Influence of adaptive thresholding on peaks detection in audio data

Tomasz Maka¹

Received: 13 May 2019 / Revised: 4 February 2020 / Accepted: 17 February 2020 /
Published online: 19 March 2020
© The Author(s) 2020

Abstract

Many audio analysis systems employ peak picking procedure to produce the final decision. A typical scheme uses a thresholding function to minimise detection errors where its form depends on the structure of the input signal. The paper covers the problem of an adaptive thresholding function estimation. Using the genetic algorithm to optimise the components of the thresholding function we have determined the level of importance of individual local statistics on the final function representation. The proposed method has been used to tune the peak detection procedure to identify the change points in an audio signal. In the result of the heuristic configuration, the best accuracy of segment boundaries have been obtained for thresholding function built on top of two local statistics of the detection function and constant value. Finally, as an example, a comparison with the state-of-the-art scheme for audio segmentation was performed.

Keywords Signal analysis · Adaptive thresholding · Genetic algorithm

1 Introduction

The increasing popularity of machine listening systems makes a demand for robust and computationally efficient algorithms dedicated to analysis, processing and recognition of audio signals. The change points in an audio signal may entail the occurrence of the components like events [18], scene-change moments [7], and sound sources activity [10]. Such elements are an integral part of the auditory scene, and they determine the ultimate effectiveness of an audio analysis system.

The decomposition process of the auditory scene often includes an analysis of abrupt changes in an audio signal to perform segmentation of audio stream. Such a stage may be challenging since it depends on the type of signal and acquisition conditions. A typical approach for searching the change points in an audio signal is based on the calculation of detection function and performing peak-picking process with thresholding stage [19]. The

✉ Tomasz Maka
tmaka@wi.zut.edu.pl

¹ Faculty of Computer Science and Information Technology West Pomeranian University of Technology, Szczecin Żołnierska 52, 71-210, Szczecin, Poland

robustness of peaks detection relies on the thresholding function, but its form is not trivial to estimate. Therefore, in real conditions, an adaptive thresholding function is often selected experimentally by using local statistical properties of the analysed signal.

The peaks detection procedure is an integral part of various signal analysis algorithms. For example in [14] an approach to detection of impulse noise in audio signals is presented. The proposed technique uses adaptive thresholding which can be obtained by averaging the detection function and scaling it. The following detection functions have been introduced and analysed including derivative, RMS and auto-regressive based methods. In the result it has been shown that the fourth derivative is accurate to determine smallest audible errors in the signal. The occurrence of peaks in the signals plays also an essential role in biological signals like QRS complexes of ECG signals or pressure analysis [1]. Another application of peaks detection includes musical sound analysis. For example, in [19] authors presents a method for determining a time start of notes in the audio signal, similarly, the work [4] proposed a neural network approach for detecting musically relevant events in an audio stream. Since peaks detection is an integral part of many signal analysis systems, there are many methods for this task with various accuracy and computational complexity. One of the robust and accurate algorithm for peaks detection was proposed in [22].

The estimation of thresholding functions is important in the audio analysis where the potential peak candidates describe task-related places or properties in the signal. The local threshold value is often used to eliminate the possible irrelevant peaks. An application of such functions in audio analysis includes onset/offset detection [2, 19], voice activity detection [15], audio denoising [24] and reference-free signal-to-noise measurement [11]. The problem of estimation of thresholding values is well known in the image processing domain, and various approaches have been already extensively studied in many works [17, 20, 21, 26]. However, according to the best knowledge of the authors, the heuristic adaptation of the thresholding function for the audio analysis tasks has not been investigated in the literature.

Our study focuses on the components selection of an adaptive thresholding scheme. We have proposed a technique to evaluate the separability of the peaks in the detection function. Then, by using a Monte Carlo simulation, the most discriminative similarity measures were selected. Using metric-based segmentation technique a feature space were selected and then the final performance has been optimized using genetic algorithm. The optimized parameters were exploited to define adaptive thresholding function used for segmentation. Finally, the obtained results have been compared with Δ BIC segmentation scheme.

The paper is organized as follows. In the next section, information about motivation and contribution is provided. Section 3 describes the problem studied in this work. The basic elements of the peaks detection technique using adaptive thresholding are presented in Sections 4–7. The detailed description of the audio data set is described in the following part. Section 9 includes an analysis of detection function, feature space selection and computational cost analysis for considered example. In the next two sections, the adaptive thresholding heuristic analysis and evaluation are presented. Finally, a summary of the study is included in the last section.

2 Motivation and contribution

The motivation behind this study is the problem of automatic generation of thresholding function for peaks selection in the data computed from audio signals. Peaks detection and selection is a significant operation in many audio analysis applications. In real situations, this operation results in many task-unrelated peaks. Therefore, together with this

operation, a stage to filter out irrelevant data is applied. In many proposed solutions, the choice of signal components or decision boundary relies on the thresholding. The calculation of the thresholding function is problematic due to connections with input signal. Therefore many systems introduce tuning phase where the thresholding value has to be selected experimentally, or a set of conditional rules with many coefficients have to be applied. Such configurations have to be corrected accordingly to the input signal and may introduce many detection errors in the detection stage. The natural way seems to be the determine such function based on the variability of the detection signal. The purpose of the study is an estimation of such function based on local changes of the signal by determining the dependencies of statistical properties of regions near analysed peaks.

The main contribution of this work includes:

- A proposition of a mechanism to establish a workflow configuration adapted to the audio domain to estimate the parameters of adaptive thresholding.
- Dynamic threshold estimation using the local changes of the detection function by a weighted sum of selected statistics. The weights were estimated utilising a genetic algorithm for all combinations of the statistical components.
- Analysis of audio feature space to determine the region boundaries for signals with segments of speech, music and environmental sounds.

3 Problem formulation

Since the localisation of the peaks in the signal is an integral part of many audio analysis tasks, the robustness of its detection is essential in the final analysis accuracy. In this work, the peaks detection procedure is performed on the signal called later as detection function. Depending on the type of a task, the detection function can represent various time and frequency properties of the audio signal. In most cases, such function contains peaks in locations reflecting changes in the source signal. Unfortunately, in practice, due to many factors like subtle changes in signal, the type of feature space, and noise, the detection stage may provide incorrect results. This situation is because false peaks can mask the real peaks or many spurious peaks may be detected making the decision process difficult. In order to prevent such a case, a thresholding function is employed to eliminate or diminish detection errors.

The problem of peaks selection using adaptive thresholding can be defined as follows. Let $A = \{a_i : i = 1, \dots, G\}$ denote a set of positions of actual peaks and $D = \{s_i : d(s_i) > T_h(s_i), i = 1, \dots, H\}$ is a set of positions of detected peaks, where $d(\cdot)$ is a detection function, $T_h(\cdot)$ is the thresholding function, G and H are the number of actual and detected peaks respectively. The goal is to maximize F_1 -score [25] between elements of A and D sets by adjusting the weights k_j and constant c of thresholding function expressed as:

$$T_h(n) = c + \sum_{j=1}^{J-1} k_j \cdot y_j \left\{ \overline{d_w} \left(\left\lfloor \frac{n}{w} \right\rfloor \right) \right\}, \quad n = 0, \dots, N-1, \quad (1)$$

where: N is the size of detection function, $\overline{d_w}(i)$ denotes i -th window of length w ($w \ll N$) of detection function, $y_j \{\cdot\}$ is the j -th statistics, J is the number of statistics, $y_j \{\overline{d_w}(\cdot)\} : \mathbb{R}^w \rightarrow \mathbb{R}^1$.

Fig. 1 Generalized scheme of peaks detection procedure with the marked selection stage

The generalized scheme of peaks detection procedure for audio signals is depicted in Fig. 1. The properties of the audio signal are emphasised at the first stage where the detection function is generated. Then, the resulting signal is normalised in the specific range, and peaks detection is performed. The next stage for selection of peaks from a group of candidates is executed using an adaptive thresholding function. Such thresholding is applied to ignore spurious peaks. Finally, the peaks detected in the selected timespan are grouped to improve the final detection process.

In the following sections, the problem of the adaptive thresholding configuration using a heuristic approach is discussed.

4 Detection function

To calculate the detection function an input signal is split into frames, then for each frame (n) an audio feature vector $\mathbf{x}_n^{(m)}$ is calculated, where m is the dimensionality of the feature vector. Finally, the obtained vectors from adjacent frames are compared by the similarity measure $\Phi(\cdot)$. The assumption is that the resulting representation should emphasise the places in audio signal related to the analysis task. Therefore, localisation of change points can be found by searching for dominant peaks in detection function $d(\cdot)$. Because the identification of possible peak candidates depends on the frame size, the detection is usually performed with a given margin window (w_s). The detection function is defined as:

$$d(n) = \Phi \left(\mathbf{x}_n^{(m)}, \mathbf{x}_{n+1}^{(m)} \right), \quad (2)$$

where: $\Phi(\cdot)$ is the similarity measure, $\Phi: \mathbb{R}^m \rightarrow \mathbb{R}^1$, $\mathbf{x}^{(m)}$ is the feature vector, m denotes feature space dimensionality, n is the number of frame, $n = 0, \dots, N - 2$ and N is the total number of frames.

An example of detection function with marked four change points (P_1 , P_2 , P_3 and P_4) is presented in Fig. 2. Because the detection function consists of many possible peak candidates, in signal analysis tasks, only dominant and task-related peaks have to be selected. For detection of such peaks, significant role plays its surrounding consists, in the ideal case, changes with values below the heights of analysed peaks. The situation is better when the heights difference between the actual and the highest peak in the surroundings is the biggest as can be observed for peaks denoted as P_1 , P_2 and P_4 . In contrast, the opposite case exists for peak P_3 (marked with a minus sign). Therefore, the constant value thresholding cannot be applied to select all four peaks with perfect detection accuracy. For this reason, the elimination of irrelevant peaks from the analysis task point of view has to exploit the statistical properties of detection function around the actual peaks.

Fig. 2 An example of detection function with labeled peaks

5 Peaks separability

The accuracy of peaks detection depends on the variability of $d(\cdot)$ signal. For this reason, we proposed a technique to evaluate the $d(\cdot)$ signal by calculating the peak separability score for known peaks in the signal shown in Algorithm 1. The score is computed as a ratio of the number of separable peaks to the total number of peaks taking local or global constraints into account.

The peak is separable locally when the values of the signal in regions on the right and left sides have smaller values than the considered peak. The global separation means that all peaks are locally separable, and the values in surroundings are lower than the lowest peak. According to the example in Fig. 2 the peak P_3 is not separable locally because there is a higher value between neighbouring peaks. The proposed algorithm calculates two types of scores: local (Ψ_l) and global (Ψ_g). Both scores may range from zero (the worst case) to one (the best case – when all peaks are separable). The situation when $\Psi_g = 1$ means that the detection of all peaks may be executed using a constant thresholding level. On the other side, the detection function with the high local score (Ψ_l) is more suitable to use with adaptive thresholding functions. The case for $\Psi_l = 0$ indicates that the input signal is highly noised.

6 Peaks detection

The aim of the peaks detection stage in terms of determining the boundaries of segments is not to detect all peaks, but only dominant peaks with the defined height. The selection criterion is always defined as the comparison with defined thresholding function. Because the segments have the non-zero length the height and its separability are crucial in the detection stage. In such a case, the one can select any peak-picking algorithm because the result is dependent on the task specificity and source signal defined as detection function.

Algorithm 1 Local (Ψ_l) and global (Ψ_g) score calculation for determining the separation level of peaks.

Inputs : $d(\cdot)$ – detection function; t_s – number of samples between actual peaks;
 w_s – number of samples in detection margin; M – number of actual peaks

Output : separation scores – Ψ_l, Ψ_g

```

1 for  $i \leftarrow 1$  to  $M$  do
2    $R(i) \leftarrow \{d(i \cdot t_s - w_s), \dots, d(i \cdot t_s + w_s)\}$ 
3 end
4  $s_1 \leftarrow 0$ ;  $s_2 \leftarrow 0$ 
5  $V \leftarrow \{\max[R(i)] : i = 1, \dots, M\}$ 
6  $\hat{u} \leftarrow \min(V)$ 
7  $P \leftarrow \{\arg \max[R(i)] : i = 1, \dots, M\}$ 
8 for  $i \leftarrow 1$  to  $M$  do
9    $W_L \leftarrow \{d(P_{i-1} + 1), \dots, d(P_i - 1)\}$ 
10   $W_R \leftarrow \{d(P_i + 1), \dots, d(P_{i+1} - 1)\}$ 
11  if  $\hat{u} > \max(W_L)$  and  $\hat{u} > \max(W_R)$  then
12     $s_1 \leftarrow s_1 + 1$ 
13  end
14  if  $V_i > \max(W_L)$  and  $V_i > \max(W_R)$  then
15     $s_2 \leftarrow s_2 + 1$ 
16  end
17 end
18  $\Psi_g \leftarrow s_1/M$ ;  $\Psi_l \leftarrow s_2/M$ 

```

In this work we proposed a detection procedure which uses the simplest form of neighbourhood criterion in the detection function $d(\cdot)$. For this purpose we defined a dedicated function $P_d(\cdot)$ in the following form:

$$P_d(n) = \Theta[T_h(n)] \cdot \prod_{p=0}^{2w-1} \Theta[d(n+1+\lfloor p/2 \rfloor) \cdot (-1)^p], \quad (3)$$

where: $\Theta(z) = \left\lfloor \frac{1}{2} + \frac{1}{2} \cdot \text{sgn}[d(n) - z] \right\rfloor$, $n = 1, \dots, N-2$, $P_d(n) \in \{0, 1\}$, $T_h(n)$ denotes thresholding function and $\text{sgn}(\cdot)$ is the signum function.

The dynamic range of $T_h(\cdot)$ function has to be lower than the variability of $d(\cdot)$ thus the values of the thresholding function should satisfy the inequality:

$$\min[d(\cdot)] < T_h(n) < \max[d(\cdot)]. \quad (4)$$

The function $P_d(n)$ returns 1 if the peak occurs in n -th frame (when it is a local maximum, and its value is higher than a threshold value), otherwise it returns 0.

7 Adaptive thresholding

In contrast to the constant thresholding, the configuration of adaptive thresholding is more challenging due to variability and adaptation to local changes of the analysed signal. The simplest form of such function may use local mean and median values of frames with additional scaling [2, 15, 27].

In our study the function is composed of statistical properties of analysed signal: mean, median, standard deviation and constant value as a compromise between robustness and computational cost. We have employed the following thresholding function estimated for the frames of the detection function $d(\cdot)$:

$$T_h(n) = \underbrace{\alpha \cdot \bar{d}(n)}_I + \underbrace{\beta \cdot \tilde{d}(n)}_{II} + \underbrace{\gamma \cdot \hat{d}(n)}_{III} + \underbrace{c}_{IV}, \quad (5)$$

where: $\bar{d}(n)$ is mean value, $\tilde{d}(n)$ denotes median, $\hat{d}(n)$ is standard deviation of the frame and c is a constant value, $n = 0, \dots, N - 1$, and N is the total number of frames. The α , β and γ are weighting coefficients.

The function $T_h(\cdot)$ uses the sum of four components denoted as I, II, III and IV in the later sections. The main problem with the $T_h(\cdot)$ function is connected with the selection of the weighting coefficients and constant value to select valid peaks in the detection function. This issue is addressed in the next parts of the work.

8 Experimental setup

A dedicated dataset with various type of audio segments was prepared to determine the coefficients of the thresholding function. The dataset¹ used in the experiments contains 15 recordings with nine consecutive segments each. Every item is 45 seconds long where the single segment is five seconds long, mono, with 44100Hz sampling rate. The data contains various types of speech (monologues, dialogues with male and female speakers), different music genres and typical environmental sounds including natural sounds, machines and others. Table 1 shows the characteristics of the audio data set.

We have decided to use recordings with equal length segments and eight change points at the same locations to simplify the analysis of detection function and rules for finding suboptimal coefficients for thresholding function. In contrary to such case the existence of variable segments lengths and the number of change points for every signal would significantly impede the analysis process. The sequences of segments and types of changes between them were selected randomly excluding three homogeneous sets containing speech, music and environmental sound excerpts. The types of transitions and its occurrence in the recordings is characterised and presented in Table 2.

The energy distribution in the time–frequency plane of all recordings used in the experiments is shown in Fig. 3. In most sets the all segments are easily distinguishable in the time–frequency plane.

To evaluate the detection results we calculated the precision rate (P) and recall rate (R):

$$P = \frac{S_C}{S_D}, \quad R = \frac{S_C}{S_A},$$

then we used the F_1 -score defined as harmonic mean of P and R [25]:

$$F_1\text{-score} = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \cdot S_C}{S_A + S_D},$$

where: S_C is the number of correctly detected peaks, S_A denotes the number of actual peaks and S_D is the number of detected peaks. The high P value and at the same time low R value denotes that a few peaks are detected and most of them are correct, in the opposite case

¹The dataset can be downloaded at <http://quefreny.org/data/#at>

Table 1 Audio test data characteristics (segment type: S – speech, M – music, E – environmental sound)

Set	Audio structure										Segments		
	1	2	3	4	5	6	7	8	9		S	M	E
1	S	M	M	E	S	S	M	E	M		3	4	2
2	E	E	S	E	S	M	E	M	M		2	3	4
3	M	S	S	M	S	E	E	S	E		4	2	3
4	S	E	M	S	S	M	M	S	E		4	3	2
5	M	M	E	E	S	M	M	S	E		2	4	3
6	E	S	M	M	E	E	E	M	E		1	3	5
7	E	E	E	M	E	M	M	M	S		1	4	4
8	S	E	M	S	S	E	M	M	S		4	3	2
9	S	M	M	M	E	S	M	E	M		2	5	2
10	S	M	S	S	S	M	E	E	E		4	2	3
11	S	S	S	S	E	M	E	E	E		4	1	4
12	M	M	M	E	E	E	S	S	E		2	3	4
13	S	S	S	S	S	S	S	S	S		9	0	0
14	M	M	M	M	M	M	M	M	M		0	9	0
15	E	E	E	E	E	E	E	E	E		0	0	9
Total:											42	46	47

many peaks are identified, and most of them are incorrect. The F_1 -score is equal to 1 for the perfect detection and 0 in the worst case.

9 Evaluation procedure

As an example of peaks detection application, we have selected a metric-based segmentation. In short, the whole process looks as follows: the similarity of two audio frames is computed and used to form the detection function, then a process of peaks picking

Table 2 Distribution of transitions in the audio database

Transition	Set															Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
S → M	2	1	1	1	1	1			2	2						11
M → S			2	2	1		1	2		1						9
S → E		1	2	2	1			2			1	1				10
E → S	1	2	1		1	1			1			1				8
M → E	2	1			1	2	1		2	1	1	1				12
E → M	1	1		1		1	2	2	1		1					10
S → S	1		1	1				1		2	3	1	8			18
M → M	1	1		1	2	1	2	1	2			2		8		21
E → E		1	1		1	2	2			2	2	2			8	21

Fig. 3 Spectrograms of all sets used in the experiments

with optional thresholding is executed to detect the possible boundaries of the segments in audio data. In all experiments, we used the normalized detection function with margin $w_s = 1000$ ms.

9.1 Detection function

We have defined a simulation domain consists of 200 features and 16 similarity measures, and then we run the Monte Carlo simulation. Using random source with unbiased uniform distribution, for each iteration, we have sampled $m = 2, \dots, 8$ dimensional feature vector and similarity measure, then calculate detection function $d(\cdot)$ and normalize it to range $[0, 1]$. For obtained function, we computed the Ψ_g score at the positions of actual change points. The number of iterations for each recording was equal to 10^4 , then from a final set, a 1000 items (where $\Psi_g > 0$) were selected, and the instances of each similarity measure were counted. The three dominant similarity measures with score $\Psi_g \geq 0.6$ are depicted in Fig. 4.

In the result of this experiment, the following three similarity measures were selected for subsequent analysis: mean absolute error (M_{MAE}), mean absolute percentage deviation (M_{MAPD}) and Kullback–Leibler divergence function (M_{KL}).

An influence of frame size on the detection function was determined in the next trial using the chosen similarity measures. Figure 5 illustrates that the highest Ψ_l score was obtained for 100 ms frame length and M_{MAE} measure. We have decided to use violin plots [13] to show the dynamical range and local density estimates of the data.

Fig. 4 The number of instances in function of ψ_g threshold for dominant measures obtained in Monte Carlo simulation

9.2 Feature space selection

In the next step, we have conducted experiments to select the audio feature type for frames comparison. For this purpose, we decided to employ a popular scheme for change points detection in audio data (ΔBIC) which utilises the Bayesian Information Criterion (BIC). In this method, the maximum likelihood principle is applied to the two following models M_1 and M_2 [8]:

$$M_1 : \mathbf{x}_1^{(m)}, \dots, \mathbf{x}_N^{(m)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_N),$$

$$M_2 : \begin{cases} \mathbf{x}_1^{(m)}, \dots, \mathbf{x}_i^{(m)} \sim \mathcal{N}(\boldsymbol{\mu}_{(i,L)}, \boldsymbol{\Sigma}_{(i,L)}) \\ \mathbf{x}_{i+1}^{(m)}, \dots, \mathbf{x}_N^{(m)} \sim \mathcal{N}(\boldsymbol{\mu}_{(i,R)}, \boldsymbol{\Sigma}_{(i,R)}) \end{cases}.$$

The ΔBIC can be treated as a likelihood criterion penalised by the complexity of the model:

$$\Delta BIC(i) = N_{(i,L)} \log |\boldsymbol{\Sigma}_{(i,L)}| - N_{(i,R)} \log |\boldsymbol{\Sigma}_{(i,R)}| - N \log |\boldsymbol{\Sigma}_N| - \lambda P(N, m), \quad (6)$$

where: $|\boldsymbol{\Sigma}|$ is determinant of covariance matrix, λ is a penalty weight, $P(N, m)$ denotes penalty function [6]. If the difference (ΔBIC) between BIC values for models M_1 and

Fig. 5 The ψ_l score for detection function computed for several frame sizes with selected similarity measures

Table 3 The F_1 -score characteristics for Δ BIC audio segmentation scheme

Feature	F_1 -score		
	Min	Max	Mean
GFCC	0	0.8571	0.514
LFCC	0	0.6667	0.3283
LPC	0.75	1	0.8995
MFCC	0	0.7692	0.3869

M_2 is positive, then the model M_2 is preferred. That situation constitutes a change point candidate in data. A simple analysis technique with sliding variable-size window is applied to detect all possible change points. Most of the implementations uses feature space with mel-frequency cepstral coefficients and log-energy (standard set for analysis, recognition and processing of speech signals). Therefore, we have compared the effectiveness of several feature spaces and their influence on the final accuracy of change point detection.

We used the penalty coefficient $\lambda = 1$ as was suggested in the original work, but it results in many false change points. In this case, the clustering stage is required to compensate for the high sensitivity of local changes in signal. Thus, the value of λ was selected experimentally to minimise the number of false results without using the clustering stage. We have chosen λ value keeping in mind the number of detecting peaks close to the number of actual peaks. For the comparison of discriminative properties of audio features, the five popular feature sets [8, 12, 16, 23] has been tested using Δ BIC scheme. The evaluation was performed for the frame size equal to 10 ms, the dimensionality $m = 24$ and the penalty coefficient $\lambda = 4$. The F_1 -scores are shown in Table 3. The best results were achieved for the linear prediction coefficients (LPC), then for gammatone frequency cepstral coefficients (GFCC) and mel-frequency cepstral coefficients (MFCC). However, the MFCC, GFCC and LFCC implementations have the higher dimensionality of feature space than $m = 24$. In this case, we employed the first 24 coefficients only.

As a result, we have chosen the LPC feature to generate a detection function with the M_{MAE} similarity measure. Next, for the whole dataset, we calculated the Ψ_l score to find the separability of change points. Local separability levels are depicted in Fig. 6 for dimensionality $m = 2, \dots, 24$ and frame size equal to 100 ms.

From Fig. 6, it is apparent that the highest Ψ_l scores with lowest variability of peaks separation occurs for $m = 2$ and $m = 3$. For the sake of this fact we selected $m = 2$ in the final form of our detection function.

9.3 Computational cost analysis

The peaks selection process can be applied to various audio analysis tasks. Thus the following considerations of computational cost are related to the specifics of the audio segmentation example. A computational assessment of used algorithms may be divided into two parts. In the first part, an adjustment of the parameters of analysed data, including the type of feature space, similarity measure and task specificity is carried out. Additionally, in case of used genetic algorithm, the number of generations and the size of the population are constant, the optimisation is realised in roughly constant time. At this stage, an estimation of adaptive thresholding function and its components is performed only once.

Fig. 6 The Ψ_l score for M_{MAE} measure with LPC feature space (dimensionality from 2 to 24)

In the second part, the proper process of peaks detection and selection is carried out. The source audio signal of length equal to Z seconds is split into equal-length frames with size R , resulting in $N = Z/R$ frames. Next, for each frame, a vector of linear prediction coefficients (LPC) with order $L = 24$ is computed and from this vector, $\bar{L}$ coefficients are selected to form a final feature vector where $\bar{L} \leq L$. In our implementation, the LPC coefficients were calculated using Levinson–Durbin recursion, where the computational complexity is proportional to L^2 [3]. Then, the resulting feature vector from every frame is compared with the adjacent frame using mean absolute error (MAE) measure giving $N - 1$ values of detection function. Such procedure requires $\bar{L} \cdot (N - 1)$ additions and subtractions and $N - 1$ division by the $\bar{L}$ value. In the next stage, peaks detection is performed, and its complexity is dependent on the type of detection algorithm. Using a simple technique described by (3) which is an equivalent of a set of comparisons, the current value has to compare with its adjacent values of detection function. The number of comparisons, in this case, is equal to $2w \cdot (N - 1)$, where w is the size of the analysed window. Obtained a group of peaks candidates a trajectory of adaptive thresholding is computing with use an average value, median, standard deviation and a constant value in $\lfloor (N + 1)/H \rfloor$ frames, where H is the size of the window used for calculation of the single value of thresholding function. The exact number of arithmetic operations of this phase is dependent on the type of statistics implementation and computational complexity of the sorting method used in the median calculation.

In the case of the ΔBIC algorithm employed in the comparison of the segmentation process, using a direct implementation presented in [8], the computational complexity is proportional to Z^3 [9].

10 Experimental results

The proper selection of the thresholding signal is not trivial, and it needs to tweak several parameters at once to get the acceptable results. That is why we performed an analysis of an adaptive thresholding function using a genetic algorithm (GA). The algorithm is based on the basic scheme and uses non-overlapping populations. At each generation, a new set

Fig. 7 The F_1 -score for all combinations of components in adaptive thresholding function

is generated by selecting from the past population and creating the new offspring. The simulation was performed for the whole dataset and all combinations of components (I, II, III and IV). The optimisation stage was run with the following configuration: population size was equal to 30, a 1000 generations were produced with mutation probability equal to 0.01, crossover probability was equal to 0.6 and F_1 -score was used as an objective function. The F_1 -scores achieved in simulation are depicted in Fig. 7.

In this case, the set of four components gave the highest mean F_1 -score. For obtained distributions of results for each configuration, we have selected final coefficients by computing a set of statistics from each distribution.

Table 4 Segmentation results for all possible configurations

Configuration	F_1 -score			Function
	Min	Max	Mean	
I	0.4615	0.8571	0.6816	max
II	0.4348	0.7	0.5717	median
III	0.2963	0.6667	0.4046	max
IV	0.4	0.7143	0.5341	mean
I + II	0.4615	1	0.6867	median
I + III	0.4615	0.875	0.6678	median
I + IV	0.5333	1	0.6875	mean
II + III	0.4615	0.875	0.6755	mean
II + IV	0.5	0.9333	0.7083	median
III + IV	0.4286	0.7143	0.5967	mean
I + II + III	0.4706	0.875	0.6828	mean
I + II + IV	0.5455	0.8889	0.7077	median
I + III + IV	0.5333	0.9412	0.7020	median
II + III + IV	0.5	1	0.7229	mean
I + II + III + IV	0.4615	1	0.6966	mean

Fig. 8 Comparison of segmentation accuracy of ΔBIC algorithm and proposed technique with heuristic adaptation of thresholding function (GA)

A change point detection procedure was executed with adaptive thresholding for every case to verify the weighting coefficients. The results are shown in Table 4, where the highlighted result was obtained for the following configuration: $\beta = 1.856708$, $\gamma = 1.590472$, $c = 0.086527$.

For comparison purposes using the same feature space, a segment's boundaries detection test was performed and compared with the technique called ΔBIC [5]. The results of this process are depicted in Fig. 8. Except for the eight set in the rest cases the proposed technique improved the detection accuracy with incomparably less computational cost.

After segment boundaries detection with the proposed approach and the ΔBIC method, an analysis of the transitions between segments containing various acoustic signals without discriminative properties for selected feature space was performed. The data set used in experiments contains 135 different parts and 120 change points between them (see Table 1). In the proposed technique for the whole data set 35 transition points were not detected

Fig. 9 Illustration of the detection process for set 13

Fig. 10 Illustration of the detection process for set 14

which represents 29.2% of the total, while 51 points were not identified in case of ΔBIC which is 42.5% of all change points. The least detected segment transitions were observed for the same type signals (S-S, M-M, E-E). In this case, 29 and 14 points for the first and second case respectively were not identified. For mixed segments, the number of undetected transitions between parts was similar in both methods and varied from 7 to 8 points.

11 Threshold analysis

For the purpose of comparing the obtained results, we have selected three homogeneous sets (speech, music, and environmental sounds). In the first case, the set contains several speech signals and the obtained detection results for optimised and final threshold are different. In

Fig. 11 Illustration of the detection process for set 15

the Fig. 9 the results of peaks detection in the case using both thresholds are presented. In the case of suboptimal thresholding, the change between the third and fourth segment was missing due to the absence of a peak in the detection function for the range. Also, between 5th and 6th segments the change was not detected despite the occurrence of the distinct peak at the boundaries. Additionally, two peaks were occurred between two first segments and at the end of the last regions which were joined at the clustering stage, resulting in F_1 -score equal to 0.8. In the result of the comparison of the final and suboptimal threshold functions, two false peaks appeared in fourth and seventh segments diminishing the F_1 -score to the 0.7368.

The structure of peaks remaining the same after applying the final thresholding in case of the set containing music segments depicted in Fig. 10. The mean-squared error for this set is equal to 0.006125, and it has the lowest value in the whole dataset. Also, all boundaries of the segments have been determined. The deterioration of detection results while replacing the optimised threshold with global final threshold was observed for a sequence of segments contained various environmental sounds as shown in Fig. 11. The detection function did not expose a peak between region two, three and four (transition between birds, kids vocalizations and stadium ambience). After applying the final threshold the boundaries of the 7th segment were not detected. The set contained environmental sounds the detection function does not reflect the segment's boundaries accurately. For this reason, the visible changes were not in transitions between the 2nd, 3rd and 4th segment. After using final thresholding function the boundaries between 6th and 7th and 7th and 8th segment were not found which have an effect on the F_1 -score by decreasing from 0.8 to 0.67. Figure 12 shows the mean-squared error values computed between thresholding functions for sub-optimal and final cases for all sets. Moreover, in Fig. 13 a comparison of thresholding functions of sub-optimal and final representations for the highest mean-squared error equal to 0.00813 (top panel) for set 1 and the lowest equal to 0.00051 (bottom panel) for set 8 is presented. Use the target coefficients of the end representation of thresholding function produces the deterioration of the detection accuracy except for set 14. Such a situation is somewhat expected due to the separate estimation of the sub-optimal threshold for individual sets. In Table 5 a comparison of adaptive thresholds optimized for given set and final estimated thresholding function is shown. Only in case of set 14, there is no worsening of the detection accuracy.

Fig. 12 Mean squared error between optimized and final threshold functions

Fig. 13 Final thresholds for the highest value of the mean-squared error: set 1 **(a)** and the lowest value set 8 **(b)**

In the remaining situations, the lowest difference between F_1 -scores were observed for the set 12, whereas the highest accuracy deterioration occurred for the set 8. An interesting case is connected with the set 8 where despite the highest resemblance between thresholding functions the accuracy is most decreased.

Table 5 Accuracy detection for the optimized and the final threshold

Set	Optimized threshold			Final threshold		
	Precision	Recall	F ₁ –score	Precision	Recall	F ₁ –score
1	1	0.875	0.9333	0.7	0.875	0.7778
2	0.8571	0.75	0.8	0.7143	0.625	0.6667
3	1	0.5	0.6667	0.6667	0.5	0.5714
4	1	1	1	0.8	1	0.8889
5	0.8571	0.75	0.8	0.6	0.75	0.6667
6	1	1	1	1	0.75	0.8571
7	0.6154	1	0.7619	0.6667	0.5	0.5714
8	1	0.75	0.8571	0.75	0.375	0.5
9	1	0.875	0.9333	0.8333	0.625	0.7143
10	0.7778	0.875	0.8235	0.625	0.625	0.625
11	1	0.652	0.7692	0.7143	0.625	0.6667
12	0.8889	1	0.9412	1	0.875	0.9333
13	0.8571	0.75	0.8	0.6364	0.875	0.7368
14	1	1	1	1	1	1
15	0.8571	0.75	0.8	1	0.5	0.6667

12 Concluding remarks

We have proposed a technique to peaks separability assessment for manually labelled data. Then, we showed how the resulting scores might be used to estimate the adaptive thresholding in peaks-picking procedure. In the result of audio feature space comparison, we have found that LPC coefficients have better discriminative properties for presented approach than other popular features including MFCC. The obtained results show that the proposed method may be employed to determine change points with low computational effort. It is important to stress that every change point detection method requires the configuration phase where the parameters are tuned to the set of source signals, especially in real acoustical conditions. After the heuristic adaptation of thresholding function, the best segmentation results occurred for two local statistics of the signal (median, standard deviation) and the constant value. We have successfully implemented the presented technique along with onset maps in our auditory scene analysis system as a fast equivalent to segmentation mechanisms which are based on the BIC criterion.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Aboy M, McNamers J, Thong T, Tsunami D, Ellenby MS, Goldstein B (2005) An automatic beat detection algorithm for pressure signals. *IEEE Trans Biomed Eng* 52(10):1662–1670
2. Bello JP, Daudet L, Abdallah S, Duxbury C, Davies M, Sandler MB (2005) A tutorial on onset detection in music signals. *IEEE Trans Acous Speech Signal Process* 13(5):1035–1047
3. Benesty J, Sondhi MM, Huang Y (2008) Springer handbook of speech processing. Springer, Berlin. <https://doi.org/10.1007/978-3-540-49127-9>
4. Bock S, Schlüter J, Widmer G (2013) Enhanced peak picking for onset detection with recurrent neural networks. In: 6th International workshop on machine learning and music – MML’2013, Prague, pp 1–4
5. Cettolo M, Vescovi M (2003) Efficient audio segmentation algorithms based on the BIC. In: IEEE International conference on acoustics, speech, and signal processing – ICASSP’2003, pp 537–540
6. Cettolo M, Vescovi M, Rizzi R (2005) Evaluation of BIC-based algorithms for audio segmentation. *Comput Speech Lang* 19(2):147–170
7. Chan CF, Yu EWM (2010) An abnormal sound detection and classification system for surveillance applications. In: 18th European signal processing conference – EUSIPCO’2010. IEEE, Aalborg, pp 1851–1855
8. Chen SS, Gopalakrishnan P (1998) Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In: DARPA Broadcast news transcription and understanding workshop, Lansdowne, pp 127–132
9. Cheng SS, Wang HM (2003) A sequential metric-based audio segmentation method via the Bayesian information criterion. In: European conference on speech communication and technology – EUROSPREECH’2003, Geneva, pp 945–948
10. Dov D, Talmon R, Cohen I (2017) Multimodal kernel method for activity detection of sound sources. *IEEE/ACM Trans Audio Speech Lang Process* 25(6):1322–1334
11. Fodor B, Fingscheidt T (2012) Reference-free SNR measurement for narrowband and wideband speech signals in car noise. In: 10. ITG Symposium speech communication. VDE, Braunschweig, pp 1–4
12. Ganchev T (2011) Contemporary methods for speech parameterization, 1st edn. Springer briefs in electrical and computer engineering. Springer, New York
13. Hintze JL, Nelson RD (1998) Violin plots: a box plot-density trace synergism. *Am Stat* 52(2):181–184
14. Kauppinen I (2002) Methods for detecting impulsive noise in speech and audio signals. In: 14th International conference on digital signal processing - DSP’2002, vol 2, Santorini, pp 967–970
15. Li X, Horaud R, Girin L, Gannot S (2016) Voice activity detection based on statistical likelihood ratio with adaptive thresholding. In: IEEE International workshop on acoustic signal enhancement – IWAENC’2016. IEEE, Xi’an, pp 1–5
16. Markel JD Jr, AHG (1976) Linear prediction of speech, communication and cybernetics, vol 12. Springer, Berlin
17. Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 11(285–296):23–27
18. Potamitis I, Ganchev T (2008) Generalized recognition of sound events: approaches and applications. In: Tsihrintzis GA, Jain LC (eds) Multimedia services in intelligent environments, vol 120. Springer, Berlin, pp 41–79
19. Rosao C, Ribeiro R, de Matos DM (2012) Influence of peak selection methods on onset detection. In: Proceedings of the 13th international society for music information retrieval conference - ISMIR’2012, Porto, pp 517–522
20. Rosin PL (1997) Edges: saliency measures and automatic thresholding. *Mach Vis Appl* 9:139–159
21. Sahoo PK, Soltani S, Wong AKC (1988) A survey of thresholding techniques. *Comput Vis Graph Image Process* 41(2):233–260
22. Scholkmann F, Boss J, Wolf M (2012) An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals. *Algorithms* 5(4):588–603
23. Shao Y, Wang D (2008) Robust speaker identification using auditory features and computational auditory scene analysis. In: IEEE International conference on acoustics, speech and signal processing – ICASSP 2008, Las Vegas, pp 1589–1592
24. Siedenburg K (2012) Persistent empirical wiener estimation with adaptive threshold selection for audio denoising. In: The 9th sound and music computing conference – SMC’2012, Copenhagen, pp 426–433
25. Theodoridis S, Koutroumbas K (2009) Pattern recognition, 4th edn. Academic Press

26. Weszka JS (1978) A survey of threshold selection techniques. *Comput Graph Image Process* 7(2):259–265
27. Yan F, Zhang H, Kube CR (2005) A multistage adaptive thresholding method. *Pattern Recogn Lett* 26(8):1183–1191

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Tomasz Maka received his PhD degree in computer science from Szczecin University of Technology in 2005. He is currently working as an assistant professor in Faculty of Computer Science and Information Technologies, West Pomeranian University of Technology, Szczecin. His scientific interests include audio analysis, machine hearing and audio feature engineering.