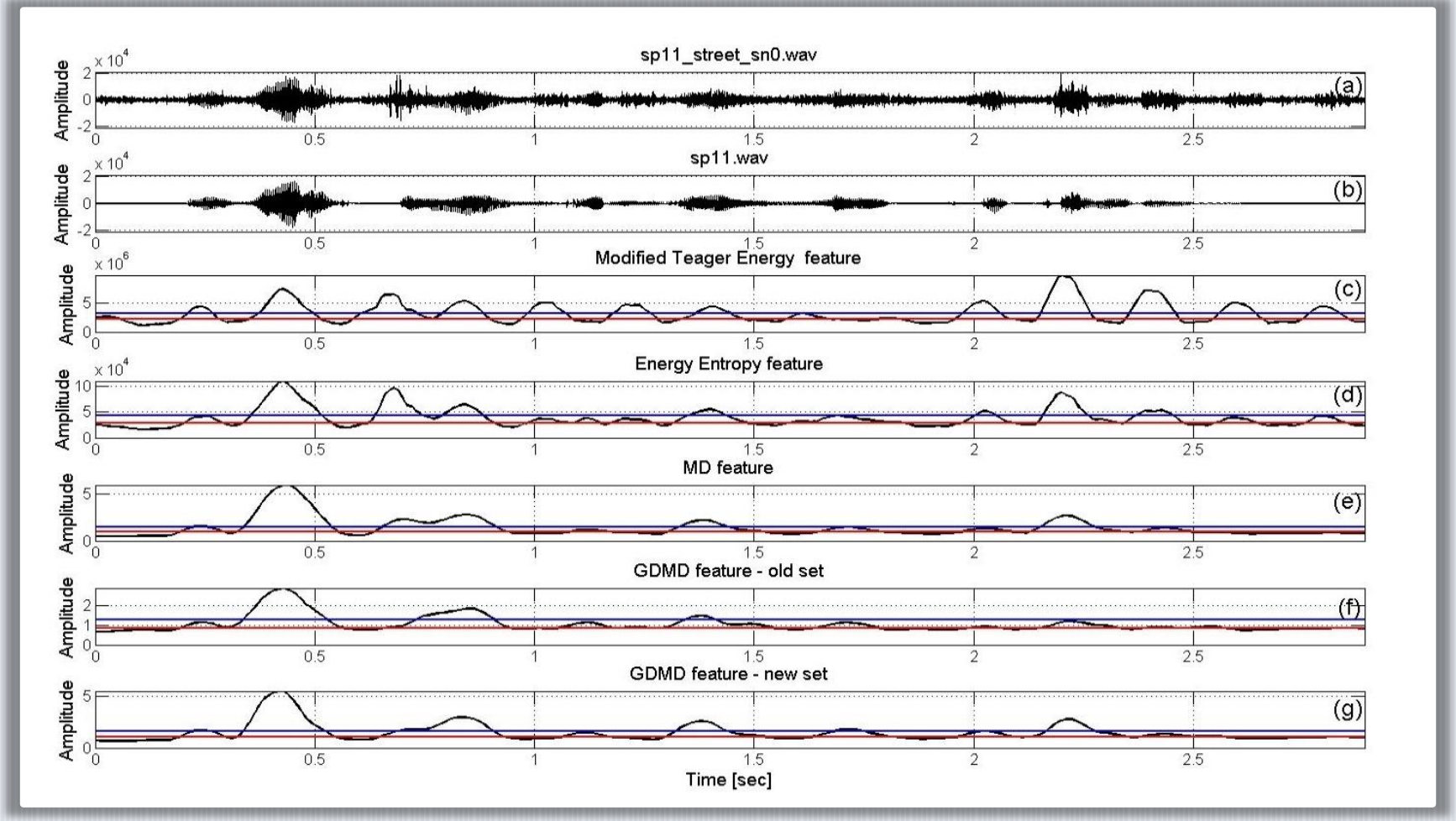

Fig.6. Examples from the NOIZEUS corpus: (a) noisy example; (b) the clean version; (c) modified Teager energy; (d) energy entropy feature; (e) MD feature; (f) GDMD feature – old set (α=0.4, γ=0.9, l_w = 8) [25, 26]; (g) GDMD feature – new set (α=0.6, γ=0.4, l_w = 32) [15] ;

FFT spectra of the noisy example from Fig.6

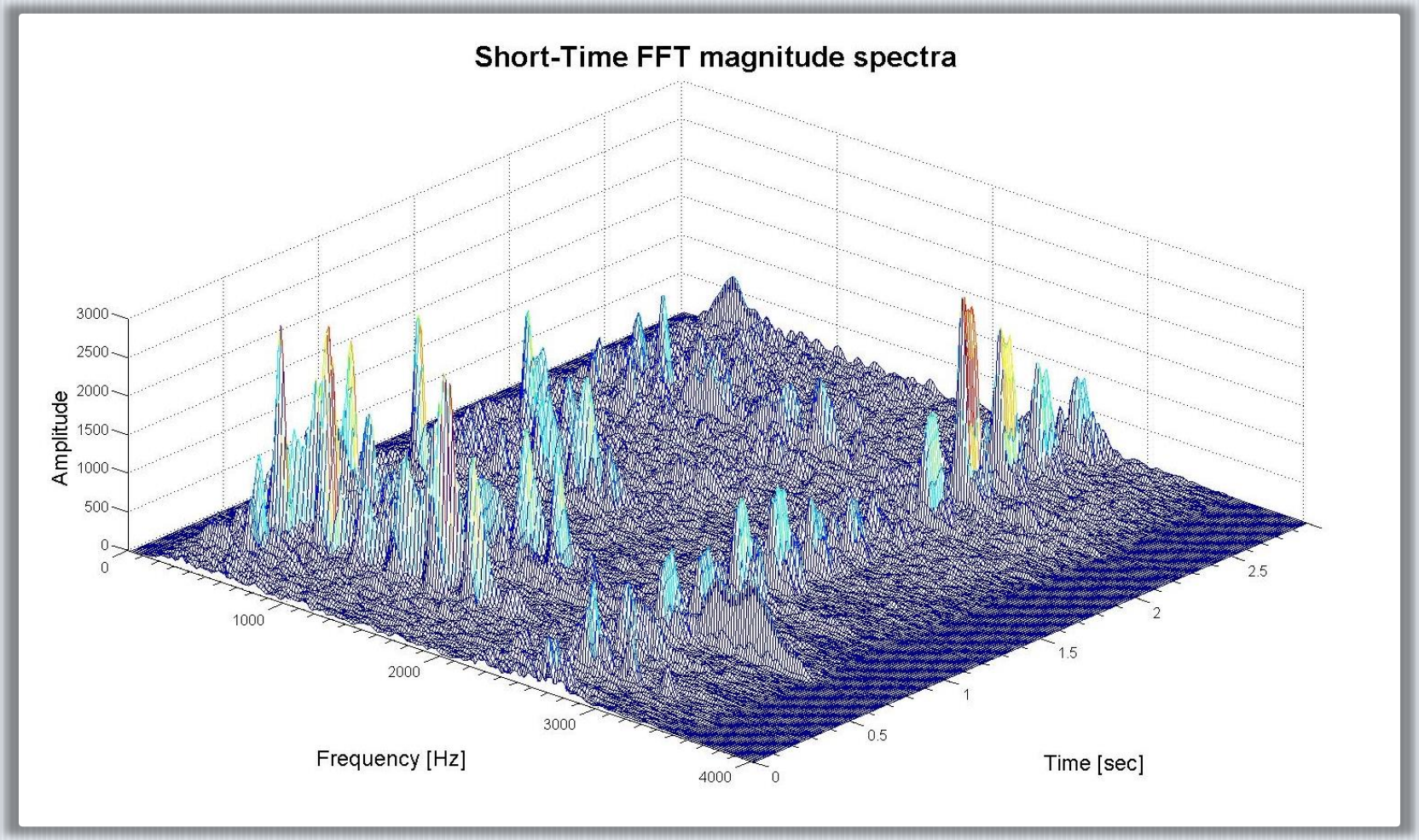

Fig.7. 3D mesh plot of the short-time FFT magnitude spectra.

MD feature of the noisy example from Fig.6

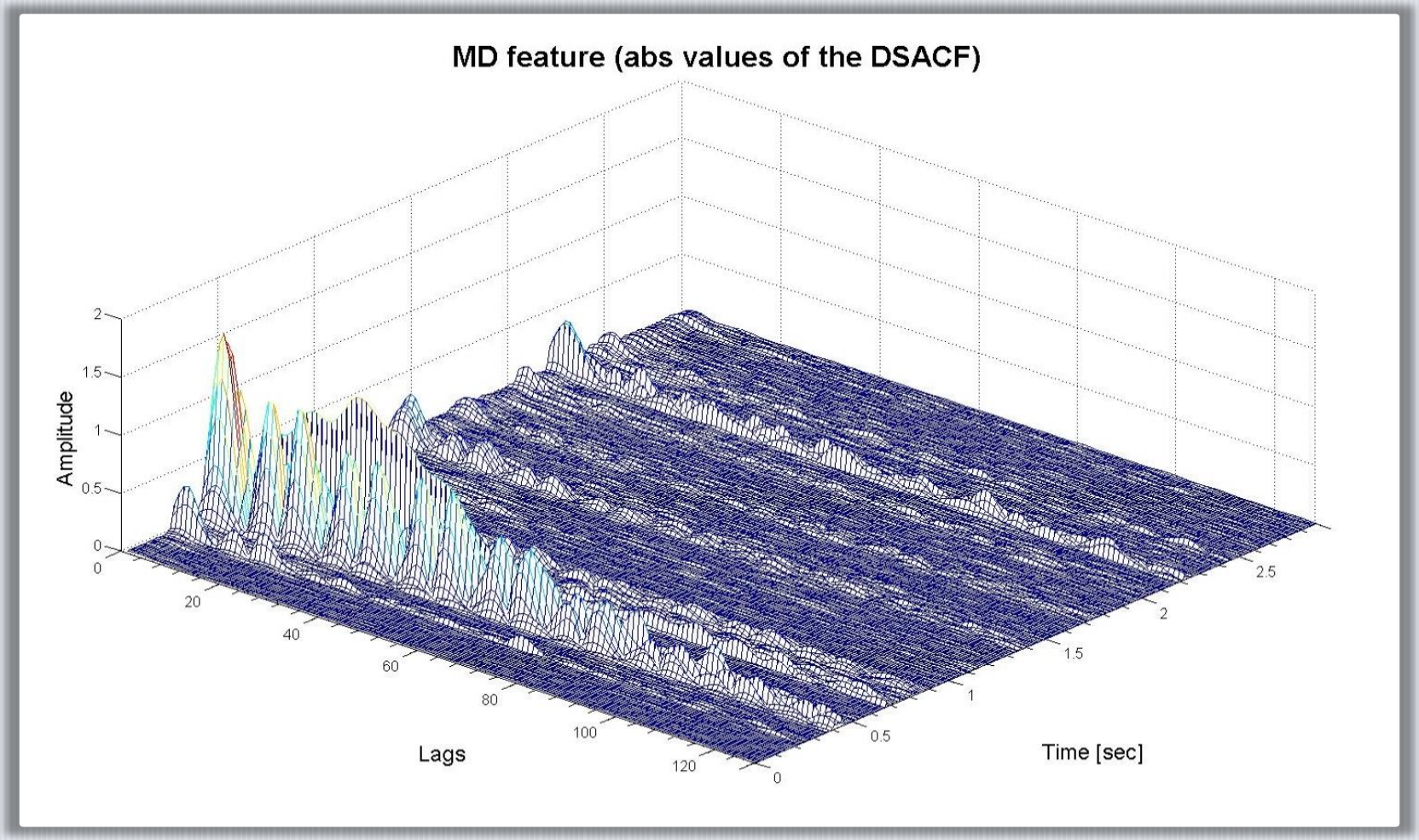

Fig.8. 3D mesh plot of the abs values of the DSACF of the FFT magnitude spectra.

GDMD feature of the noisy example from Fig.6

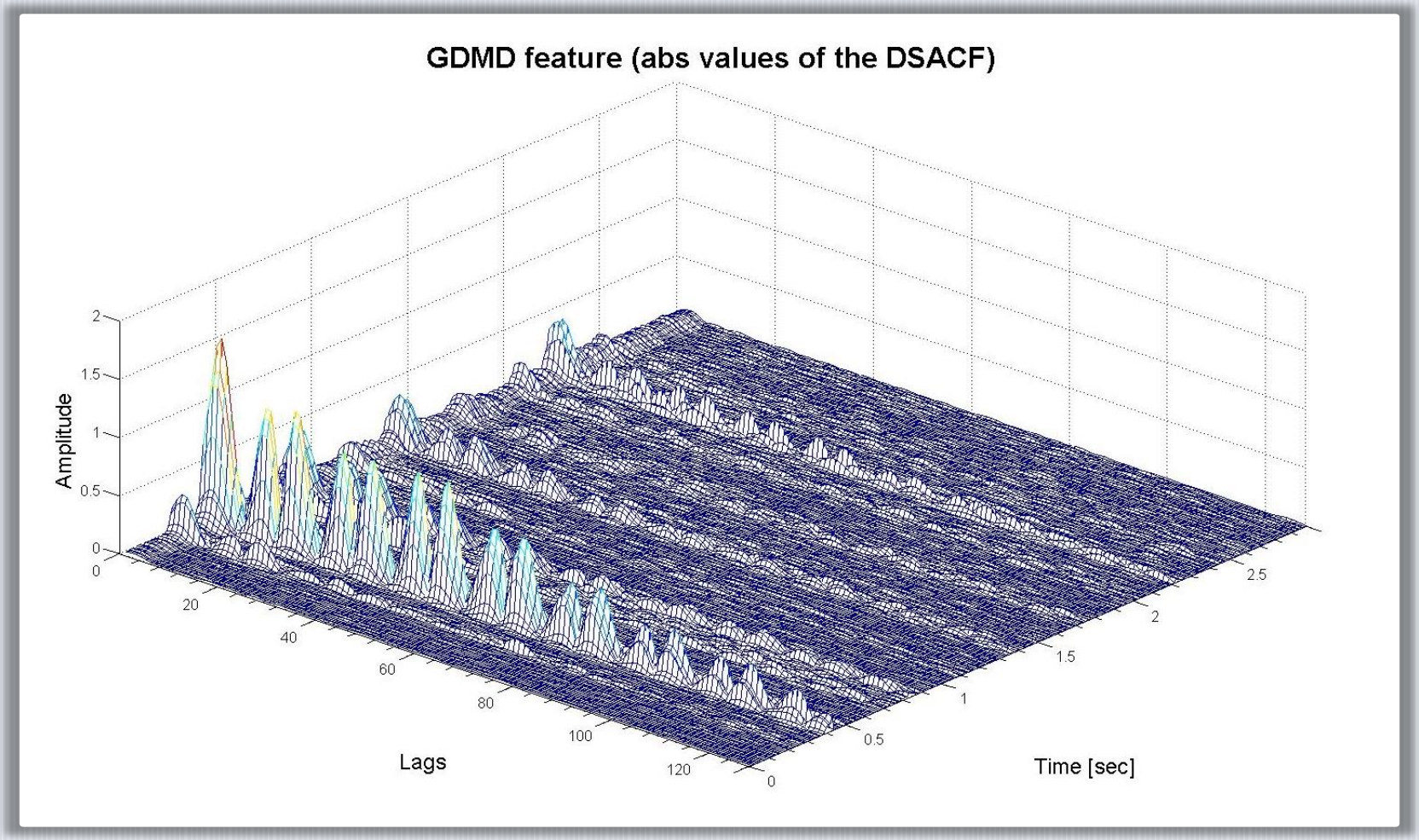

Fig.9. 3D mesh plot of the abs values of the DSACF of the MGDS with new set of parameters α=0.6, γ=0.4, l_w = 32 [15].

EXPERIMENTS

Two experiments were carried out [14, 15]:

- **Experiment No.1** - The ED accuracy was evaluated in terms of frame difference between manually labelled and detected endpoints [21].
- **Experiment No.2** - The ED features performance is evaluated in terms of speaker verification rate.
- **Experimental data -** Selected from the BG-SRDat corpus [11]:
  - Bulgarian language and telephone speech;
  - Recorded in real-world environment - street pay phones;
  - Single phrase - length is about 2 seconds;
  - Used data - 262 records of the utterances collected from 12 male speakers with different numbers of records per speaker – from 16 up to 34;
  - The phrase is (in Bulgarian): *Zdravei Manolov. Kak se chuvstvash dnes?;*
  - Its English meaning is : *Hello Manolov! How are you today?.*
  - The pronunciation (roughly) is: *[zdra`vei:] [ma`nolov]! [kak] [se] [`tʃuvstvaʃ] [dnes]?*
  - **Important note:** The phrase starts with voiced fricative *'z'* and ends with unvoiced fricative*'s'*.

EXPERIMENT No.1 – ED accuracy evaluation

The frames difference $D_B(s)$ between manually labelled and detected by algorithm beginning points is defined as

$$D_B(s) = M_B(s) - ED_B(s) \tag{12}$$

where $M_B(s)$ is the manual labelled beginning point; $ED_B(s)$ is the beginning point obtained by endpoint detection algorithm and $s=1\ldots S$ is the number of files.

The frames difference $D_E(s)$ for the ending points is

$$D_E(s) = M_E(s) - ED_E(s) \tag{13}$$

where $M_E(s)$ is manually labelled ending point and $ED_E(s)$ is the detected ending point.

The histograms for D_B and D_E for the ED features are shown in Figs.10 and 11 [15].

EXPERIMENT No.1 – ED accuracy evaluation

The histograms of D_B

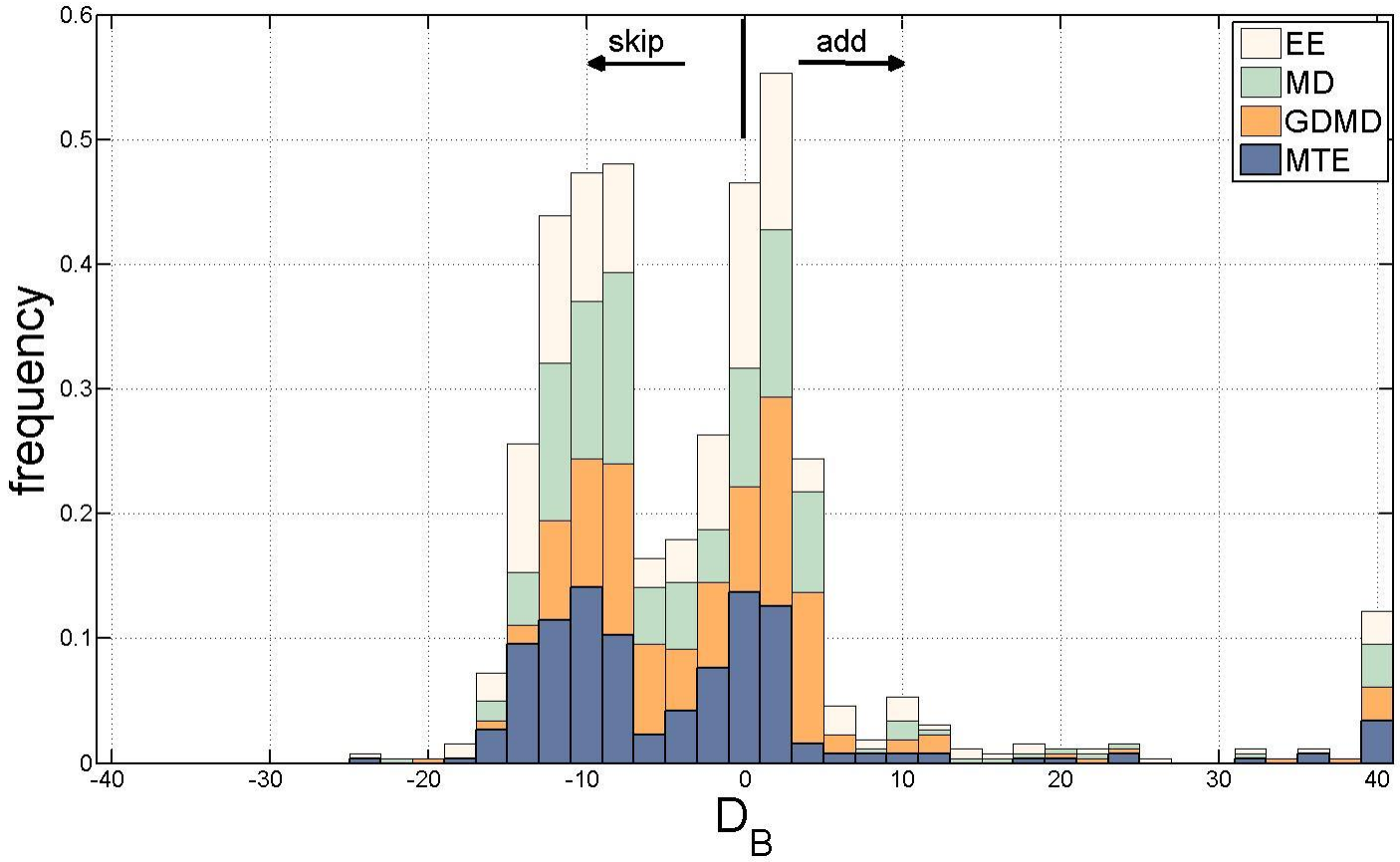

Fig.10. The histograms of the differences D_B for the ED features.

EXPERIMENT No.1 – ED accuracy evaluation

The histograms of D_E

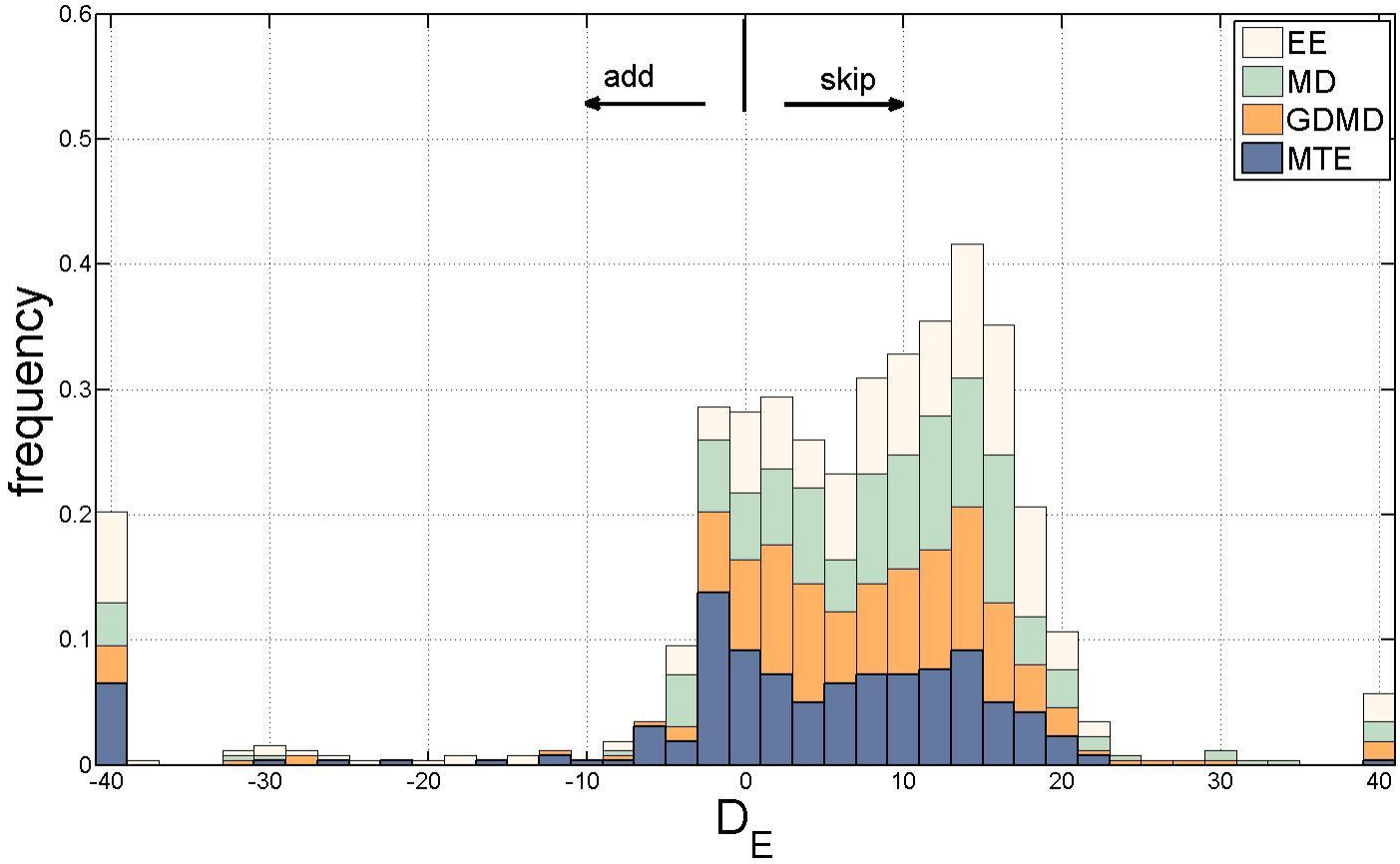

Fig.11. The histograms of the differences D_E for the ED features.

EXPERIMENT No.1 – ED accuracy evaluation

Each value in the Table 1 shows the rate of the distribution (in %) less than the 5-frames and the 10-frames differences, respectively [15].

Table 1. The rate of the distribution (in %)

№	Features	ABS(D_B)		ABS(D_E)	
		≤ 5	≤ 10	≤ 5	≤10
1	MTE	39.69	68.70	37.02	61.83
2	EE	41.22	67.55	20.99	44.27
3	MD	40.45	74.80	29.00	51.52
4	GDMD	49.23	83.20	34.73	56.87

EXPERIMENT No.2 – DTW speaker verification

For each ED feature a separate DTW speaker verification task was carried out, i.e. a single classifier was considered.

Experimental setup [13]:

- **Pre-processing step**
  - Frames - 30 ms, shift 10 ms, Hamming window;
  - Features – MFCC-14 coefficients, FFT size – 512;
- **DTW step**
  - DTW - normalize-wrap method with constrained endpoints conditions [10];
  - Local distance – root power sum cepstral distance [27];
  - Training data – 10 utterances per speaker randomly selected;
  - Speaker template - obtained by averaging (after dynamic time warping alignment) of his training utterances [21];
- **Verification scheme**
  - Cohort normalization [2];
  - To set the individual thresholds, the training set is used directly as a validation set;
  - 142 client accesses (CA) & 1562 impostor accesses (IA) per trial;
  - 5 trials are used – total 710 CA and 7810 IA;

EXPERIMENT No.2 – DTW speaker verification

Results [15]:

FRR - False Rejection Rate; FAR - False Acceptance Rate; HTER - Half Total Error Rate; CI – confidence interval for HTER. Confidence values δ and standard deviations σ obtained from the Z_{HTER}-test [1] are shown in Table 3. With [A, B] are noted the two endpoints detection features A and B being tested.

Table 2. Speaker verification results

№	Features	FRR[%]	FAR[%]	HTER[%]	95% CI
1	**Manual**	6.90	4.98	5.94	±0.0096
2	**MTE**	11.83	10.47	11.15	±0.0123
3	**EE**	14.08	12.48	13.28	±0.0133
4	**MD**	10.56	8.06	9.31	±0.0116
5	**GDMD**	8.30	7.31	7.80	±0.0105

Table 3. Confidence values

	[GDMD, MD]	[GDMD, MTE]
δ	93.88	99.99
σ	0.0080	0.0082

EXPERIMENT No.2 – DTW speaker verification

Speaker verification performance

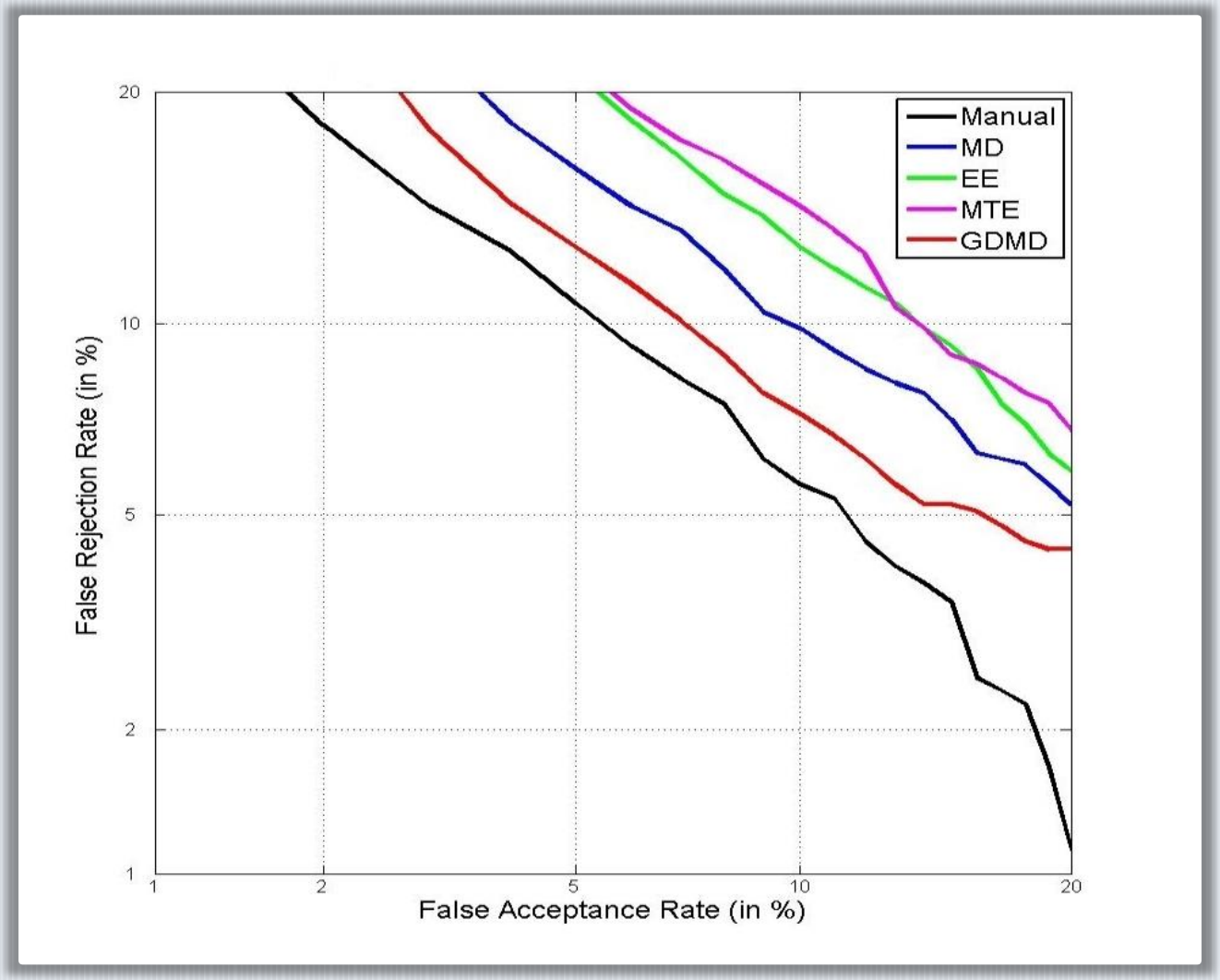

Fig.12. DET curves [8] for different ED features

CONCLUSIONS

Based on the experimental results the following conclusions are made:

- ***The GDMD feature demonstrates the best performance in the endpoint detection tests based on verification rate. This is due to the minimal number of the serious endpoint detection errors obtained for this feature.***

- ***Based on the Z_{HTER}-test the GDMD feature is statistically significantly different from the MD and MTE features;***

- ***The GDMD feature has two drawbacks: a need to adjust a few parameters in the MGDS estimation and increased number of computation (in comparison with the MD feature).***

REFERENCES

1. Bengio, S., Mariethoz, J.: A Statistical Significance Test for Person Authentication, ODYSSEY - The Speaker and Language Recognition Workshop, pp.237-244 (2004).
2. Burileanu, C., Moraru, D., Bojan, L., Puchiu, M., Stan, A.: On Performance Improvement of a Speaker Verification System Using Vector Quantization, Cohorts and Hybrid Cohort-World Models, International Journal of Speech Technology, No.5, pp.247–257 (2002).
3. Gerven, S., Xie, F.: A comparative study of speech detection methods, Eurospeech, pp. 1095-1098 (1997).
4. Gu, L., Zahorian, S.: A new robust algorithm for isolated word endpoint detection, IEEE ICASSP, vol.IV, pp.4161-4164 (2002).
5. Huang, L., Yang, C.: A Novel Approach to Robust Speech Endpoint Detection in Car Environment, IEEE ICASSP, pp.1751-1754 (2000).
6. Jia, C., Xu, B.: An Improved Entropy based Endpoint Detection Algorithm, ISCSLP, pp. 96-100 (2002).
7. Li, Q., Zheng, J., Tsai,A., Zhou, Q.: Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition, IEEE Transaction on SAP, vol.10, No.3, pp.146-157 (2002).
8. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET Curve in Assessment of Detection Task Performance, Eurospeech, pp.1895-1898 (1997).
9. Mesa-Navarro, J., Moreno-Bilbao, A., Lleida-Solano, E.: An Improved Speech Endpoint Detection System in Noisy Environments by Means of Third-Order Spectra, IEEE Signal Processing Letters, vol.6, No.9, pp.224-226 (1999).
10. Myers, C., Rabiner, L., Rosenberg, A.: Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition, IEEE Transactions on ASSP, vol.28, No.6, pp.623-635 (1980).
11. Ouzounov, A.: BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels, Cybernetics and Information Technologies, vol.3, No.2, pp.101-108 (2003).
12. Ouzounov, A.: A Robust Feature for Speech Detection, Cybernetics and Information Technologies, vol.4, No.2, pp.3–14 (2004).
13. Ouzounov, A.: Cepstral Features and Text-Dependent Speaker Identification - A Comparative Study, Cybernetics and Information Technologies, vol.10, No.1, pp.1-12 (2010).
14. Ouzounov, A.: Telephone Speech Endpoint Detection Using Mean-Delta Feature, Cybernetics and Information Technologies, vol.14, No.2, pp.127-139 (2014).
15. Ouzounov A.: Noisy Speech Endpoint Detection Using Robust Feature, Springer International Publishing Switzerland 2014, V. Cantoni et al. (Eds.): BIOMET 2014, LNCS 8897, pp. 105–117 (2014).
16. Ramirez, J., Segura, J., Benítez, C., De la Torre, A., Rubio, A.: Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information, Speech Communication, vol.42, No.3-4, pp.271-287 (2004).
17. Ramirez, J., Yelamos, P., Gorriz J., Seguraet, J.: SVM-based speech endpoint detection using contextual speech features, Electronics Letters, vol.42, No.7, pp.426-428 (2006).
18. Seok, J., Bae, K.: A Novel Endpoint Detection using Discrete Wavelet Transform, IEICE Transaction on Inf. & Syst., vol.E82-D, No.11, pp.1489-1491 (1999).
19. Shin, W., Lee, B., Lee, Y., Lee, J.: Speech/non-speech classification using multiple features for robust endpoint detection, IEEE ICASSP, pp.1399-1402 (2000).
20. Wu, B. F., Wang, K. C.: Robust Endpoint Detection Algorithm based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments, IEEE Transactions on SAP, vol.13, No.5, pp.762-775 (2005).
21. Yamamoto, K., Jabloun, F., Reinhard, K., Kawamura, A.: Robust Endpoint Detection for Speech Recognition Based on Discriminative Feature Extraction, IEEE ICASSP, vol.I, pp.805-808 (2006).

REFERENCES

21. Zelinski, R., Class, F.: A Learning Procedure for Speaker–Dependent Word Recognition System based on Sequential Processing of Input Tokens, IEEE ICASSP, pp.1053-1056 (1983).
22. Zhao H., Zhao, L., Zhao, K., Wang, G.: Voice Activity Detection based on Distance Entropy in Noisy Environment, Fifth International Joint Conference on INC, IMS and IDC, pp.1364-1367 (2009).
23. Zhang, Z., Furui, S.: Noisy Speech Recognition based on Robust End-point Detection and Model Adaptation, IEEE ICASSP, vol.1, pp.441-444 (2005).
24. Zhu, J. and Chen, F.: The Analysis and Application of a New Endpoint Detection Method based on Distance of Autocorrelated Similarity, Eurospeech, pp.105-108 (1999).
25. Hegde, R., Murthy, H., Gadde, V.: Significance of the Modified Group Delay Feature in Speech Recognition, IEEE Transactions on Audio, Speech and Language Processing, vol. 15, No.1, pp.190-202 (2007).
26. Murthy, H., Gadde, V.: The modified group delay function and its application to phoneme recognition, IEEE ICASSP, vol.1, pp.68-71 (2003).
27. Tohkura, Y.: A Weighted Cepstral Distance Measure for Speech Recognition, IEEE Transactions on ASSP, vol.35, No. 10, pp.1414-1422 (1987).
28. Hu, Y. and Loizou, P.: Subjective evaluation and comparison of speech enhancement algorithms, *Speech Communication*, 49, pp.588-601 (2007).
29. http://ecs.utdallas.edu/loizou/speech/noizeus/
30. Renevey, Ph. and Drygajlo, A.: Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech, pp.1883-1886 (2001).

Thanks for your time!

For any questions, please
send an email to atanas@iinf.bas.bg