B U L G A R I A N A C A D E M Y O F S C I E N C E S
INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Atanas Petrov Ouzounov

SPEECH DETECTION IN SPEAKER RECOGNITION SYSTEMS

ABSTRACT OF PhD THESIS

Consultant: Assoc. Prof. Georgi Gluhchev

Sofia
2020

Keywords: speaker recognition, speaker verification, speaker identification, voice activity detection, endpoint detection, group delay spectrum, spectral autocorrelation function, finite state machine, multilayer perceptron.

Introduction

Biometrics is the science of recognizing individuals through analysis by technical means of his physical or behavioral traits. It assumes that many of these traits (modalities) are strictly individual. The following physical traits are considered: voice, face, iris, fingerprints, palm veins, hand geometry, palm prints, ear shape, and respectively behavioral such as signature, handwriting style, keyboard dynamics, gait and more [Kisku et al., 2014].

In the last decade, biometric technologies became a rapidly developing area (in the US and China), and their deployment is in various fields – from grocery stores, airports to government institutions. The need for biometric solutions drives enormous investments in research leading to the development of new algorithms for features extraction and classification and design of advanced applications.

Voice is one of the primary modalities and the most accessible biometric trait, because of the widespread use in recent years of mobile phones and voice over Internet (VoIP) applications. This fact gives the voice a significant advantage over other biometric traits. It leads to the development of many more applications in the field of voice biometrics than in other modalities.

Currently the applications in voice biometrics can be divided into three main groups [Jain et al., 2008]:

- speaker detection (speaker spotting) - detecting a speaker through analysis of multiple calls (e.g. in call centers);
- speaker verification (voice authentication) - a typical application is remote access control by phone (e.g. bank transactions);
- forensic speaker recognition;

A trendy area is the mobile voice biometrics, i.e. the development of biometric applications for mobile devices (phones, tablets, etc.). The main problem in this area is the operation of biometric devices in dynamically changing environment.

In fact, the applications listed above are always based on a system (local or remote) for speaker recognition. No matter what is the task – text-dependent or text-independent, verification or identification, this system must include one a mandatory algorithm (module), namely a voice activity detector. It separates speech fragments in the received audio stream and sends the information about them for further processing in the system. Actually, its functioning is crucial for the whole system. This is because the speaker's voice model only uses the speech fragments and the separation accuracy has a significant impact on the final decision of the biometric system.

The rapidly development of biometric technologies (including voice biometrics) worldwide, determines the topic of thesis - voice activity detection in speaker recognition systems as extremely up-to-date research.

Purpose of the thesis

The development of robust features for voice activity detection algorithms intended for speaker recognition with telephone speech is the purpose of the thesis.

Tasks of the thesis

The following tasks have been formulated to achieve the goal of the dissertation:

1. To develop robust features for speech detection, which based on the properties of the spectral autocorrelation function and the group delay spectrum.

2. To develop an approach for short phrase endpoint detection that includes two algorithms - for adaptive thresholds settings and a finite state machine.
3. To develop endpoint detection algorithms that uses the proposed features and to study them experimentally in fixed-phrase speaker verification tasks.
4. To develop voice activity detection algorithms that uses the proposed features and to study them experimentally in text-independent speaker identification tasks.

Research methodology

The recommended methodology is based on methods and approaches from the following areas:

- linear algebra – linear transformations, etc.;
- digital signal processing - correlation analysis, spectral analysis and others;
- pattern recognition - neural networks, hidden Markov models and others.

Content of the dissertation

The dissertation consists of a glossary of terms, introduction, five chapters, contributions, and dissertation publications and citations and references. The first chapter is entitled *"Speech Detection: A Review"*, chapter two -"*Speech detection features based on the properties of SACF and GDS*, chapter three -*"Algorithms for endpoint detection in fixed-phrase speaker verification. The experimental study"*, fourth -*"VAD algorithms in text-independent speaker identification. The experimental study"* and fifth -*" BG-SRDat – Telephone speech corpus intended for speaker recognition"*. The main content is set out on 164 pages, 48 figures and 27 tables are included. The list of references includes 151 sources.

Chapter 1. Speech detection: A Review

1.1. Introduction

Speech detection is defined as the process of localization of speech among different types of non-speech events. Non-speech events are all audio events accompanying the realization of the speech message but not related to the information it carries. These non-speech events may be from the surrounding environment (street noise, background conversations, etc.), from the communication channel or sound artifacts generated by the speaker (sigh, cough, etc.).

Speech detection is referred to in various terms, the most common of these being Voice Activity Detection (VAD) [Tuononen, 2008]. As a speech detection sub-task and sometimes as a separate type of detection the Endpoint Detection (ED), i.e. defining the boundary points of the speech message is considered. It locates only the border points (start and end) of a message while pausing inside the word or phrase is not marked (if they are up to a certain length). In most cases, ED- algorithms are used in text-dependent speaker recognition task with words or short phrases.

The speech detector is a separate step in the biometric pre-processing system. The main goal in developing this kind of algorithm is to achieve robustness of their decision, i.e. the segmentation of the speech sequence does not change regardless of signal quality and environmental conditions variations [Nautsch et al., 2016].

1.2. VAD algorithms

VAD algorithms contain three main modules: feature extraction, classifier, and hangover scheme [Ramirez et al., 2007]. Frequently used features are based on - spectral divergence [Ramirez et al., 2004], group delay functions [Krishnan et al., 2006], autocorrelation functions

[Ghaemmaghami et al., 2010a], periodic and aperiodic components [Ishizuka et al., 2010], delta-phase spectrum [McCowan, 2012], formants [Yoo et al., 2015], polynomial regression of the Mel spectrum [Disken, 2017], i-vectors [Yamamoto et al., 2017].

The decision module (classifier) uses different approaches according to the task and the type of used speech data. For example, for text-dependent speaker recognition with corpus RSR2015 [Alam et al., 2014] the sequential Gaussian mixture model [Ying et al., 2011] have been used. In text-independent speaker verification in NIST 2008 SRE (Speaker Recognition Evaluation) a multilayer perceptron [Ganapathy et al., 2011] is used. With the same type of speaker verification and NIST 2016 SRE has used a deep learning neural network [Yamamoto et al., 2017].

1.2.1. Features used in VAD and ED algorithms

The text describes the features used in VAD and ED algorithms. The material in this section is mainly based on the review published in [Graf et al., 2015].

1.2.2. Classifiers used in VAD and ED algorithms

The main classifiers used in VAD algorithms are the Gaussian mixture model, support vectors machine, and method with i-vectors.

1.2.3. VAD algorithms in speaker recognition systems

1.2.3.1. Study of VAD algorithms in text-independent speaker verification system for NIST SRE

In work [Mak et al., 2014], the VAD algorithms have been specially adapted for NIST 2010 SRE. A feature of these speaker verification tests is the quality of the records. In a considerable part of them, the SNR is about 5 dB. Two systems are used for speaker verification. The former uses the GMM-SVM approach [Campbell et al., 2006a] and the latter is with the i-vectors method [Dehak et al., 2011]. The following VAD algorithms were tested in the paper. These are AE-VAD (used signal energy), ASR-VAD (segmented data obtained by speech recognition system and provided by NIST [NIST, online]), GMM-VAD (algorithm using model with Gaussian mixtures [Fukuda et al., 2010], SM-VAD (Sohn algorithm [Sohn et al., 1999]), SS + SM-VAD (SM-VAD using spectral subtraction), SS + AE-VAD (AE-VAD using spectral subtraction).

With the GMM-SVM system, the used SM-VAD detector performs better than GMM-VAD for NIST SRE interview data. The main reason is a large amount of pre-segmented speech needed for GMM training. The spectral subtraction dramatically improves the accuracy of the AE-VAD signal energy detector and has little effect on SM-VAD accuracy. In the statistical model, the background noise is taken into consideration in the calculation of the estimation function, and in this case, the spectral subtraction is not sufficient enough. Best results for both criteria – EER and minDCF - were obtained at SS + AE-VAD.

In the system that using i-vectors four versions of SM-VAD has been tested. It is assumed that the distribution of the Fourier coefficients can be respectively with Gaussian (basic algorithm), with Laplace and with Gamma distribution. The fourth test has a Gaussian distribution but spectral subtraction was used in the pre-processing step. Experiments show that SM-VAD with Gamma distribution demonstrates better results than the underlying algorithm at EER (Equal Error Rate) criterion.

1.2.3.3. VAD algorithms based on MLP

The proposed algorithm [Ganapathy et al., 2011] is based on the posterior probabilities of the phonemes in English obtained at the outputs of a multilayer perceptron (MLP). In MLP training are used features obtained by the frequency domain linear prediction method (FDLP) [Ganapathy et al., 2010]. Thus a 420-dimensional feature vector is obtained. MLP training has been implemented with CTS (conversational telephone speech) data [Hain et al., 2005] containing telephone calls lasting 180 hours. Speaker verification is based on the GMM-UBM system with i-vectors and GPLDA [Garcia-Romero et al., 2011]. To train, UBM uses data from NIST 2004 SRE, Switchboard II Phase III and NIST 2006 SRE. In training mode, the VAD

algorithm provided by NIST is used. In speaker verification tests the following VAD algorithms are implemented: with adaptive energy signal [Reynolds et al., 2005], with Mel-cepstrum, with time-frequency modulation [Mesgarani et al., 2006]. The MLP1- proposed algorithm uses as features are used Mel cepstrum with CMS, while the MLP2 uses features are obtained by FDLP. Verification accuracy is estimated by EER and detection accuracy by the total mean error of FAE and MDE calculated for all pronunciations. The experimental results show that the highest verification rate was achieved using the MLP2 VAD algorithm. It is interesting to note that in MLP2, a minimum EER is obtained even when training and testing are done with different languages.

1.3. Endpoint detection algorithms

1.3.1. Introduction

ED algorithms include two main steps – features extraction and decision. In the first stage, one or more speech features are calculated, for example - signal energy [Li et al., 2012], spectral entropy [Zhang et al., 2013], [Zhang et al., 2016], time-frequency parameters [Kyriakides et al., 2011], wavelets [Yali et al., 2014], Mel cepstrum [Cao et al., 2017] and others. On the second stage, the most commonly used are finite state machine [Chung et al., 2014] or classifiers - neural networks [Wu et al., 2012], Hidden Markov Models (HMM) [Zhang et al., 2005], Support Vector Machines (SVM) [Feng et al., 2016] and others.

1.3.2. Algorithms for endpoint detection

1.3.2. 4 Li algorithm using Teager energy

An ED-algorithm using the Teager Energy Operator (TEO) as a feature has been proposed [Li, 2012]. Unlike traditional energy, this type of energy contains information not only about the amplitude but also about the frequency characteristics of the signal. In order to determine the endpoints, threshold values and corresponding logical rules are introduced. Tests were made with speech with additive noises selected from NOISEX-92 [Noisex, online]. The main disadvantage of this ED–algorithm in noisy speech signals is the unsatisfactory detection of endpoints when there are fricative sounds. Notwithstanding this fact, compared to the traditional energy, the results obtained demonstrate the advantages of the TEO.

1.4. Conclusion

Based on the review, it can be concluded that the combination of sources providing various information is a successful strategy in developing algorithms for speech detection in a real-world environment. This involves a fusion of different representations of the speech signal, a fusion of multiple feature streams in one VAD algorithm, and a combination of different VAD algorithms. In turn, these VAD algorithms can be built with different classifiers, which give an opportunity for greater adaptability of the detection when changing environmental conditions.

Chapter 2. Speech detection features based on the properties of SACF and GDS

2.1. Speech detection features using a spectral autocorrelation function

2.1.1. Introduction

The work proposes to form speech detection features using the properties of the spectral autocorrelation function. The main idea is to achieve peak enhancement of the harmonic components in the spectral autocorrelation function using the approximation of its first derivative.

2.1.2. Spectral autocorrelation function. Properties.

Spectral AutoCorrelation Function (SACF) can be calculated in different ways according to the purposes for which it is used. It is accepted that the SACF is defined as discrete quantities of the magnitude spectrum (or power spectrum) with spectral resolution as in the Fourier transform used to obtain the spectrum. If $|X(k)|$ is the magnitude spectrum of the speech

obtained by FFT for the current segment, the biased estimate of the autocorrelation function $R_A(l)$ is defined as [Klapuri, 2000]

$$R_A(l) = \frac{2}{K} \sum_{k=0}^{K/2-l-1} |X(k)|.|X(k+l)|, \tag{2.1}$$

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

2.1.3. Delta spectral autocorrelation function

In this work, a parameter for speech detection based on the first-order derivative of the spectral autocorrelation function is proposed. Since this derivative has no analytical form, it can only be approximated by finite differences. However, applying the first-order finite difference to real signals leads to increased noise since these differences are, in fact, a high-pass filtration. To avoid this problem, an idea similar to this one described in [Rabiner et al., 2010] but implemented in different way is proposed. In [Rabiner et al., 2010] the first derivative in time of a cepstral contour is represented as an orthogonal polynomial approximation of the contour calculated within a specific time area. In this case, the first-order polynomial coefficient describes the slope (i.e., the first derivative in time) of the cepstral trajectory for a given time segment. These orthogonal first-order polynomial coefficients are known as delta cepstral coefficients or just a delta cepstrum.

Another interpretation of the delta cepstrum is proposed in [Fukuda et al., 2010], where it is considered as a sequence obtained at the output of a noncausal FIR filter. Its transfer function $H(z)$ is defined as

$$H(z) = \frac{\sum_{k=1}^{K} k.(z^k - z^{-k})}{2\sum_{k=1}^{K} k^2} \tag{2.5}$$

The derivative of the spectral autocorrelation considered as an output signal of the filter in (2.5) is proposed in the thesis. It is named Delta Spectral Autocorrelation Function (DSACF) and is calculated using SACF values within the current segment (intra-frame processing). In the text, SACF is referred to as $R(n,l)$ regardless of how it is calculated - by the amplitude or by the power spectrum. The spectrum type is specified further in the text.

DSACF $\Delta R(n,l)$ for the n^{th} segment is calculated by SACF $R(n,l)$ according to

$$\Delta R(n,l) = \frac{\sum_{q=1}^{Q} q.(R(n,l+q) - R(n,l-q))}{2\sum_{q=1}^{Q} q^2} \tag{2.6}$$

where $l = 0,...,L$ is the number of lags of the SACF; Q is typically between 2 and 5, i.e. filter length from 5 to 11 lags and $n = 0,...,N-1$, N is the number of segments. It is accepted $R(n,l) = 0$ for $l < 0$ and $l > L$, i.e. the first and last few values of $\Delta R(n,l)$, should not be subject to analysis as boundary conditions influence them. In Fig. 2.3 waveforms of three signals, their normalized SACFs and their corresponding DSACFs (Q= *3*) up to lag $L = 100$ are shown.

In the DSACF in Fig 2.3 (g) strong positive and negative peaks that are difficult to interpret are observed. To be overcome this it is proposed to use the idea of the second FFT described in [Wang et al., 2001], but applied not to the amplitude spectrum but to the spectral and delta spectral autocorrelation functions. In this way, it is possible to make direct comparison between the two spectra (2nd FFT spectrums) and determine the effect of the application of the delta filter (2.5) to the spectral autocorrelation function.

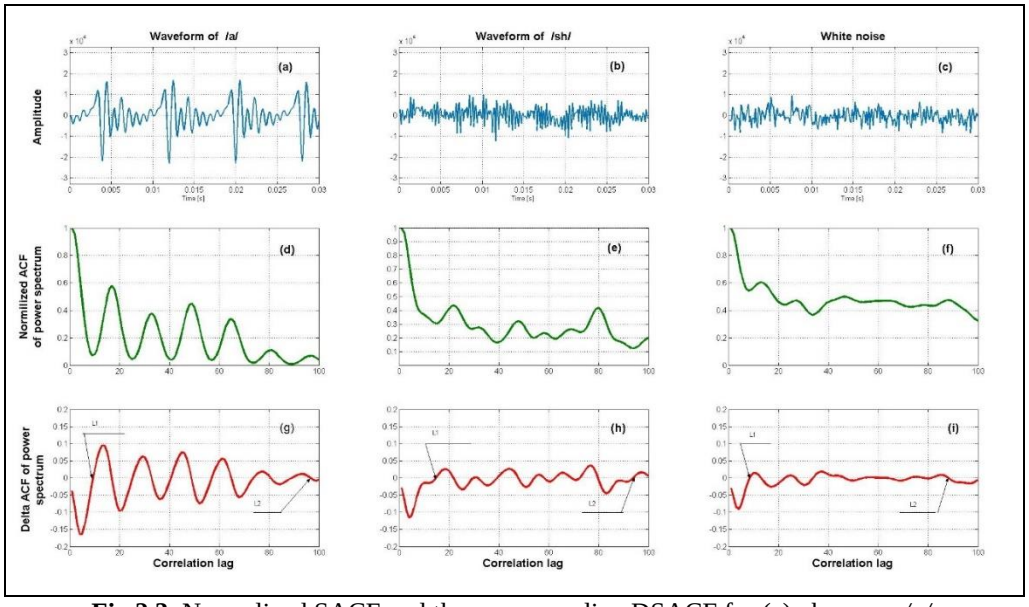

Fig.2.3. Normalized SACF and the corresponding DSACF for (a) phoneme */a/*, (b) fricative */sh/* and (c) white noise.

The Fig. 2.4 shows the amplitude spectrum of the part of the phoneme *'a'* (sampled frequency 8 kHz) and amplitude spectra of SACF, filter in (2.5) and DSACF - $|S_R(\Omega)|$, $|H(\Omega)|$ and $|S_{\Delta R}(\Omega)|$, respectively. The graphics are obtained with the following parameters - $K = 3$ (formula (2.6) - filter length 7 lags), FFT - 512 points, and the unbiased spectral autocorrelation function obtained by

$$R(l) = \begin{cases} \dfrac{1}{K/2 - |l|} \displaystyle\sum_{k=0}^{K/2-1-l} |X(k)|.|X(k+l)|; \; l \geq 0; \; l \leq L; \\ R(-l) \qquad\qquad\qquad\qquad ; l < 0 \end{cases} \tag{2.7}$$

where K is the number of FFT points, $L = K/4$ and $|X(.)|$ is the amplitude spectrum.

In Figs. 2.5 and 2.6 the block diagrams of the algorithms for calculating of the $|S_R(\Omega)|$ and $|S_{\Delta R}(\Omega)|$ are shown.

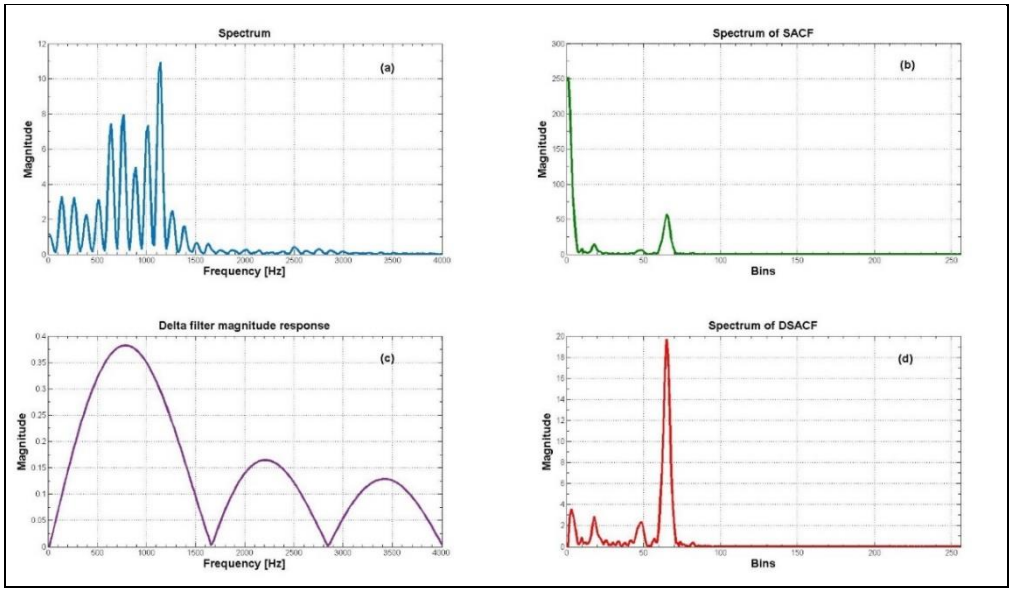

Fig. 2.4. Magnitude spectra of: (a) phoneme /a/, (b) SACF, (c) delta filter and (d) DSACF.

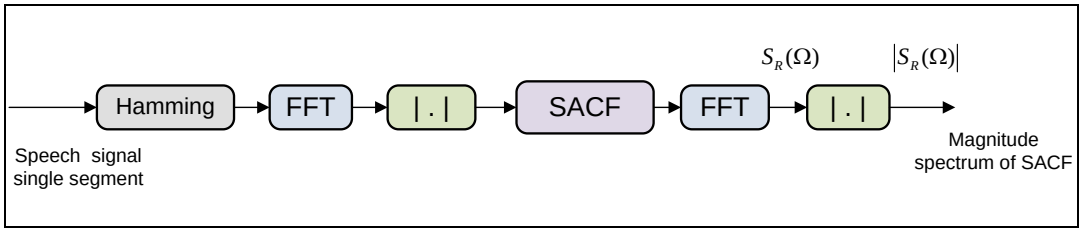

Fig.2.5. Block diagram of the algorithm for calculating of $|S_R(\Omega)|$.

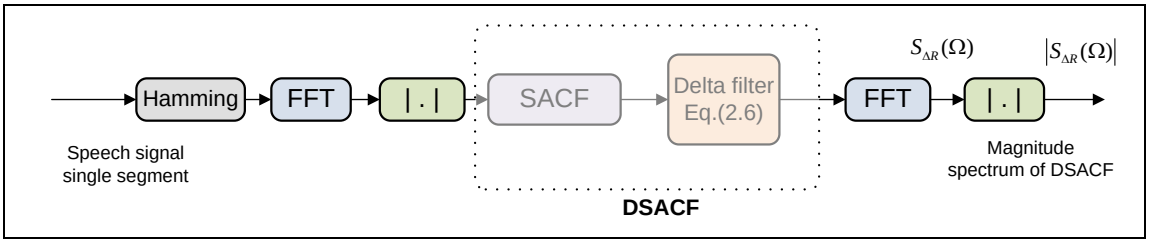

Fig.2.6. Block diagram of the algorithm for calculating of $|S_{\Delta R}(\Omega)|$.

The fundamental frequency (pitch) in the phoneme segment shown in Fig. 2.4 (a) is about 125 Hz. At a sampling rate of 8000 Hz and FFT with 512 points, the difference between the peaks of the pitch in Fig. 2.4 (a) is 8 spectrum bins. By applying 2nd FFT to SACF and DSACF and according to the calculations in [Akant et al., 2010] the peak corresponding to the fundamental frequency in the spectrum is at 64 bin, as seen in Fig. 2.4 (b) and (d).

The delta filter with magnitude response shown in Fig.2.4 (c) can be regarded as a set of three band filters. In the figure, the amplitude is linear in order to be suitable for comparison with the other two amplitude spectra - the SACF and the DSACF. If the amplitude is logarithmic and the points are determined at -3 dB relative to the maximum (0 dB) the values of the frequencies are shown in Table 2.1. With f_L and f_H are noted the cutoff frequencies, f_0 is the central frequency and *B* are the bandwidth of the bandpass filters.

Table 2.1. Frequencies of the delta filter

BPF	f_L [Hz]	f_0 [Hz]	f_H [Hz]	B[Hz]
1	383	772	1179	796
2	1904	2193	2501	597
3	3111	3405	3699	588

The first filter is essential in the filtering of the SACF. The level at its center frequency f_0 is higher than that of the first and second filters, respectively, by about 7.3 dB and 9.5 dB. This filter reduces the components in the spectra of SACF close to the DC term and corresponding to the envelope energy of the spectral autocorrelation function. Furthermore, there is a sharp peak in the DSACF spectrum that corresponds to the energy of fundamental frequency harmonics in the spectral autocorrelation function – as seen in Fig.2.4 (d).

In Fig. 2.9 are shown the described above magnitude spectrums but calculated for speech with additive white noise at SNR=5 dB.

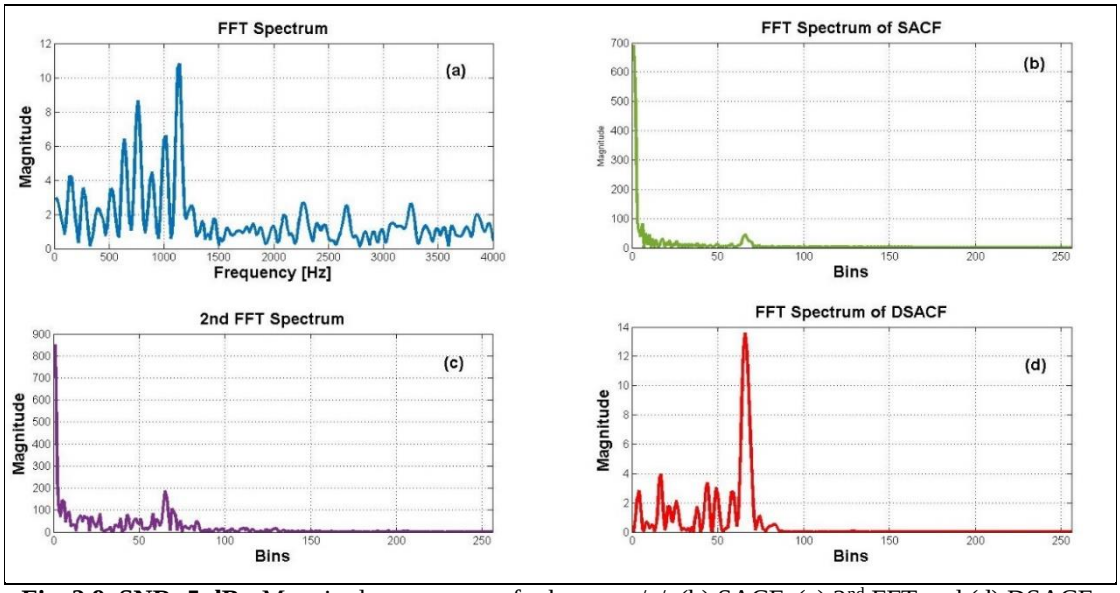

Fig. 2.9. SNR=5 dB - Magnitude spectrums of: phoneme /a/, (b) SACF, (c) 2nd FFT and (d) DSACF.

Comparing the spectra with the clear and noisy signals shown in Figs. 2.4 and 2.9 the following will be established. First, the idea of [Wang et al., 2001] to emphasize the pitch peak for the noisy signals is confirmed. In Fig. 2.9 (c) the peak in the 2nd FFT spectrum located at 64 bin corresponds to the fundamental frequency of 125 Hz. Second, when comparing the spectra of SACF, respectively - Figs. 2.4 (b) and 2.9 (b) and of the DSACF - Figs. 2.4 (d) and 2.9 (d), it is found that the peak in the SACF spectrum is reduced to a much greater extent than the corresponding peak in the delta spectral autocorrelation function. Moreover, the peak in the DSACF spectrum for the noisy signal is more pronounced even than in the 2nd FFT spectrum. These facts are arguments for using the DSACF as the basis for the formation of robust speech detection features.

2.1.4. Mean-Delta (MD) features

2.1.4-A. Motivation

The characteristics of the DSACF described in §2.1.3 underlie the features suggested in the dissertation. As can be seen in Fig. 2.3 (g) (h) (e) DSACF has significant positive and negative peaks even for the fricative consonant *"sh"*. This property, on the one hand, and on the other hand, the shape of the spectrum of the DSACF for noisy signals shown in Fig. 2.9 (d), are the starting points for the formation of the features suggested in the dissertation. The author assumes that if a parameter is formed that for the current segment is a summary estimate of the number and magnitude of the peaks in the DSACF, then this parameter can be successfully used as a speech detection feature, especially for noisy speech signals. In the dissertation, this assumption was confirmed experimentally for two versions of the DASCF calculated respectively by the Fourier amplitude spectrum and by the modified group delay spectrum.

The speech detection features suggested in Chapter 2 are formed by the DSACF and not by its spectrum. The direct use of the DSACF spectrum (i.e., the application of a second FFT) to develop speech detection features and the evaluation of their effectiveness in speaker recognition systems is a subject of future research.

2.1.4.1. Mean Delta feature

The first proposed feature is called the Mean-Delta (MD) feature and is intended for use in time contour analysis. For n^{th} segment the MD feature $m_d(n)$ is defined as

$$m_d(n) = F\left(\sum_{l=0}^{L}\left|\Delta R(n,l)\right|\right) \tag{2.13}$$

where $\Delta R(n,l)$ is the causal part of DSACF. In the formula (2.13) with *F (.)* is denotes an additional transformation which is defined according to the features of the speech detection algorithm. The so-called *basic* algorithm for calculation of the MD feature will be presented here. For n^{th} segment, it has the form:

- compute the magnitude spectrum $|X(k)|$ of the Hamming-windowed speech signal via the Fast Fourier Transform (FFT) of size *K*;
- apply mean normalization (the mean vector of the amplitude spectrum is calculated over all segments in the file)

$$|\hat{X}(n,k)| = \frac{|X(n,k)|}{\frac{1}{N}\sum_{n=1}^{N}|X(n,k)|} \tag{2.14}$$

  where *N* is the number of segments in the utterance (file);
- compute the unbiased spectral autocorrelation function with lags *L=K/4* using the normalized amplitude spectrum

$$R(n,l) = \frac{1}{K/2-|l|}\sum_{k=0}^{K/2-1-l}|\hat{X}(n,k)|.|\hat{X}(n,k+l)|; \quad l \geq 0;\ l \leq L; \tag{2.15}$$

- compute delta spectral autocorrelation function $\Delta R(n,l)$ according to (2.6) with *Q = 3*;
- smooth the time contour of the delta spectral autocorrelation function (for each lag) using the Long-Term Spectral Envelope (LTSE) algorithm with parameter *J = 3* [Ramirez et al., 2004]. The smoothed version of $\Delta R^s(n,l)$ is noted as

$$\Delta R^s(n,l) = \max\{\Delta R(n+j,l)\}_{j=-J}^{j=+J}. \tag{2.16}$$

- compute the MD parameter $m_d(n)$ as

$$m_d(n) = \left[\sum_{l=0}^{L}|\Delta R^s(n,l)|\right]^{0.5} \tag{2.17}$$

- smoothing of $m_d(n)$ contour by a moving average filter;

In Fig. 2.10 the block diagram of the above algorithm for calculating of the MD feature is shown.

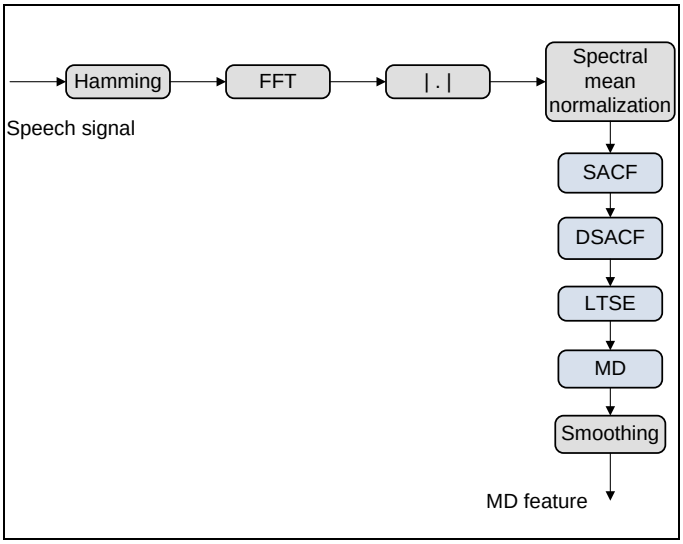

Fig.2.10. Block diagram of the algorithm for calculating of the MD feature.

2.1.4.2. Basic mean delta feature

The second feature is named as Basic Mean-Delta (BMD) and is intended for speech detection in recognition algorithms, i.e. the parameter is defined in vector form. For n^{th} segment BMD feature $m_{BMD}(n)$ is defined as follows:

- compute the magnitude spectrum $|X(k)|$ of the Hamming-windowed speech signal via the FFT with size *K*;
- apply mean normalization (the mean vector of the amplitude spectrum is calculated over all segments in the file) according (2.14);
- compute the unbiased spectral autocorrelation function with lags *L=K/4* using the normalized amplitude spectrum according (2.15)
- compute delta spectral autocorrelation function $\Delta R(n,l)$ according to (2.6) with *Q=3*;
- smoothing of the time contour of the delta spectral autocorrelation function (for each lag) using the Long-Term Spectral Envelope (LTSE) algorithm with parameter *J = 3* [Ramirez et al., 2004]. The smoothed version of $\Delta R^s(n,l)$ is noted as

$$\Delta R^s(n,l) = \max\left\{\Delta R(n+j,l)\right\}_{j=-J}^{j=+J}. \quad (2.18)$$

- divide the total number of lags *L* in DSACF by *V* equal in length and non-overlap ranges as follows

$$\{L_1, L_2\}...\{L_v, L_{v+1}\}...\{L_{2V-1}, L_{2V}\} \quad (2.19)$$

- determine the size of the BMD vector $m_{BMD}(n)$ by the number of *V* ranges in the form

$$m_{BMD}(n) = \{m_{BMD}(n,1),...,m_{BMD}(n,v),...,m_{BMD}(n,V)\} \quad (2.20)$$

- v^{th} component in $m_{BMD}(n,v)$ is defined as

$$m_{BMD}(n,v) = \log\left[\max\left\{\left|\Delta R^s(n,m)\right|\right\}_{m=L_v}^{m=L_{v+1}}\right] \quad (2.21)$$

2.1.4.3. Modified mean delta feature

The third feature is called Modified Mean-Delta (MMD) feature and is intended for speech detection by recognition algorithms. It is defined in a manner similar to the basic MD feature in § 2.1.4.2. The difference is that a rectangular window is applied on the lags sequence. This window length is *Y* lags and is shifted by step of *U* lags so the number of steps is *V* and it determines the size of the MMD vector. For n^{th} segment MMD parameter $m_{MMD}(n)$ is

$$m_{MMD}(n) = \{m_{MMD}(n,1),...,m_{MMD}(n,v),...,m_{MMD}(n,V)\} \quad (2.22)$$

where $m_{MMD}(n,v)$ has the form

$$m_{MMD}(n,v) = \log\left[\max\left\{\left|\Delta R^s(n,m)\right|\right\}_{m=(v-1)*U}^{m=(v-1)*U+Y}\right] \quad (2.23)$$

2.2. Speech detection features based on the group delay spectrum

2.2.1. Introduction

This item includes the description of the Group Delay Spectrum (GDS) [Murthy et al., 2011] and an analysis of its variations for speech signals with additive noise. This analysis is indirectly done by using of the Projection Distortion Measure (PDM) based on the additive spectral model [Mansour et al., 1989]. Here a feature called the Group Delay Mean Delta (GDMD) that combines Modified GDS (MGDS) [Hegde et al., 2007] and the MD feature discussed in § 2.1.1.4 is proposed.

2.2.2. Group Delay Spectrum

2.2.3. Study of the GDS for speech with additive noise

2.2.4. Group Delay Mean Delta feature

MGDS $\tau_m(\omega)$ is defined as

$$\tau_m(\omega) = \left(\frac{\tau(\omega)}{|\tau(\omega)|} \right) (|\tau(\omega)|)^{\alpha} \tag{2.44}$$

where

$$\tau(\omega) = \left(\frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}} \right) \tag{2.45}$$

and $S(\omega)$ is the cepstral-smoothed version of the FFT spectrum $|X(\omega)|$. The parameters α and γ vary from 0 to 1 ($0 < \alpha \leq 1$) and ($0 < \gamma \leq 1$). These two parameters and cepstral-smoothed spectrum in denominator were introduced to reduce the amplitude peaks and to limit the dynamic range in the MGDS. To control the degree of cepstral smoothing in $S(\omega)$ a cepstral lifter with a length l_w is used.

A new feature called Group Delay Mean Delta (GDMD) - a feature that is intended for speech detection by contour analysis is proposed. It uses the Mean-Delta approach proposed in §2.1.4, but in this case, the spectral autocorrelation function is defined not with the FFT spectrum but with the modified GDS defined in (2.44). The main purpose of this combination is to use the properties of the GDS and to achieve enhancement of the peaks in the delta spectral autocorrelation function. Two modifications of the GDMD feature are proposed. For n^{th} segment, the proposed GDMD features are calculated in three steps (for the sake of the clarity the index *n* is omitted in some formulas):

A. Step 1. Calculation of MSGS according to [Hegde et al., 2007] as follows:

- let $x(n)$ is the speech signal in the current segment, $n = 1, \ldots, N$ is the number of samples in the segment;
- apply FFT to the sequences *x(n)* and *nx(n)* and obtain the corresponding spectra *X(k)* and *Y(k)*;
- compute $|S(k)|$ - cepstrallly smoothed spectrum of $|X(k)|$ using low-order cepstral lifter l_w;
- compute the MGDS $\tau_m(k)$ as

$$\tau_m(k) = [\text{sign}].\left| \frac{X_R(k)Y_R(k) + Y_{\text{Im}}(k)X_{\text{Im}}(k)}{S(k)^{2\gamma}} \right|^{\alpha}, \tag{2.46}$$

where *[sign]* is the sign of the term

$$\frac{X_R(k)Y_R(k) + Y_I(k)X_I(k)}{S(k)^{2\gamma}} \tag{2.47}$$

Parameters α, γ and l_w are adjusted according to the particular requirements.

B. Step 2. Calculation of MD feature, using MGDS $\tau_m(k)$ (2.46) as follows:

- compute the average MGDS – averaged over all frames in the utterance;
- obtain the mean normalized MGDS $\tau_m^n(k)$ by dividing the frame MGDS by the average MGDS;
- compute the non-normalized unbiased spectral autocorrelation function $R_m(l)$ using the mean normalized MGDS $\tau_m^n(k)$

$$R_m(l) = \frac{1}{K/2 - l} \sum_{k=0}^{K/2-l} \tau_m^n(k)\tau_m^n(k+l); \;\; l \geq 0; l \leq L; \tag{2.48}$$

where *K* is the size of FFT, $l = 0, \ldots, L$, L is the number of correlation lags, and $L = K/4$.

- compute the delta spectral autocorrelation function $\Delta R_m(n,l)$ according to (2.7) using $R_m(n,l)$ with delta window Q as (here index n is included)

$$\Delta R_m(n,l) = \frac{\sum_{q=1}^{Q} q.(R_m(n,l+q) - R_m(n,l-q))}{2\sum_{q=1}^{Q} q^2} \tag{2.49}$$

- perform a contour smoothing for delta spectral autocorrelation function $\Delta R_m(n,l)$ by using J-order long-term spectral envelope algorithm [Ramirez et al., 2004]. The obtained smoothed version of $\Delta R_m(n,l)$ is denoted as $\Delta R_m^s(n,l)$

$$\Delta R_m^s(n,l) = \max\left\{\Delta R_m(n+j,l)\right\}_{j=-J}^{j=+J}. \tag{2.50}$$

- compute the GDMD parameter $m_{gd}(n)$ using $\Delta R_m^s(n,l)$ as

$$m_{gd}(n) = \left[\sum_{l=0}^{L}\left|\Delta R_m^s(n,l)\right|\right] \tag{2.51}$$

C. Step 3-1. Compute and smooth the lin-GDMD contour:

- compute the lin-GDMD parameter $m_{gd-lin}(n)$ using $m_{gd}(n)$ according to

$$m_{gd-lin}(n) = \left[m_{gd}(n)\right]^{0.5} \tag{2.52}$$

- smooth the m_{gd-lin} contour by a moving average filter;

C. Step 3-2. Compute and smooth the log-GDMD contour:

- normalize the $m_{gd}(n)$ contour in (2.51) and obtain the final contour $m_{gd}^*(n)$ as

$$m_{gd}^*(n) = \left|m_{gd}(n) - m_{gd}^{\min}\right|, \tag{2.53}$$

where $m_{gd}^{\min} = \min_n\{m_{gd}(n)\}$.

- compute log-GDMD according to

$$m_{gd-log}(n) = log(1 + m_{gd}^*(n)) \tag{2.54}$$

- smooth the $m_{gd-\log}$ contour by moving average filter;

Fig. 2.12 shows the block diagram of the above algorithm for computing the GDMD feature. The normalization done in (2.53) and (2.54) is proposed because the minimum values obtained in the GDMD contour are always less than 1, i.e., direct use of a log function is not appropriate.

2.3. Conclusions

The first part of Chapter 2 discusses some of the characteristics of the spectral autocorrelation function obtained by the FFT spectrum. A method is proposed in which, by applying a delta filter to the spectral autocorrelation function, the so-called delta spectral autocorrelation function is obtained. In the second part of the chapter has made a qualitative study of the effect of the additive noise on the GDS. On the one hand, based on the delta spectral autocorrelation function properties alone, and, on the other, by combining it with the modified group delay spectrum, a total of five speech detection features have been proposed. These are the features - MD, log -GDMD, lin-GDMD, BMD, and MMD. The first three are for detection by contours analysis and the last two for detection by recognition algorithms.

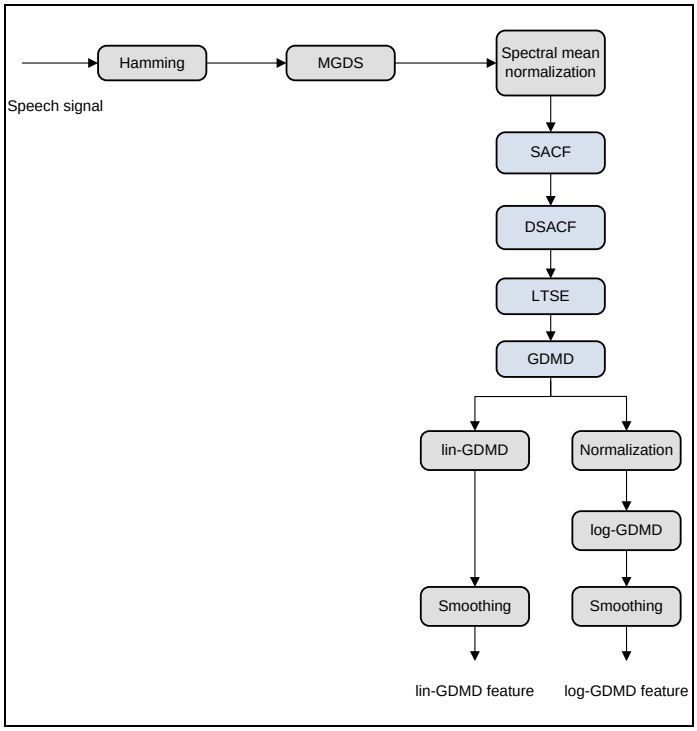

Fig. 2.12. Block diagram of the algorithm for the GDMD features computing.

Chapter 3. Algorithms for endpoint detection in fixed-phrase speaker verification. The experimental study

3.1. Introduction

In this chapter, a comparative experimental analysis is conducted using the proposed in the previous chapter features designed for contour-based speech detection. The following features are selected as references: Energy-Entropy (EE) feature [Huang et al., 2000]; Spectral Entropy with Normalized frame Spectrum (SENS) [Renevey et al., 2001]; Modified Teager's Energy (MTE) [Gu et al., 2002] and Long-Term Spectral Divergence (LTSD) [Luengo et al., 2010]. In the experiments endpoint detector including thresholds setting algorithm and finite state machine is used. Various versions of this endpoint detector are developed according to the contour features.

It should be noted that a *detector*, *endpoint detector*, *and an algorithm for endpoint detection* are used synonymously in Chapter 3. This is done to make the text clearer.

In order to estimate the performance of the endpoint detection algorithms, three experiments are carried out. In the first one, the Euclidean distances between two Z-normalized feature contours – for clean and noisy versions of the testing phrase [Chen et al., 2005] are calculated. The goal is to estimate the difference between contours caused by the noise. The speech samples from SpEAR corpus are used [SpEAR, online].

In the second one, the endpoint accuracy was evaluated in terms of frames differences between hand-labeled and detected endpoints. The speech samples from two corpora are used – in Bulgarian from BG-SRDat [Ouzounov, 2003] and in English from TIDIGITS [Dan Ellis, online].

In the third experiment, the performance of the endpoints detection algorithms in terms of the recognition rate is estimated via two fixed-text speaker verification applications. The first application is based on the Dynamic Time Warping (DTW) algorithm [Theodoridis et al., 2010] while the second one uses the left-to-right HMM [Gales et al., 2008]. The verification results are compared to those obtained by manual endpoint detection. Here the speech examples are selected only from the Bulgarian corpus BG-SRDat.

The Z_{HTER} – test method proposed in [Bengio et al., 2004] is used to assess the difference (in the statistical sense) between the endpoint detectors by using the obtained verification rate.

3.2. Reference features

The reference features listed above are described.

3.3. Analysis of Z-normalized contours

3.4. Endpoint detection algorithms

Most often, when developing endpoint detectors for short phrases, two algorithms work together - the first one for thresholds setting (fixed or adaptive) and the second - a finite state machine [Li et al ., 2002], [Abdulla et al ., 2009], [Chung et al ., 2014]. The paper proposes an approach for developing such a detector, including an algorithm for calculating adaptive thresholds and a deterministic finite state machine. In most cases, such ED-detectors are *ad hoc* solutions. In the course of the research, it was found to be extremely difficult to reproduce decisions based on heuristic rules accurately. Therefore, in the thesis, the efficiency of the proposed finite state machine is compared with the hangover algorithm, which is well described in the standard [ETSI, 2007]. Based on the proposed approach (and depending on the characteristics of the features contours), three algorithms for endpoint detection have been developed.

3.4.2. Adaptive thresholds settings algorithm

To reduce the detection errors due to the use of fixed thresholds scheme an adaptive algorithm that uses two pairs of thresholds is proposed. The first pair is intended for detection of the starting point, while the second one – for the ending point. In other words, two thresholds – low and high – are set using the contour characteristics in the beginning part of the speech record, and the state automaton uses these thresholds for starting point detection only. In a similar way, the two other thresholds – low and high – are set using the contour specifics in the ending part, and they are used only for detection of the ending point. The critical issue in this algorithm is how to define the beginning and ending parts in speech record based only on the contour features. In order to do this, it is proposed to use the contour peaks analysis.

Let $P=\{p_i\}, i=1,\ldots,G,$ is the set of peaks, where G is the total number of peaks in analyzed contour. Each peak is defined as $p_i=(v_i,l_i)$ where v_i is the peak value and l_i is the location of the peak, i.e., the number of contour frame where the peak is placed. Let define new set $Q_M=\underset{v}{\text{sort}}\{P\}$ obtained after sorting the peaks over the peak values v_i in descending order and select the first M of them and $M<<G$. Let define $l_{\min}$ and $l_{\max}$ where $l_{\min}=\min_l\{Q_M\}$ and $l_{\max}=\max_l\{Q_M\}$. The position of the splitting point l_{spl}, i.e., the point that divides the contour into two parts – beginning and ending – is defined as $l_{\text{spl}}=l_{\min}+\kappa(l_{\max}-l_{\min})$.

In the proposed algorithm, a single initial threshold is computed for each part of the contour. By using this threshold, two additional averages m_{down} and m_{up} are estimated. The first average is calculated from the contour values smaller than the threshold, while the second one – from the values equal to or higher than it. In such way, the pairs of thresholds $T_{\text{beg}}^{\text{low}}, T_{\text{beg}}^{\text{high}}$ for beginning part of the contour and $T_{\text{end}}^{\text{low}}, T_{\text{end}}^{\text{high}}$ for the ending one are defined. The thresholding algorithm is as follows.

Step 1. Compute the contour values $C(n)\geq 0; n=1,\ldots N$, N is number of frames.

Step 2. Find contour peaks $P=\{p_i\}$; $p_i=(v_i,l_i)$; v_i is the peak value, l_i is the location of the peak and $i=1,\ldots,G$, G is the number of peaks.

Step 3. Find $Q_M=\underset{v}{\text{sort}}\{P\}$ in descending order and select the first M peaks; $M<<G$.

Step 4. Compute $l_{\min} = \min_{l}\{Q_M\}$ and $l_{\max} = \max_{l}\{Q_M\}$.

Step 5. Compute splitting point

$$l_{\text{spl}} = l_{\min} + \kappa(l_{\max} - l_{\min}). \tag{3.22}$$

Step 6. Compute initial thresholds for the beginning and ending part of the contour:

$$\begin{aligned} T_{\text{beg}}^{\text{init}} &= \text{E}\{C(n)\};\ \ n = 1,\ldots,l_{\text{spl}}, \\ T_{\text{end}}^{\text{init}} &= \text{E}\{\text{C}(n)\};\ \ n = l_{\text{spl}}+1,\ldots,N. \end{aligned} \tag{3.23}$$

Step 7. Compute additional average values for the beginning part:

$$m_{\text{beg}}^{\text{down}} = \frac{\sum_{n=1}^{l_{\text{spl}}} C(n)w(n)}{\sum_{n=1}^{l_{\text{spl}}} w(n)},\ \ w(n) = \begin{cases} 1 & \text{if } C(n) < T_{\text{beg}}^{\text{init}}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.24}$$

$$m_{\text{beg}}^{\text{up}} = \frac{\sum_{n=1}^{l_{\text{spl}}} C(n)v(n)}{\sum_{n=1}^{l_{\text{spl}}} v(n)},\ \ v(n) = \begin{cases} 1 & \text{if } C(n) \geq T_{\text{beg}}^{\text{init}}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.25}$$

Step 8. Compute additional average values for the ending part:

$$m_{\text{end}}^{\text{down}} = \frac{\sum_{n=l_{\text{spl}}+1}^{N} C(n)w(n)}{\sum_{n=l_{\text{spl}}+1}^{N} w(n)},\ \ w(n) = \begin{cases} 1 & \text{if } C(n) < T_{\text{end}}^{\text{init}}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.26}$$

$$m_{\text{end}}^{\text{up}} = \frac{\sum_{n=l_{\text{spl}}+1}^{N} C(n)v(n)}{\sum_{n=l_{\text{spl}}+1}^{N} v(n)},\ \ v(n) = \begin{cases} 1 & \text{if } C(n) \geq T_{\text{end}}^{\text{init}}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.27}$$

Step 9. Compute pair of thresholds for the beginning part:

$$\begin{aligned} T_{\text{beg}}^{\text{low}} &= m_{\text{beg}}^{\text{down}} + \alpha_1(m_{\text{beg}}^{\text{up}} - m_{\text{beg}}^{\text{down}}), \\ T_{\text{beg}}^{\text{high}} &= \max(T_{\text{beg}}^{\text{init}}, \beta_1 T_{\text{beg}}^{\text{low}}). \end{aligned} \tag{3.28}$$

Step 10. Compute pair of thresholds for the ending part:

$$\begin{aligned} T_{\text{end}}^{\text{low}} &= m_{\text{end}}^{\text{down}} + \alpha_2(m_{\text{end}}^{\text{up}} - m_{\text{end}}^{\text{down}}), \\ T_{\text{end}}^{\text{high}} &= \max(T_{\text{end}}^{\text{init}}, \beta_2 T_{\text{end}}^{\text{low}}). \end{aligned} \tag{3.29}$$

The parameters $\alpha_1, \beta_1, \alpha_2, \beta_2, \kappa$ and M are adjusted according to the particular requirements.

3.4.3. Finite-state automation

In the book about Bulgarian phonetics [Tilkov et al., 1977] it is claimed that no word begins with more than four consonants, and no word ends with more than three consonants. The preliminary experiments with a limited set of Bulgarian words (selected from [Tilkov et al., 1977]) have shown that the voiced fragments can be preceded (in the beginning) and followed (in the end) by unvoiced ones with a length of about 200-400 and 400-600 ms, respectively. It is worth to point out that for the English language, it is claimed that no word begins with more than three consonants, and no current word ends with more than four consonants [Roach, 2009]. Besides, it is claimed that the mentioned above time fragments for the English language are about 300 and about 500 ms, respectively [Ghaemmaghami et al., 2010b]. The comprehensive analysis of this issue, however, is clearly outside the scope of this study.

These two time constants are applied in the developed state automaton for defining of the pre-voiced and post-voiced fragments where the beginning and ending points will be searched.

The proposed ED algorithm is based on eight-state automaton with states: INIT, SCAN_DATA, SCAN_START, MAYBE_IN, SCAN_END, MAYBE_OUT, END_FOUND and END. A specific feature of the proposed state automaton is that in some circumstances, an error may occur. If this is happened the ED algorithm stops, and the particular file is ignored in the further processing steps. The errors occur in four cases:

- when the utterance ends outside the audio file – error ERR_TOOLONG;
- when the SNR is very low – error ERR_LOWSPEECH;
- when the current thresholds do not allow the starting or ending points to be found – errors ERR_BAD_BEG_THRS, ERR_BAD_END_THRS;
- when the estimated length of the utterance is less than MinLengthTime – error ERR_TOOSHORT.

This error mechanism is designed to prevent cases when inappropriate speech data have been entered in the recognition system. Protection from so-called inappropriate pronunciation or sound artifacts is an essential step in the real-time voice verification systems over telephone lines.

The finite state machine based decision logic applied to the ED is shown in Fig. 3.12. The parameters $T_{\text{SCAN\_START}}$, $T_{\text{MAYBE\_IN}}$, $T1_{\text{SCAN\_END}}$, $T2_{\text{SCAN\_END}}$ and $T_{\text{MAYBE\_OUT}}$ are state timers. Each one of the time constants *MaxQuietTime, UpTime1, UpTime2, MiddleTime, MinLengthTime, EndTime, BegTime, MaxStateTime* determines the length of the interval after which the state transition is performed. In this algorithm two types of so-called Endpoint Candidates (EC) are proposed. These EC are segment numbers that are likely candidates for the ending point, which is selected from them by logical rules. The results from the proposed ED algorithm (with adaptive thresholds) applied on the log-GDMD feature contour for a noisy speech example selected from the "Lombard Speech" section in the SpEAR database are illustrated in Fig. 3.13. The state transition-timing diagram is shown in Fig. 3.13 (c). Along the contour in Fig. 3.13 (d) are marked important details in the temporal execution of the proposed algorithm as: hand-labelled and estimated endpoints, splitting point, endpoint candidate type-1, pairs of thresholds $T_{\text{beg}}^{\text{low}}$, $T_{\text{beg}}^{\text{high}}$ for beginning part and $T_{\text{end}}^{\text{low}}$, $T_{\text{end}}^{\text{high}}$ for ending part one.

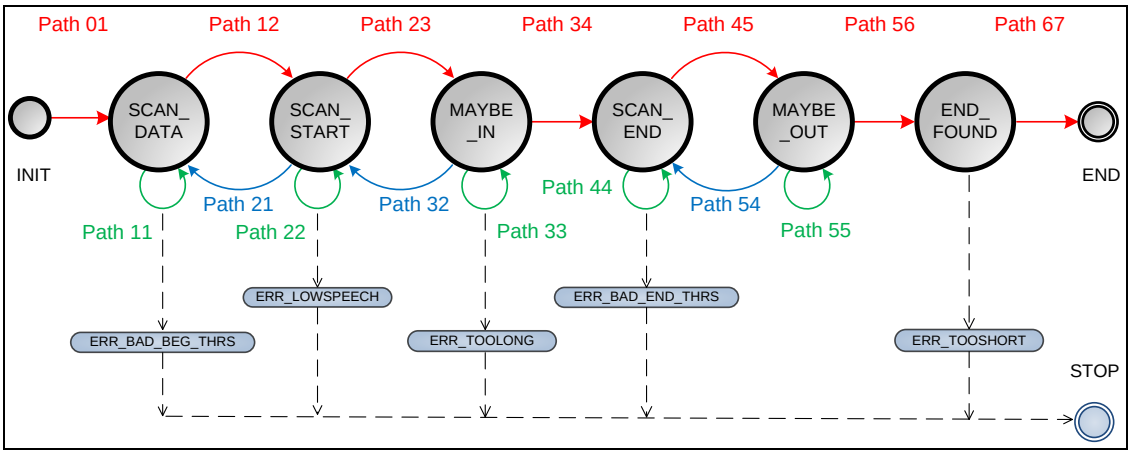

Fig.3.12. Finite state machine based decision logic diagram.

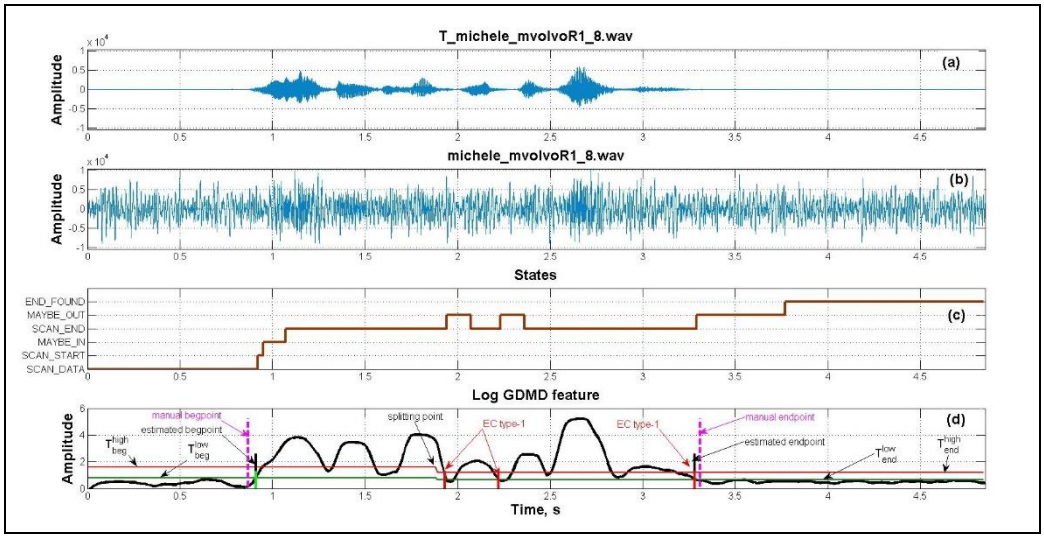

Fig. 3.13. Example from the SpEAR database: (a) clean signal; (b) noisy version; (c) the state transition timing diagram; (d) log-GDMD feature contour with marked some details in temporal execution of the algorithm.

3.5. Endpoint detectors

3.5.1. GDMD-E detector

This detector is a combination of the log-GDMD feature (§ 2.2.4), adaptive thresholds algorithm (§ 3.3.2) and the finite state machine (§ 3.3.3). The block diagram of the proposed ED-algorithm is shown in Fig. 3.15.

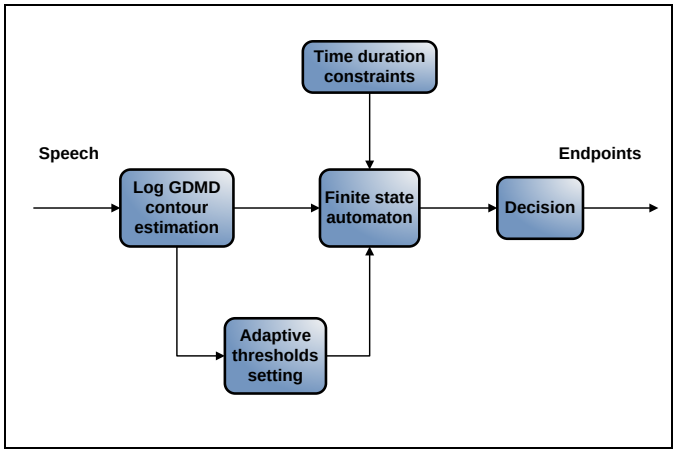

Fig. 3.15. The block diagram of the GDMD-E detector.

3.5.2. LTSD-E detector

This detector is proposed to test the performance of the LTSD feature alone. Typically, VAD-LTSD algorithm includes its own adaptive threshold and hangover scheme [Ramirez et al., 2004]. Here, the LTSD-E detector is designed using the LTSD feature contour and proposed in this paragraph, adaptive thresholds and finite state machine.

3.5.3. GDMD-H detector

This detector is proposed in order to study the join operation of the hangover algorithm and log-GDMD feature contour. The block diagram of the proposed ED-algorithm is shown in Fig. 3.17.

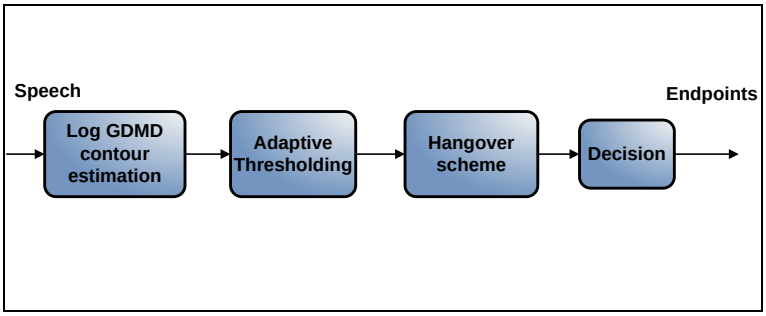

Fig. 3.17. The block diagram of the GDMD-H detector.

3.6. Experiments

3.6.1. Speech data

The speech data used in the experiments are selected from the BG-SRDat corpus [Ouzounov, 2003] and the TIDIGITS corpus [Dan Ellis, online]. In the first experiment – accuracy evaluation – the data are chosen from both data sources, while in the second experiment – verification task –they are selected only from the former one. From BG-SRDat corpus short records are selected. The length of the utterance is about 2 sec, and the length of the single record (file) is about 2.5-3 sec. The speech data used in the study include 337 files collected from 18 male speakers. From the TIDIGITS corpus (in English) are selected examples containing spoken digit strings. The speech data used in the study include 84 files collected from 3 male and three female speakers. The hand labeling of the endpoints for all speech data is done in order to have reference endpoints for comparative purposes.

3.6.2. Algorithms settings

The endpoint detectors parameters are tuned in the study only in the endpoint accuracy experiments. Thus leads to a maximum rate of distribution for frame differences less than 10-frames. The tuned parameters are later used in the speaker verification tests. All adjustments are performed experimentally using a trial-and-error approach.

3.6.3. Detection accuracy estimation

In this experiment the endpoints accuracy was evaluated in terms of frames difference between hand-labelled and detected endpoints [Yamamoto et al., 2006]. The frames difference $D_B(s)$ between hand-labelled and detected beginning points is defined as (for each utterance)

$$D_B(s) = M_B(s) - ED_B(s), \tag{3.30}$$

where $M_B(s)$ is the hand-labelled beginning point; $ED_B(s)$ is the beginning point obtained by endpoint detection algorithm and $s = 1,\ldots,S$ is the number of utterances. The frames difference for ending points $D_E(s)$ is defined as

$$D_E(s) = M_E(s) - ED_E(s), \tag{3.31}$$

where $M_E(s)$ is the hand-labelled ending point; $ED_E(s)$ is the ending point obtained by endpoint detection algorithm.

Detection accuracy analysis is performed by plotting histograms of the frames differences - separately for beginning and ending points. The data points (phrases) from each corpus used for the histograms' creation are 84 and 262 and the final numbers of bins are 9 and 19, respectively. These numbers are the averages of the number of bins calculated for each feature by the Scott's standard reference rule [Scott, 2010].

In Table 3.4 the statistical information of the histograms is presented – each value shows the rate of distribution in percentages for all test conditions. The absolute values of the differences are denoted in Table 3.4 as $|D_B|$ and $|D_E|$ while with $\overline{D}$ are denoted their average values for the particular feature and the corresponding frame difference.

Table 3.4. Rate of the distribution

Speech corpus		BG-SRDat					
No.	**Features & adapt2thr**	$\|D_B\|$		$\|D_E\|$		$\bar{D}$	
		≤ 5	**≤ 10**	**≤ 5**	**≤10**	**≤ 5**	**≤ 10**
1	**log-MTE-E**	56.10	71.37	**55.72**	77.09	55.91	74.23
2	**log-EE-E**	49.61	65.26	36.64	58.01	43.12	61.64
3	**log-MD-E**	**60.30**	80.91	54.96	74.42	**57.63**	77.67
4	**log-GDMD-E**	54.96	**87.02**	51.90	**78.24**	53.43	**82.63**
5	**LTSD-E**	41.60	84.35	37.02	67.17	39.31	75.76
6	**log-GDMD-H**	47.32	**87.02**	35.11	61.06	41.22	74.04
7	**LTSD-H**	45.80	85.11	24.42	46.56	35.11	65.83

The best results (based on $\bar{D}$) are obtained for log-scale features with combination with the adaptive thresholds algorithm. The following four ED are selected: log-GDMD-E& adapt2thr, log-GDMD-H&adapt2thr, LTSD-E&adapt2thr and LTSD-H described in [Luengo et al., 2010, §2] and for them are plotted commonly *stacked* histograms in Figs. 3.18-3.19. Two labels – *skip* and *add* – are added to the histograms. They are used to denote the areas in histograms corresponding to the skipped or the added frames.

The used phrase begins with the following two phonemes *'з'* and *'д'* (it is the Bulgarian word 'здравей-*'zdravei'*). The stacked histogram in Fig.3.18 has two modes. This occurrence is based on the fact that for some records all algorithms skip the voiced fricative *'з'* and set the beginning point at the voiced stop consonant *'д'* (after the voice bar). These errors correspond to the left mode with a difference of about [-5] frames. The right mode (difference about [+5] frames) is a result of added noisy segments before the first phoneme *'з'*, because of the log-scale feature which amplifies low-level contour values. The phrase ended with unvoiced fricative *'с'*, which is difficult to detect in telephone records due to its noise-like characteristics. In this case, in Fig. 3.19 there is a maximum at frame difference [- 5] frames. This means that adding of noisy frames at the end of the phrase is observed. In the histogram a significant value exists at frames difference equal or greater than [-20] as the contribution of the LTSD-H detector being the largest.

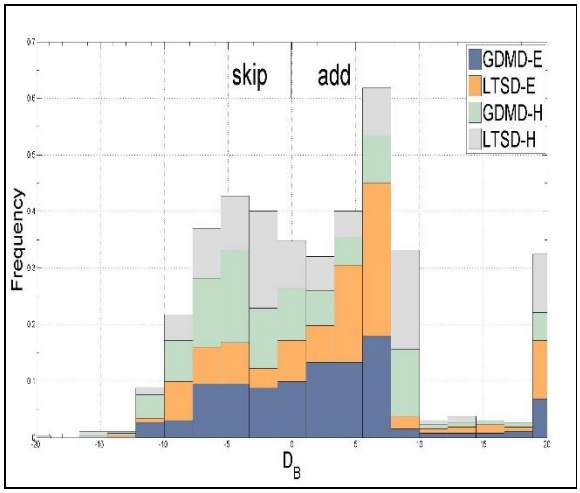

Fig. 3.18. The histograms of DB - BG-SRDat

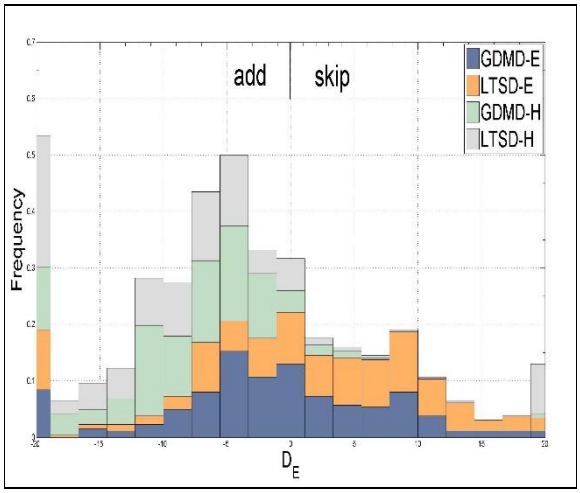

Fig.3.19. The histograms of DE - BG-SRDat

3.6.4. Text-dependent speaker verification

The performance of various endpoint detectors described in § 3.5.3 is compared via the verification results, i.e., for each ED algorithm, a separate speaker verification task is carried out. The additional verification task is done with hand-labelled endpoints. Two different

algorithms are applied in speaker verification tasks - DTW and HMMs. The tests are conducted with short phrases selected from the BG-SRDat corpus.

3.6.4.1. Pre-processing

In the pre-processing step, the Hamming-windowed frames of 30 ms with a rate of 10 ms are used. The number of the Mel-Frequency Cepstral Coefficients (MFCC) is 14 (with 24 Mel filters of equal area), and the cepstral normalization is applied for each file separately [Ganchev, 2011].

3.6.4.2. Speaker verification via DTW

3.6.4.3. Speaker verification via HMM

The phrase modeling is done by a whole-phrase continuous HMM [Buyuk et al., 2012]. The selected model is with a left-to-right topology with no skip state and the output distributions are represented as a mixture of Gaussians with diagonal covariance matrices. Well-known Baum-Welch Algorithm [Gales, et al., 2008] carries out the HMM training. In the verification, the individual speaker's thresholds are used. They are estimated by using the world (or background) model as a non-speaker model. The speaker's score is obtained by computing the log-likelihood ratio of the particular utterance using the speaker and world models. The verification thresholds are set a priori based on distributions of the scores from claimed speakers and impostors [Munteanu et al., 2010].

3.6.4.4. Speech data used in verification

The speech data used in the speech verification experiments include 337 records of the phrase collected from 18 male speakers. The more significant part - 262 records from 12 speakers (these data are the same for both applications) is intended for models forming (training set), for thresholds settings (validation set) and testing (verification set). Because the speech corpus is small, the same data set is used for training and validation [Bengio et al., 2004]. The rest of the data – 75 records from 6 speakers are selected for the UBM training in the HMM test.

The 5×2 fold cross validation method is applied in order to make efficient use of all available data [Kuncheva, 2014]. Overall results are computed as weighted means of the outcomes from the five repetitions. In the verification mode, there are 142 client accesses or False Rejection (FR) tests and 1562 impostor accesses or False Acceptance (FA) tests. After five runs, the total tests are for false rejection – 710, and false acceptance – 7810.

3.6.4.5.Experimental results

It is known that for limited real-world data, the single value error is not a reliable estimation of the speaker verification performance [Bengio et al., 2004]. Since this is our case, it was decided to apply the methodology for performance estimation of the speaker verification proposed in [Bengio et al., 2004]. The verification results are presented as rate ratios – False Rejection Rate (FRR), False Acceptance Rate (FAR), and the Half Total Error Rate (HTER). The Z_{HTER} -test method proposed in [Bengio et al., 2004] is applied to verify whether the given classifier is statistically significantly different from another. The minimal verification error (HTER = 8.42%) for hand-labelled utterances is obtained for the left-to-right HMM with 35 states and 2 Gaussian mixtures, and this topology is used in all experiments. The HMM speaker verification results are shown in Table 3.10– the rates and the confidence intervals for the HTERs.

Table 3.10. HMM speaker verification errors

No.	Endpoint detector	FRR, %	FAR, %	HTER, %	95%CI
1	**Manual**	15.63	1.21	8.42	±0.0134
2	**log-GDMD-E**	18.45	0.98	9.71	±0.0143
3	**LTSD-E**	22.25	1.20	11.72	±0.0153
4	**log-GDMD-H**	18.45	1.02	9.73	±0.0143
5	**LTSD-H**	22.53	1.04	11.78	±0.0154

3.7. Conclusions

Based on the experiments, the following conclusions are made. **The first**, the log-GDMD-based endpoint detectors always (in all tests) perform better than the LTSD-based ones. **The second**, in the endpoint detection accuracy tests, the state automaton with the adaptive threshold scheme outperforms the hangover scheme for the same features. **The third**, in speaker verification tests for the same features, the state automaton with adaptive threshold scheme mostly outperforms the hangover scheme in terms of verification rate, but the difference between them is not statistically significant.

CHAPTER 4. VAD algorithms in text-independent speaker identification. The experimental study

4.1. Introduction

Voice Activity Detection (VAD) is the task of determining the existence of speech fragments in the audio stream, and it plays a crucial role in any speech processing system. It is a binary classifier. Despite the widespread use of VAD algorithms, no universal algorithm has been developed to work in a real-world environment reliably.

In this chapter, a comparative experimental analysis is carried out of the effectiveness of the features proposed in Chapter 2. For each feature (reference or proposed by the author), a separate detector is formed, that becomes part of a text-independent speaker identification system implemented via the MLP classifier.

The experimental studies were carried out with two different speech detection algorithms - they will be referred to in the text as VAD-1 and VAD-2. The development of two separate VAD algorithms is required because in Chapter 2 two types of features are proposed – in scalar (VAD-2) and vector form (VAD-1).

VAD-1 is accepted to use a multilayer perceptron as a classifier, and a binary decision is obtained by the output neuron value thresholding. The VAD-2 uses time contours and thresholds (similar to the algorithms discussed in Chapter 3), and in this case, the binary decision is obtained by the feature contour thresholding. Only speech segments obtained by the VADs decisions are sent to the speaker recognition MLP classifier [Kitaoka et al., 2007].

In order to validate the performance of the VAD algorithms, two experiments are carried out. In the first one, the accuracy is evaluated in terms of frames differences between hand-labelled and detected fragments endpoints. The tests are carried out with speech data from the following corpora - TIDIGITS [Dan Ellis, online], NOIZEUS [NOIZEUS, online], and BG-SRDat [Ouzounov, 2003]. In the second experiment, the performance of the VAD algorithms in terms of the recognition rate is estimated via an MLP-based text-independent speaker identification system. The tests are done with data from the BG-SRDat corpus.

4.2. Reference features

In VAD-1 the reference features are Multi-Band Spectral Entropy - MBSE [Misra et al., 2005]; Frequency-Filtering parameter (FF) [Macho et al. 2001]; Relative Spectral Difference - RSD [Macho et al., 2001] and Index-weighted Mel- Frequency Cepstral Coefficients - IW-MFCC) [Ganchev, 2011]. In VAD-2 the reference contours are obtained by Sohn algorithm [Sohn et al., 1999] (discussed in Chapter 1 - §1.2.3.1.1) - here its VoiceBox version is used [VoiceBox, online]; by Wu algorithm [Wu et al., 2006] and by LTSD algorithm discussed in Chapter 3 (§3.2.4) [Ramirez et al., 2004].

4.3. VAD errors

The VAD errors are determined by comparing the manually determined speech fragments endpoints with those obtained by the corresponding detector. They are used to evaluate the properties of different VAD algorithms. Common errors are described in [Davis et al., 2006]. They are: Front-End Clipping (FEC), Mid-speech Clipping (MSC), OVER (overhang**),** Noise Detected as Speech (NDS), Back-End Clipping (BEC), Front-end adding (FEA) and Speech

Detected as Noise (SDN). The detection accuracy is defined in two ways - correctly recognized speech segments or Speech Hit Rate (SHR) and correctly recognized non-speech segments or Non-speech Hit Rate (NHR).

4.4. Performance assessment

The performance of the voice activity detectors as binary classifiers can be assessed by using the ROC (Receiver Operating Characteristics) curves and through the Confusion Matrix (CM). Most often, however, scalar values are introduced, to represent in general form the characteristics of the ROC-curves and CM. Here, similar values are used - in ROC analysis - F-measure and AUC (Area Under Curve) [Fawcett, 2006] and for the CM - Confusion Entropy (CEN) [Wei et al., 2010], [Delgado et al., 2019].

4.4.1. ROC analysis

4.4.2. Confusion matrix

4.5. Text-independent speaker identification

4.6. Speech detector VAD-1

4.6.1. Features selection

In VAD-1, four reference features and two proposed by the author are used. The features proposed by the author are - BMD and MMD feature in § 2.1.4.2-3. The reference ones are MBSE in § 4.2.2, FF in § 4.2.3, RSD in §4.2.3 and IW-MFCC in §4.2.4.

4.6.2. Multilayer perceptron

The MLP is with structure 14-20-1. The network has 20 neurons in one hidden layer and a single output neuron. The activation functions of the neurons are hyperbolic tangent function. The RProp algorithm with the most typical parameter settings is applied according to the recommendation in [Demuth et al., 2009] and [LeCun et al., 2012]. The network is trained in batch mode.

4.6.3. Thresholding

The binary decision is obtained by thresholding of the output neuron value. It is accepted to apply the Otsu's threshold [Kisku et al., 2014], and its calculation is done separately for each file.

4.6.4. Speech data used for VAD-1

The speech data for detection are separated into three groups - for training, validation and testing. The first group includes 24 files and the second – 12 files. For training 70000 speech frames and about 40000 non-speech ones are used. The validation frames are twice smaller. Hand-labelled data are used as targets.

4.6.5. Estimation of detection accuracy

In Table 4.1, the values of AUC, F- measure, VAD- errors, and VAD-accuracy are shown. They are calculated as weighted averages over all 270 tested files.

4.7. Speaker identification system with VAD-1

The text-independent speaker identification system includes three modules - pre-processing, the MLP classifier, and supra-segments decision scheme. The block diagram of the MLP-based speaker identification system with details about VAD-1 algorithm is shown in Fig. 4.3.

4.7.1. Pre-processing

The preprocessing module includes two sub-modules – VAD (§ 4.6) and Mel-cepstrum extractor, with 14 cepstral coefficients obtained by 24 Mel-filters of equal area. Only frames marked as a speech by VAD algorithm are processed.

4.7.2. Multilayer perceptron

The number of speakers is 12, and the architecture of the MLP is 14-120-12. The input vector size is 14, the number of hidden layer neurons is 120, and the number of output neurons is 12. The activation function for all neurons is a hyperbolic tangent. The training is implemented using the RProp algorithm in batch mode with most typical parameter settings according to the recommendations in [Demuth et al., 2009]. To compensate for the effect of MLP random

initialization, the multiple runs scheme is applied and 5x10 scheme is adapted here [Kuncheva, 2014].

4.7.3. Speech data

Speaker recognition data is selected from the BG-SRDat corpus and is divided into three groups – for training, validation and testing. It is accepted to use the same number of speech segments for each class in the training mode. The number of speech segments from one file is limited up to 1300. For each speaker two files or 2600 speech segments are used. These segments are obtained by random selection from all speech segments contained in both files. Validation data is chosen in a similar way, except that only one speaker file is used (i.e. 1300 segments).

4.7.4. Decision scheme

Recognition is accomplished by supra-segment analysis. The length of one supra-segment is 200 segments (2 seconds). It shifts without overlapping along with the speech segments in the test file. The speaker is identified for each supra-segment separately by finding a maximum class value in the mean vector obtained by averaging over all MLP output vectors for the given supra-segment.

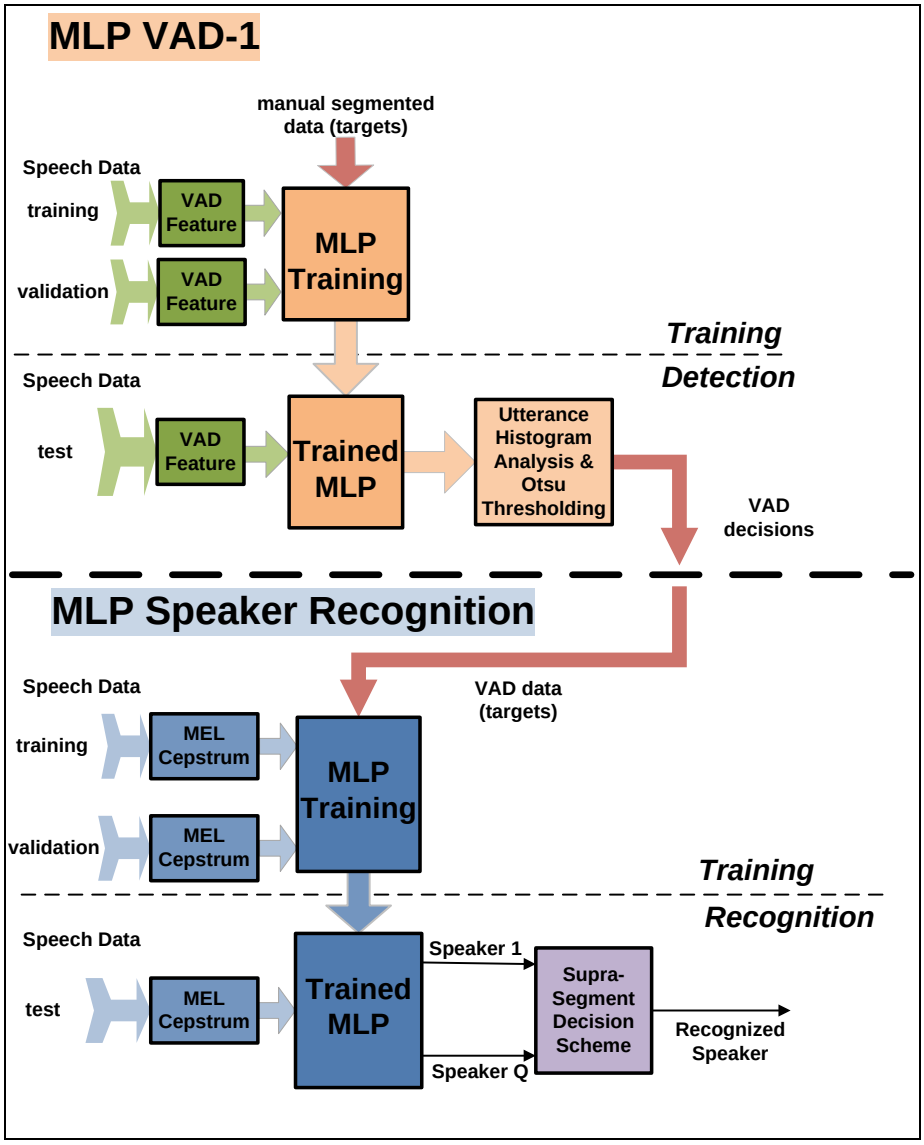

Fig.4.3. The block diagram of the MLP-based speaker identification system with details about VAD-1 algorithm.

Table 4.1. BG-SRDat – VAD-errors and VAD-accuracy in percentage and F-measure and AUC values

		Features					
No.	**Errors**	**BMD**	**MMD**	**IW-MFCC**	**RSD**	**FF**	**MBSE**
1	**NDS**	**3.0460**	3.0811	5.3943	5.5603	5.4806	7.0924
2	**SDN**	1.2672	1.3016	**0.1380**	0.2218	0.2188	0.4281
3	**FEA**	7.3957	7.6403	3.2107	3.4770	3.3037	**2.9677**
4	**MSC**	4.9555	**4.6617**	8.5554	8.0090	7.8394	11.6752
5	**OVER**	4.1887	4.2373	2.5303	2.5641	**2.2968**	2.7926
6	**FEC**	2.8267	**2.6551**	3.1080	3.1519	3.2129	4.4374
7	**BEC**	4.7662	**4.5610**	4.8369	5.2852	5.3564	6.1939
	Accuracy	**BMD**	**MMD**	**IW-MFCC**	**RSD**	**FF**	**MBSE**
1	**SHR**	86.0038	**86.6680**	83.3382	83.3661	83.3671	77.2084
2	**NHR**	81.3985	80.9349	**88.0417**	87.0674	87.7306	85.6802
3	**F-Measure**	0.8753	**0.8780**	0.8768	0.8745	0.8762	0.8334
4	**AUC**	0.9028	0.9043	**0.9245**	0.9183	0.9221	0.8701

4.7.5. Experimental results

In Table 4.8 the confusion entropy for each feature, for each speaker and its overall values are shown. The recognition error is included in the last row of the table. Table 4.8 shows the consistency of recognition error and total entropy for the first two and the last two features. Notable in this case is the fact that the smallest recognition error and minimum overall CEN for BMD and MMD are partly in line with the results obtained in accuracy detection tests, namely the maximum values of F- measure and SHR observed in Table 4.1 for the MMD feature.

Table 4.8. Overall entropy and entropy for each speaker

		Features					
No.	**CEN**	**BMD**	**MMD**	**IW-MFCC**	**RSD**	**FF**	**MBSE**
1	**Sp #1**	0.1149	0.1378	**0.1041**	0.1353	0.1997	0.1902
2	**Sp #2**	0.1270	**0.1229**	0.1333	0.1487	0.1835	0.2165
3	**Sp #3**	0.0354	**0.0321**	0.0611	0.0645	0.0710	0.0805
4	**Sp #4**	**0.2719**	0.2903	0.3649	0.2841	0.3849	0.3276
5	**Sp #5**	**0.2437**	0.2874	0.2779	0.2615	0.2724	0.2944
6	**Sp #6**	**0.2622**	0.2962	0.3577	0.3444	0.3613	0.3579
7	**Sp #7**	0.1795	**0.1592**	0.1687	0.1652	0.1667	0.1928
8	**Sp #8**	**0.1244**	0.1373	0.2195	0.2139	0.2008	0.3118
9	**Sp #9**	**0.1557**	0.1663	0.2089	0.2763	0.2802	0.3668
10	**Sp #10**	**0.2628**	0.2690	0.4029	0.4254	0.4195	0.4100
11	**Sp #11**	0.1062	0.1340	**0.1061**	0.1301	0.1233	0.1207
12	**Sp #12**	0.0615	**0.0448**	0.0463	0.0590	0.0749	0.0567
13	**Overall CEN**	**0.1582**	0.1686	0.1917	0.1913	0.2124	0.2191
14	**Recog.Err.[%]**	**14.46**	15.94	18.87	19.63	21.83	25.19

4.8. Speech detector VAD-2

The VAD-2 algorithm proposed in the paper includes two steps – feature extraction and thresholding scheme.

4.8.1. Feature extraction

Here three reference features and one proposed by the author are used. A separate speech detector is designed for each feature. The feature proposed by the author is - log-GDMD in §2.2.4 and the reference ones are $\Gamma(m)$ in formula (1.25) - obtained by Sohn's algorithm in

§1.2.3.1.1; SAE feature - by the Wu's algorithm in §4.2.1 and the LTSD feature described in §3.2.4.

4.8.2. Thresholding scheme

It is accepted to use thresholds obtained by the algorithms, proposed by the author in §3.4.1 and §3.4.2 - fixed and adaptive thresholds, respectively. Fixed thresholds are applied to the log-GDMD, Sohn and SAE. The LTSD algorithm includes its threshold. The adaptive threshold in §3.4.2 is used only for the log-GDMD feature (it is noted separately).

4.8.3. Speech data used for VAD-2 accuracy estimation

The speech data used in the VAD-2 accuracy estimation are different from that in the VAD-1. The main reason is that VAD-1 is implemented as a classifier and it requires training, validation and testing data - that's why it used data only from BG-SRDat. For VAD-2, which has threshold logic, it is accepted, except data from BG-SRDat corpus, to use data in English from two corpora - Dan Ellis [Dan Ellis, online] and NOIZEUS [NOIZEUS, online].

4.8.4. Study of detection accuracy

The detection results obtained for the NOIZEUS (Table 4.11) show that the maximum AUC value is obtained for the log-GDMD feature in four of the eight types of noise. For other four types of noise, the maximum AUC value belongs to the LTSD feature. According to the results shown in Table 4.12 (Dan Ellis' corpus), the AUC value is maximum at the log-GDMD feature.

4.9. Speaker identification system with VAD-2

The speaker identification system used in VAD-2 analysis is the same as that described in §4.7.

4.9.1. Speaker identification results

For analyzed features in Table 4.13 are shown the values of CEN and recognition errors (in percentages) for different thresholds (BG-SRDat data).

Table 4.11. Corpus NOIZEUS - AUC values for different types of noise at SNR = 5 dB

	Features	Airport	Babble	Car	Exhibition	Restaurant	Station	Street	Train
1	**log-GDMD**	0.8011	0.8103	0.8280	**0.8492**	0.8198	**0.8199**	**0.7890**	**0.8156**
2	**Sohn**	0.7806	0.776	0.7603	0.8295	0.8036	0.7703	0.7782	0.7763
3	**Wu**	0.7511	0.7885	0.8239	0.8341	0.7996	0.7973	0.7600	0.8016
4	**LTSD**	**0.8112**	**0.8119**	**0.8361**	0.8420	**0.8228**	0.8139	0.7633	0.7944

Table 4.12. Corpus DanEllis – AUC values

	Features	AUC
1	**log-GDMD**	**0.9068**
2	**Sohn**	0.8958
3	**Wu**	0.7802
4	**LTSD**	0.7988

Table 4.13. Corpus BG-SRDat – values of CEN and recognition error in percentages

No.	Features		fixed1thr				
			0.1	**0.2**	**0.3**	**0.4**	**0.5**
1	**log-GDMD**	**Err**	15.19	14.07	**13.88**	14.00	16.93
		CEN	0.1594	0.1437	**0.1436**	0.1463	0.1719
2	**Wu**	**Err**	19.87	**16.38**	19.16	18.04	20.74
		CEN	0.1933	**0.1718**	0.1832	0.1857	0.2026
3	**LTSD + HangETSI**	**Err**	17.79				
		CEN	0.2438				
5	**log-GDMD + adapt1thr**	**Err**	13.35				
		CEN	0.1432				
6	**Sohn**		**fixed1thr**				
			0.3	**0.4**	**0.5**	**0.6**	**0.7**
		Err	18.81	19.94	19.07	**17.63**	17.69
		CEN	0.1860	0.2002	0.1913	**0.1831**	0.1868

4.10. Conclusions

The following conclusions based on the obtained experimental results are made:

• **VAD-1** - Speaker identification error, as well as CEN has the minimum values for the proposed features BMD and MMD.

• **VAD-2** - In most tests, the log-GDMD feature outperforms those based on algorithms of Sohn, Wu, and LTSD. It is necessary to note that the LTSD feature is adaptive to the varying noise levels, and Sohn's algorithm included the noise estimation procedure, while log-GDMD relies only on the internal robustness of its two components - the modified group delay spectrum and the delta spectral autocorrelation function. This robustness is based on the properties of the derivatives - negative derivative of the Fourier transform phase and the first derivative of the spectral autocorrelation function.

CHAPTER 5. BG-SRDat – Telephone speech corpus intended for speaker recognition

5.1. Introduction

Here the BG-SRDat corpus (Bulgarian language Speaker Recognition DATa) containing speech recorded over telephone lines (landline, cellphones, VoIP) and including short phrases, reading text and conversations in Bulgarian and only short phrases in English is described. The main efforts were to build a corpus with great variety in the characteristics of the communication environment, i.e., different telephone lines, different speaker locations, different background noise and more. The corpus contains 630 records of different duration, collected by 40 male speakers. The initial version of this corpus is described in [Ouzounov, 2003].

5.2. General characteristics of speech corpora for speaker recognition

5.3. Brief description of popular speaker recognition corpora

5.4. Description of the BG-SRDat

According to the type of speech material, the BG-SRDat can be considered as consisting of five modules (Speech Data 1, 2… 5), which are:

- SD1 (BG) - contains reading newspaper text - average record duration of the reading text is about 40 seconds. Two types of records of the same text have been made, respectively by a microphone (26 speakers with 28 records) and by landline phone (30 speakers with 60 records) - 26 of the speakers are identical in both type of records;
- SD2 (BG) - contains a short phrase – there are 373 records from 20 speakers made by landline phones and cellphones. The phrase is (with Latin letters): *"Zdravei Manolov. Kak se chuvstvash dnes?"*. Its English meaning is "*Hello Manolov! How are you today?"*. The author proposes this phrase mainly for the reason that it contains consonants predominantly. The phrase includes 31 phonemes, 10 of which are vowels and 21 consonants - this phoneme content makes it difficult to recognize the speakers in phone lines;
- SD3 (BG) - contains reading newspaper texts (different paragraphs) - average record duration of the reading text is about 80 seconds. There are 14 records from 10 speakers made by landline phones and cellphones. The paragraphs are selected in such a way to achieve some degree of lexical diversity;
- SD4 (BG) – contains conversations (talks about random topics) with a maximum length of about 7 minutes. There are 4 records from 4 speakers made by cellphones and VoIP;
- SD5 (EN) - contains a short phrase in English. There are 150 records from 9 speakers made by landline phones.

In the corpus description are included four attributes: 1) type of speech (fixed-phrase, reading text, etc.); 2) number and time separation of sessions; 3) recording environments; and 4) files description.

5.4.2. Number of sessions and the period between them

- SD1 - at least two sessions per speaker have been made with only one record per session. The interval between sessions is about three months.
- SD2 - the speech material contains at least ten sessions per speaker with at least two records per session. In each session, the records are made in one day, but the calls are from phone numbers placed at the different locations in Sofia and the country (for landline phones). The interval between sessions is about a week.
- SD3 - for some of the speakers are made two sessions with one record per each with different texts. The interval between sessions is about a week.
- SD4 - there is only one record per session.
- SD5 - the procedure is the same as that used with SD2.

5.4.3. Recording environments

The main efforts were to build a corpus with great variety in the characteristics of the communication environment, i.e., different telephone lines, different speaker locations, different background noise, and more. As a result, the speakers make phone calls from different places - quiet/noisy office, halls, telephone booths located on noisy streets, and more. When the cellphones are used, the calls are made through different model mobile phones and speakers, in most cases, move near high-traffic boulevards or highways.

5.4.4. Files description

Each file (a record) is formed a metadata structure that contains information as speaker ID, speech data type, calling place, phone type, noise type and more. Now, only cellphone data structures are gradually entered into MySQL database.

5.5. Application of BG-SRDat

The corpus has been used for fixed-text speaker verification, text-independent speaker identification, and speech detection. The primary trend in the future development of the corpus will be its transformation into a cellphone speech data corpus.

Author's contributions to the thesis

Scientific contributions:

1. A method for so-called delta spectral autocorrelation function estimation by applying a delta filter on the spectral autocorrelation function is proposed. The efficiency of this filtration, which significantly enhances the harmonic structure of the speech signal in the frequency domain has been demonstrated (Chapter 2, §2.1.3).
2. An approach for speech detection features estimation based on the properties of delta spectral autocorrelation function is proposed. Based on this approach three features are developed. The first one (MD feature) is in scalar form and is intended for speech detection by contour analysis, in contrast the others (BMD and MMD features) which are vectors and are intended for speech detection by recognition algorithms (Chapter 2, §2.1.4).
3. A theoretical analysis of the group delay spectrum for noisy speech signals has been performed. This analysis was indirectly done, by examining the arguments of the projection distortion measures based on the additive spectral model (Chapter 2, §2.2.3).

4. An approach for speech detection features estimation based on the combination of the delta spectral autocorrelation function and the modified group delay spectrum is proposed. Based on this approach two features (lin-GDMD and log-GDMD) are developed. They are intended for speech detection by contour analysis (Chapter 2, §2.2.4).
5. An approach for short phrase endpoint detection is proposed which includes algorithm for adaptive thresholds settings and finite state machine (Chapter 3, §3.4.2-3).

Scientific and applied contributions:

1. A comparative experimental analysis is done for the features proposed in Chapter 2 and some reference ones. The comparison is based on the Euclidean distance between Z-normalized contours calculated for each feature and clear and noisy speech signals (Chapter 3, §3.3).
2. Three algorithms for contour-based endpoint detection based on the proposed approach are developed (Chapter 3, §3.5).
3. A comparative experimental analysis of the accuracy of the developed algorithms is performed by histogram analysis of the differences between the hand-labelled and detected endpoints. The experiments are implemented with noisy speech data in Bulgarian and English (Chapter 3, §3.5 and §3.6.3).
4. A comparative experimental analysis is done for the features proposed in Chapter 2 and the reference ones. The comparison is based on the recognition errors obtained in two systems for text-dependent speaker verification based on the DTW and HMM algorithms, respectively. The used speech data in these experiments are selected from telephone speech corpus in Bulgarian (Chapter 3, §3.6.4).
5. Two algorithms for voice activity detection (VAD-1 and VAD-2) are developed using the parameters, from Chapter 2. The VAD-1 uses for speech detection the MLP classifier, whereas VAD-2 uses contour analysis (Chapter 4, §4.6, and §4.8).
6. A comparative experimental analysis is done for the features proposed in Chapter 2 and the reference ones used in VAD-1 and VAD-2. The comparison is based on the segment detection errors and the binary classification accuracy. The experiments are implemented with noisy speech data in Bulgarian and English (Chapter 4, §4.6.5, and §4.8.4).
7. A comparative experimental analysis is carried out for the features proposed in Chapter 2 and the reference ones used in VAD-1 and VAD-2. The comparison is based on the recognition errors obtained in the text-independent speaker identification tasks implemented by the MLP classifier. The used speech data in these experiments are selected from the Bulgarian corpus (Chapter 4, §4.7, and § 4.9).

Conclusions and ideas for future work

In the thesis, a method for so-called delta spectral autocorrelation function estimation is proposed. This function was obtained by applying a delta filter to the spectral autocorrelation function. Five speech detection features based only on its properties from one side, and by combining it with the modified group delay spectrum, on the other side, are proposed. These features are MD, BMD, MMD, log-GDMD, and lin-GDMD. The proposed features are used in ED- and in VAD-algorithms. These algorithms are included in the speech detectors, which are parts of the text-dependent and the text-independent speaker recognition systems. The performance of these detectors was experimentally compared with that obtained by speech detectors with reference features. The comparison was made in two stages. In the first stage, a comparison was made using detection accuracy while in the second one by using the

recognition errors. In ED-algorithms used in fixed-phrase speaker verification tasks dominant in terms of detection accuracy and minimum recognition error is the log-GDMD feature. In text-independent speaker identification tasks, the minimum recognition error was obtained when using VAD algorithms with BMD and log-GDMD parameters, respectively.

Future work in speech detection will focus on the development of the hybrid VAD-algorithms. This involves a fusion of different representations of speech signal, a fusion of multiple feature streams in one VAD algorithm, and combination of different VAD algorithms. In turn, these VAD algorithms can be built with different classifiers, which give an opportunity for greater adaptability of the detection in changed environmental conditions.

Acknowledgments

I would like to express my sincere gratitude to my consultant Assoc. Prof. Georgi Gluhchev for useful discussion and guidance during the preparation of my dissertation. Also, I would also like to thank Assoc. Prof. Bozhan Zhechev for the comments and useful suggestions.

I would also like to thank my family, because, without their faith and support, this project would never be completed.

List of printed scientific publications on the dissertation:

1. **Ouzounov A**., Cepstral Features and Text-Dependent Speaker Identification - A Comparative Study, *Cybernetics and Information Technologies,* vol. 10, No. 1, 2010, pp. 1-12, **Referenced in Web of Science.**
2. **Ouzounov A.,** Telephone Speech Endpoint Detection using Mean-Delta Feature, *Cybernetics and Information Technologies,* vol. 14, No. 2, 2014, pp. 127-139; **Referenced in Scopus, SJR=0.138**.
3. **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; **Referenced in Scopus, SJR=0.252.**
4. **Ouzounov A.**, Mean-Delta Features for Telephone Speech Endpoint Detection, In: Proc. of the International Conference Automatics & Informatics, 2015, pp.185-188.
5. **Ouzounov A.,** Mean-Delta Features for Telephone Speech Endpoint Detection, Information Technologies and Control, No. 3-4, 2014, pp. 36-43.
6. **Ouzounov A.,** LTSD and GDMD features for Telephone Speech Endpoint Detection, *Cybernetics and Information Technologies,* vol. 17, No. 4, 2017, pp. 114-133; **Referenced in Scopus, SJR=0.204.**

Citations of dissertation publications

Cited article:

Ouzounov A., Cepstral Features and Text-Dependent Speaker Identification – A Comparative Study, Cybernetics and Information Technologies, vol.10, No.1, 2010, pp.1-12, Referenced in Web of Science.

Citations:

1. Mishra P., Agrawal S., Recognition Of Voice Using Mel Cepstral Coefficient & Vector Quantization, *International Journal of Engineering Research and Applications (IJERA,)* Vol. 2, Issue 2, Mar-Apr 2012, pp.933-938.: ISSN: 2248-9622. *http://www.ijera.com/papers/Vol2_issue2/FA22933938.pdf*
2. Mishra P., Agrawal S., Recognition of Speaker Using Mel Frequency Cepstral Coefficient & Vector Quantization, *International Journal of Science, Engineering and Technology Research (IJSETR*), Volume 1, Issue 6, December 2012, pp.12-17, ISSN: 2278 – 7798. *http://ijsetr.org/wp-content/uploads/2013/08/IJSETR-VOL-1-ISSUE-6-12-17.pdf*
3. Mishra P., Agrawal S., Recognition of Speaker Using Mel Frequency Cepstral Coefficient & Vector Quantization for Authentication, *International Journal of Scientific & Engineering Research (IJSER),* Volume 3, Issue 8, August 2012, pp.1-6, ISSN 2229-5518. *https://www.ijser.org/onlineResearchPaperViewer.aspx?Recognotion-of-Speaker-Useing-Mel-Cepstral-Coefficient-Vector-Quantization-for-Authentication.pdf*
4. Бакина И. Г., Морфологическое сравнение изображений гибких объектов на основе циркулярных моделей при биометрической идентификации личности по форме ладони, Диссертация - кандидат физико-математических наук, Московский государственный университет имени М. В. Ломоносова, 2011. Научная библиотека диссертаций и авторефератов disserCat: http://www.dissercat.com/content/morfologicheskoe-sravnenie-izobrazhenii-gibkikh-obektov-na-osnove-tsirkulyarnykh-modelei-pri
5. Jain A. and O. Sharma, A Vector Quantization Approach for Voice Recognition Using Mel Frequency Cepstral Coefficient (MFCC): A Review, *International Journal of Electronics & Communication Technology*,vol.4, Issue SPL-4, 2013, pp.26-29, ISSN : 2230-7109 (Online), | ISSN : 2230-9543 (Print). http://www.iject.org/vol4/spl4/c0128.pdf
6. Chen Y., E. Heimark, D. Gligoroski, Personal Threshold in a Small Scale Text-Dependent Speaker Recognition, International Symposium on Biometrics and Security Technologies (ISBAST), July, 2013, pp.162-170, DOI: 10.1109/ISBAST.2013.29, Print ISBN: 978-0-7695-5010-7. Publisher: IEEE. *http://ieeexplore.ieee.org/xpl/abstractReferences.jsp?tp=&arnumber=6597684&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6597684*
7. Kumar, R.C.P., D.A. Chandy, Audio retrieval using timbral feature, International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), March 25-26, 2013, pp. 222-226, DOI: 10.1109/ICE-CCN.2013.6528497, Print ISBN: 978-1-4673-5037-2. Publisher: IEEE. http://ieeexplore.ieee.org/xpl/abstractReferences.jsp?tp=&arnumber=6528497&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6528497
8. Thakur A., R. Kumar, A. Bath, and J. Sharma, Automatic Control of Instruments Using Efficient Speech Recognition Algorithm, International Journal of Electrical & Electronics Engineering, Vol. 1, Spl. Issue 1, March, 2014, pp.16-22, e-ISSN: 1694-2310, p-ISSN: 1694-2426. http://ijeee-apm.com/Uploads/Media/Journal/20140328142123_ACCT_14_IG_52.pdf
9. Kumar C. P. R., S. Suguna and J. Becky Elfreda, Audio Retrieval based on Cepstral Feature, *International Journal of Computer Applications* 107(8):28-33, December 2014. DOI:10.5120/18774-0079; ISSN 0975 – 8887. https://research.ijcaonline.org/volume107/number8/pxc3900079.pdf
10. Kamil I., K. Oyeyiola, Comparative Study on the Performance of Mel-Frequency Cepstral Coefficients and Linear Prediction Cepstral Coefficients under different Speaker's

Conditions, *International Journal of Computer Applications*, March 2014, vol.90, No.11, pp.38-42. DOI: 10.5120/15767-4460; ISSN 0975 – 8887. https://research.ijcaonline.org/volume90/number11/pxc3894460.pdf

11. R. Christopher Praveen Kumar and S. Suguna, Analysis of MEL based features for audio retrieval, *ARPN Journal of Engineering and Applied Sciences*, Vol. 10, No. 5, March 2015, pp.2167-2171. ISSN 1819-6608. http://www.arpnjournals.com/jeas/research_papers/rp_2015/jeas_0315_1735.pdf
12. Patil C. and G. Dhoot, A Security System by using Face and Speech Detection, International Journal of Current Engineering and Technology, vol.4, No.3 (June 2014), pp.2176-2182; E-ISSN 2277 – 4106, P-ISSN 2347 – 5161. https://inpressco.com/wp-content/uploads/2014/06/Paper1922176-2182.pdf
13. Jain, A., Sharma, O.P., Evaluation of MFCC for speaker verification on various windows, Int. Conf. on Recent Advances and Innovations in Engineering (ICRAIE), 2014, pp.1-6, Print ISBN:978-1-4799-4041-7; DOI:10.1109/ICRAIE.2014.6909144; Publisher: IEEE https://ieeexplore.ieee.org/document/6909144
14. Abdelmajid, L., K. Mohamed, Extracting Multi Band Approach of Acoustic Vectors Extractors: Using HMM Classifier, International Journal of Science Technology & Engineering, Vol. 3, No.2, 2016, pp. 305-310; ISSN (online): 2349-784X. https://www.ijste.org/articles/IJSTEV3I2095.pdf
15. Bakina, I., Person recognition by hand shape based on skeleton of hand image, Pattern Recognition and Image Analysis, 2011, vol.21, issue 4, pp.694-704. Print ISSN 1054-6618, Online ISSN 1555-6212, DOI:10.1134/S1054661811040031. https://link.springer.com/article/10.1134/S1054661811040031
16. Trabelsi I., M. Bouhlel, Learning vector quantization for adapted Gaussian mixture models in automatic speaker identification, Journal of Engineering Science and Technology Vol. 12, No. 5 (2017) 1153 – 1164. ISSN: 1823-4690. http://jestec.taylors.edu.my/Vol%2012%20issue%205%20May%202017/12_5_1.pdf
17. Sharma R., R. Bhukya, S. Prasanna, Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification, Speech Communication, Elsevier B.V., vol. 96, 2018, pp. 207-224. https://doi.org/10.1016/j.specom.2017.12.001, ISSN 0167-6393.
18. Кралева, Р., Разпознаване на реч: Корпус от говорима детска реч на български език, Университетско издателство „Неофит Рилски“, 2019; ISBN: 978-954-00-0199-9. https://www.researchgate.net/publication/335739119_Razpoznavane_na_rec_Korpus_ot_govorima_detska_rec_na_blgarski_ezik

Cited article:

Ouzounov A., Telephone Speech Endpoint Detection Using Mean-Delta Feature, Cybernetics and Information Technologies, vol. 14, No. 2, 2014, pp. 127-139, Referenced in Scopus, SJR=0.138.

Citations:

1. Guo Yu, Zhang Erhua, Liu Chi, An endpoint detection algorithm based on frequency-domain characteristics and transition fragment judgment, Journal of Shandong University (Engineering Science), 2016, Vol. 46 Issue (2), pp. 57-63; DOI: 10.6040/j.issn.1672-3961.2.2015.147. https://caod.oriprobe.com/articles/48238509/An_endpoint_detection_algorithm_based_on_frequency.htm

2. Sopon P., J. Polpinij, T. Suksamer, Speech-Based Thai Text Retrieval, The 11th National Conference on Computing and Information Technology, NCCIT'2015, pp. 259-264. http://202.44.34.144/nccitedoc/admin/nccit_files/NCCIT-20150810110354.pdf
3. Li, L., Y.Wang, X.Li, An Improved Wavelet Energy Entropy Algorithm for Speech Endpoint Detection, Journal of Computer Engineering, 2017, vol. 43, No. 5, pp. 268-274, DOI:10.3969/j.issn.1000-3428.2017.05.043; ISSN: 1000-3428. http://manu55.magtech.com.cn/Jwk_ecice/EN/abstract/abstract27746.shtml#
4. Roy T., T.Marwala, S. Chakraverty, Precise detection of speech endpoints dynamically: A wavelet convolution based approach, Communications in Nonlinear Science and Numerical Simulation, Elsevier B. V., 2019, vol. 67, pp. 162-175; ISSN: 1007-5704. https://doi.org/10.1016/j.cnsns.2018.07.008
5. Zhang, X., Q. Xiong, Y. Dai and X. Xu, Voice Biometric Identity Authentication System Based on Android Smart Phone, 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, Dec. 2018, pp. 1440-1444; Electronic ISBN:978-1-5386-8339-2; USB ISBN:978-1-5386-8338-5; Print on Demand(PoD) ISBN:978-1-5386-8340-8. https://doi.org/10.1109/CompComm.2018.8780990;
6. Li, Y., C. L.-F. Cheng, P.-L. Zhang, Speech Endpoint Detection based on Improved Spectral Entropy, Journal of Computer Science, 2016, vol.43, No.11A, pp.233-236; ISSN 1002-137X. http://journal.jsjkx.com/jsjkx/ch/reader/view_abstract.aspx?file_no=201611A053&flag=1

Cited article:

Ouzounov A., Noisy Speech Endpoint Detection Using Robust Feature, Springer International Publishing Switzerland 2014, V. Cantoni et al. (Eds.): BIOMET 2014, LNCS 8897, p. 105–117. Referenced in Scopus, SJR=0.252.

Citations:

1. Li, Y., C. L.-F. Cheng, P.-L. Zhang, Speech Endpoint Detection based on Improved Spectral Entropy, Journal of Computer Science, 2016, vol. 43, No. 11A, pp. 233-236; ISSN 1002-137X. http://journal.jsjkx.com/jsjkx/ch/reader/view_abstract.aspx?file_no=201611A053&flag=1

References

1. Abdulla·, W., Z. Guan, H. Sou, Noise Robust Speech Activity Detection, IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2009, pp. 473-477.
2. Akant, K., R. Pande, S. Limaye, Accurate Monophonic Pitch Tracking Algorithm for QBH and Microtone Research, *The Pacific Journal of Science and Technology*, 2010, vol. 11. No. 2, pp. 342-352.
3. Alam, M., P. Kenny, P. Ouellet, T. Stafylakis, P. Dumouchel, Supervised/Unsupervised Voice Activity Detectors for Text dependent Speaker Recognition on the RSR2015 Corpus, Odyssey 2014: The Speaker and Language Recognition Workshop, pp. 123-130.
4. Bengio, S., J. Mariethoz, A Statistical Significance Test for Person Authentication, In: Proc. of ODYSSEY, The Speaker and Language Recognition Workshop, 2004, pp. 237-244.
5. Buyuk, O., M. Arslan, Model Selection and Score Normalization for Text-Dependent Single Utterance Speaker Verification, Turkish Journal of Electrical Engineering & Computer Sciences, vol. 20, 2012, No. Sup. 2, pp. 1277-1295.
6. Campbell, W., D. Sturim, D. Reynolds, Support Vector Machines Using GMM Supervectors for Speaker Verification IEEE Signal Processing Letters, 2006a, vol. 13, No. 5, pp. 308-311.
7. Cao, D., X, Gao, L. Gao, An Improved Endpoint Detection Algorithm Based on MFCC Cosine Value. Wireless Personal Communications, 2017, vol. 95, pp. 2073–2090.
8. Chen, L. M. Ozsu, V. Oria, Robust and Fast Similarity Search for Moving Object Trajectories, SIGMOD '05: Proceedings of the ACM SIGMOD international conference on Management of data, 2005, pp. 491–502.
9. Chung, H., S. J. Lee, Y. K. Lee, Weighted Finite State Transducer–Based Endpoint Detection Using Probabilistic Decision Logic, ETRI Journal, 2014, 36, pp. 714-720.
10. Dan Ellis's Home Page, Sound Examples for Projects, Columbia University; https://www.ee.columbia.edu/~dpwe/sounds/, last accessed August 2017.
11. Davis, A., S. Nordholm, R. Togneri, Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, 2006, pp. 412-424.
12. Delgado, R., J. Nunez-Gonzalez, Enhancing Confusion Entropy (CEN) for binary and multiclass classification, PLOS ONE, 2019, 14 (1), pp. 1-30.
13. Demuth, H., M. Beale, M. Hagan, Matlab Neural Network Toolbox 6: User's Guide, The MathWorks Inc., 2009.
14. Disken, G., Z. Tufekci, U. Cevik, A robust polynomial regression-based voice activity detector for speaker verification, EURASIP Journal on Audio, Speech, and Music Processing, 2017, vol. 23, pp. 1-16.
15. Dehak, N., P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-End Factor Analysis for Speaker Verification, *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, vol. 19, No. 4, pp. 788-798.
16. ETSI, Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms. ETSI ES 202 050 V1.1.5 (2007-01). Annex A.3. Stage 2 – VAD Logic, pp. 42-43.
17. Fawcett, T., An Introduction to ROC analysis, *Pattern Recognition Letters*, vol. 27, No. 8, 2006, pp. 861–874.
18. Feng, Z., J. Feng, F. Dai, The Application of Extreme Learning Machine and Support Vector Machine in Speech Endpoint Detection, International Journal of Control and Automation, 2016, 9(12), pp.191-202.
19. Fukuda, T., O. Ichikawa, M. Nishimura, Long-Term Spectro-Temporal and Static Harmonic Features for Voice Activity Detection, *IEEE Journal of Selected Topics in Signal Processing*, 2010, vol. 4, No. 5, pp. 834-844.

20. Gales, M., S. Young. The Application of Hidden Markov Models in Speech Recognition, Journal Foundations and Trends in Signal Processing, vol. 1, 2008, No 3, pp. 195-304.
21. Ganchev, T., Contemporary Methods for Speech Parameterization, Springer Briefs in Speech Technology, Springer-Verlag, New York, 2011.
22. Ganapathy, S., S. Thomas, H. Hermansky, Temporal envelope compensation for robust phoneme recognition using modulation spectrum, *Jnl. Acoust. Soc. of America*, Vol. 128 (6), 2010, pp. 3769-3780.
23. Ganapathy, S., P. Rajan, H. Hermansky, Multi-layer Perceptron based Speech Activity Detection for Speaker Verification, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011, pp. 321-324.
24. Garcia-Romero, D., C. Espy-Wilson, Analysis of I-vector Length Normalization in Speaker Recognition Systems, Interspeech, 2011, pp. 249-252.
25. Ghaemmaghami, H., B. Baker, R. Vogt, S. Sridharan, Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. In Proceedings of Interspeech, 2010a, pp. 3118-3121.
26. Ghaemmaghami, H., D. Dean, S. Sridharan, I. McCowan, Noise Robust Voice Activity Detection Using Normal Probability Testing and Time-Domain Histogram Analysis, In: Proc. of IEEE ICASSP, 2010b, pp. 4470-4473.
27. Graf, S., T. Herbig, M. Buck, G. Schmidt, Features for voice activity detection: a comparative analysis, *EURASIP Journal on Advances in Signal Processing*, 2015:91, pp.1-15.
28. Gu, L., Zahorian, S.: A new robust algorithm for isolated word endpoint detection, IEEE ICASSP, 2002, vol. 4, pp. 4161-4164.
29. Hain, T. et al., The Development of AMI System for Transcription of Speech in Meetings, Proc. MLMI, 2005, pp. 344-356.
30. Hegde, R., H. Murthy, V. Gadde, Significance of the Modified Group Delay Feature in Speech Recognition, *IEEE Transactions on ASLP*, vol. 15, 2007, No 1, pp. 190-202.
31. Huang, L. and C. Yang, A Novel Approach to Robust Speech Endpoint Detection in Car Environment, In Proc. of the IEEE ICASSP'2000, pp. 1751-1754.
32. Ishizuka, K., T. Nakatani , M. Fujimoto, N. Miyazaki, Noise robust voice activity detection based on periodic to aperiodic component ratio, *Speech Communication*, 52, 2010, pp. 41–60.
33. Jain, A., P. Flynn, A. Ross, Handbook of Biometrics, Springer, 2008.
34. Kinnunen, T., P. Rajan, A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data, Proceedings of the IEEE ICASSP, 2013, pp. 7229-7233.
35. Kisku, D. (Ed.), Gupta, P. (Ed.), Sing, J. (Ed.). Advances in Biometrics for Secure Human Authentication and Recognition. Boca Raton: CRC Press, 2014.
36. Kitaoka, N., K. Yamamoto, T. Kusamizu, S. Nakagawa, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, S. Nakamura, Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance, ASRU-2007, pp. 607-612.
37. Klapuri, A., Qualitative and quantitative aspects in the design of periodicity estimation algorithms, 10th European Signal Processing Conference, 2000, pp. 1-4.
38. Krishnan, S., Padmanabhan, R., Murthy, H.: Robust Voice Activity Detection using Group Delay Functions, In: IEEE International Conference on Industrial Technology, 2006, pp. 2603-2607.
39. Kuncheva, L. Combining Pattern Classifiers: Methods and Algorithms, 2nd Edition, John Wiley & Sons, 2014.
40. Kyriakides, A., C. Pitris, A. Fink, A. Spanias, Isolated word endpoint detection using Time-Frequency Variance Kernels, 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Publisher: IEEE, pp. 1041-1045.
41. LeCun, Y., L. Bottou, G. Orr, K.-R. Müller, Efficient Backprop, Neural Networks, Tricks of the Trade, Lecture Notes in Computer Science LNCS 7700, Springer Verlag, 2012, pp. 9-48.

42. Li, J., P. Zhou, X. Jing, Z. Du, Speech Endpoint Detection method based on TEO in Noisy Environment, *Procedia Engineering*, Elsevier Ltd., vol.29, 2012, pp. 2655-2660.
43. Luengo, I., E. Navas, I. Odriozola, I. Saratxaga, I. Hernaez, I. Sainz, D. Erro, Modified LTSE-VAD Algorithm for Applications Requiring Reduced Silence Frame Misclassification, In: Proc. of International Conference on Language Resources and Evaluation (LREC'10), 2010, pp. 1539-1544.
44. Macho, D., C. Nadeu, Comparison of Spectral Derivative Parameters for Robust Speech Recognition, Eurospeech 2001, pp. 205-208.
45. Mansour, D., B. Juang, A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition, *IEEE Transaction on ASSP*, 1989, 37, No. 11, pp. 1659-1671.
46. Mak, M., H. Yu, A study of voice activity detection techniques for NIST speaker recognition evaluations, *Computer Speech and Language*, 2014, vol. 28, pp. 295–313.
47. McCowan, I., D. Dean, M. McLaren, R. Vogt, S. Sridharan, The Delta-Phase Spectrum with Application to Voice Activity Detection and Speaker Recognition, *IEEE Trans. Audio, Speech and Language Processing*, 2012, vol. 19, no. 7, pp. 2026–2038.
48. Misra, H.; Ikbal, S.; Sivadas, S.; Bourlard, H., Multi-resolution Spectral Entropy Feature for Robust ASR, In the Proceedings of the ICASSP, 2005, pp. 253 – 256.
49. Munteanu, D., S. Toma, Automatic Speaker Verification Experiments Using HMM, In: Proc. of 8th International Conference on Communications, 2010, pp. 107-110.
50. Murthy, H., B. Yegnanarayana, Group delay functions and its applications in speech technology, *Sadhana - Indian Academy of Science, Springer Nature*, vol. 36, part 5, 2011, pp. 745-782.
51. Nautsch, A., R. Bamberger, C. Busch, Decision Robustness of Voice Activity Segmentation in unconstrained mobile Speaker Recognition Environment, 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1-7.
52. NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms, https://ecs.utdallas.edu/loizou/speech/noizeus/, last accessed February 2019.
53. NIST SRE, NIST Speaker Recognition Evaluation, https://www.nist.gov/itl/iad/mig/speaker-recognition, last accessed September 2019.
54. Oppenheim, A., R. Schafer, J. Buck, Discrete-Time Signal Processing, Prentice Hall, 2nd ed., 1999.
55. Ouzounov, A., BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels, *Cybernetics and Information Technologies*, 2003, vol. 3, No. 2, pp. 101-108.
56. Padmanabhan, R., S. Krishnan, H. Murthy, A pattern recognition approach to VAD using modified group delay, in Proc. 14th National conference on Communications, 2008, pp. 432–437.
57. Rabiner, L. and R. W. Schafer, Theory and Application of Digital Speech Processing, Prentice Hall Press, NJ, 2010.
58. Ramirez, J., C. Segura, C. Benitez, A. Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information, *Speech Communications*, 2004, vol. 42, pp. 271–287.
59. Ramirez, J., J. Gorriz, J. Segura, Voice Activity detection. Fundamentals and Speech Recognition System Robustness, In M. Grimm and Kroschel, *Robust speech recognition and Understanding*, 2007, pp. 1-22.
60. Renevey, Ph. and A. Drygajlo, Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech, 2001, pp. 1883-1886
61. Reynolds, D. et al., The 2004 MIT Lincoln laboratory speaker recognition system, Proc. ICASSP, 2005, pp. 177-180.
62. Roach, P., English Phonetics and Phonology: A Practical Course, 4th Ed. Cambridge University Press, 2009.
63. Scott, D., Scott's Rule, WIREs Computational Statistics, vol. 2, 2010, pp. 497-502.
64. Singer. H., T. Umezaki, F. Itakura, Low Bit Quantization of the Smoothed Group Delay Spectrum for Speech Recognition, Proceedings of ICASSP, 1990, pp. 761-764.
65. Sohn, J., N. Kim, and W. Sung, A statistical model based voice activity detection, *IEEE Signal Processing Letters*, 1999, vol. 6, No. 1, pp. 1-3.

66. SpEAR Database, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, http://www.cslu.ogi.edu/nsel/data/SpEAR_lombard.html, last accessed September 2016.
67. Theodoridis, S., K. Koutroumbas, An Introduction to Pattern Recognition: A MATLAB Approach, Academic Press, 2010.
68. Tuononen, M., R. Hautamäki, P. Fränti. Automatic voice activity detection in different speech applications. In Proceedings of the 1st international conference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop, e-Forensics, 2008, pp. 12:1-12:6.
69. VOICEBOX: Speech Processing Toolbox for MATLAB, last accessed September 2018, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
70. Wang, M., Bitzer, D., McAllister, D., Rodman, R., Taylor, J. An Algorithm for V/UV/S Segmentation of Speech. Proceedings of the 2001 International Conference on Speech Processing, pp. 541-546.
71. Wei, J.-M., X.-J. Yuan, Q.-H. Hub, S.-Q. Wang, A novel measure for evaluating classifiers, Elsevier, *Expert Systems with Applications*, 2010, vol. 37, pp. 3799–3809.
72. Wu, B.-F., K.-C. Wang, Robust Endpoint Detection Algorithm based on the Adaptive Band-Partitioning Spectral Entropy in Adverse Environments, *IEEE Transactions on SAP*, vol. 13, No. 5, 2005, pp. 762-775.
73. Wu, B.-F. and K.-C. Wang, Voice Activity Detection based on Auto-Correlation Function using Wavelet Transform and Teager Energy Operator, in *Computational Linguistics and Chinese Language Processing*, 2006, vol. 11, No. 1, pp. 87-100.
74. Wu, G., et al., Fuzzy Neural Networks for Speech Endpoint Detection, In: Proc. of 2012 International Conference on Fuzzy Theory and Its Applications, pp. 354-356
75. Yali, C. et al. A Speech Endpoint Detection Algorithm Based on Wavelet Transforms. – In: Proc. of 26th Chinese Control and Decision Conference (CCDC), 2014, pp. 3010-3012.
76. Yamamoto, K., F. Jabloun, K. Reinhard, A. Kawamura, Robust Endpoint detection for Speech recognition based on Discriminative Feature Extraction, Proc. ICASSP, 2006, pp. 805-808.
77. Yamamoto, H., K. Okabe, and T. Koshinaka, Robust i-vector extraction tightly coupled with voice activity detection using deep neural networks, in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 600– 604.
78. Ying, D., Y. Yan, J. Dang, F. Soong, Voice Activity Detection Based on an Unsupervised Learning Framework, IEEE Trans. on ASLP, 2011, vol. 19, no. 8, pp. 2624– 2633.
79. Yoo, I.-C., H. Lim, D. Yook, Formant-based Robust Voice Activity Detection, *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, No. 12, 2015, pp. 2238-2245.
80. Zhang, Z., S. Furui, Noisy Speech Recognition Based on Robust End-Point Detection and Model Adaptation, Proceedings of IEEE ICASSP, Vol. 1, 2005, pp. 441-444.
81. Zhang, T., H. Huang, L. He, M. Lech, A Robust Speech Endpoint Detection Algorithm Based on Wavelet Packet and Energy Entropy, In: Proc. of 3rd International Conference on Computer Science and Network Technology, 2013, pp. 1050-1054.
82. Zhang, Y., K. Wang, B. Yan, Speech endpoint detection algorithm with low signal-to-noise based on improved conventional spectral entropy, 12th World Congress on Intelligent Control and Automation (WCICA), 2016, pp. 3307-3311.
83. Zhu, D., K. Paliwal, Product of Power Spectrum and Group Delay Function for Speech Recognition, In Proceedings of 2004 IEEE International Conference on ASSP, pp. 125-128.
84. Тилков, Д., Т. Бояджиев, Българска фонетика, София, Наука и Изкуство, 1977.