Computing and Informatics, Vol. 36, 2017, 857–886, doi: 10.4149/cai_2017_4_857

CLASSIFICATION OF BROADCAST NEWS AUDIO DATA EMPLOYING BINARY DECISION ARCHITECTURE

Jozef Vavrek, Peter Feciľak, Jozef Juhár, Anton Čižmár

Technical University of Košice
Faculty of Electrical Engineering and Informatics
Letná 9
042 00 Košice, Slovakia
e-mail: {jozef.vavrek, peter.fecilak, jozef.juhar, anton.cizmar}@tuke.sk

Abstract. A novel binary decision architecture (BDA) for broadcast news audio classification task is presented in this paper. The idea of developing such architecture came from the fact that the appropriate combination of multiple binary classifiers for two-class discrimination problem can reduce a miss-classification error without rapid increase in computational complexity. The core element of classification architecture is represented by a binary decision (BD) algorithm that performs discrimination between each pair of acoustic classes, utilizing two types of decision functions. The first one is represented by a simple rule-based approach in which the final decision is made according to the value of selected discrimination parameter. The main advantage of this solution is relatively low processing time needed for classification of all acoustic classes. The cost for that is low classification accuracy. The second one employs support vector machine (SVM) classifier. In this case, the overall classification accuracy is conditioned by finding the optimal parameters for decision function resulting in higher computational complexity and better classification performance. The final form of proposed BDA is created by combining four BD discriminators supplemented by decision table. The effectiveness of proposed BDA, utilizing rule-based approach and the SVM classifier, is compared with two most popular strategies for multiclass classification, namely the binary decision trees (BDT) and the One-Against-One SVM (OAOSVM). Experimental results show that the proposed classification architecture can decrease the overall classification error in comparison with the BDT architecture. On the contrary, an optimization technique for selecting the optimal set of training data is needed in order to overcome the OAOSVM.

Keywords: Support vector machine, audio classification, broadcast news data, binary decision trees, binary decision architecture

Mathematics Subject Classification 2010: 68T10

1 INTRODUCTION

The current technologies for automatic processing of massive audio-visual data have recorded significant development in past few years. This fact can be explained by a continuous growth of multimedia data available on the internet and various audio-visual databases. Therefore, the efficient management of audio and video content is becoming inevitable. There is a need to implement various content-based techniques in order to process these data automatically. Content-based analysis of audio data is the key component for applications like automatic speech recognition, classification and retrieval. It is a contribution for the solution process in discrimination between various acoustic classes in audio stream by searching through frequency and time domain of the signal amplitude and finding parameters that capture all the spectral and temporal variations. Such processing of audio content improves the overall discrimination power between all acoustic classes and even more optimize input data for the process of classification by using a predefined type of classifier. These techniques are widely applied in systems that process audio data with rich audio content.

The motivation to our research is the content-based indexing and retrieval with the possible use in automatic systems for retrieving queries in broadcast news (BN) databases. We have focused on filtering of non-speech segments and training acoustic models for pure speech and non-pure speech in order to improve speech recognition accuracy in the first step. The retrieving accuracy of spoken queries and utterances is thus conditioned by a transcription produced at the output of automatic speech recognition (ASR) system. The retrieving itself is then conducted by comparing decoded text strings. The paper is therefore focused on content-based classification of BN data utilizing a robust classification architecture along with sufficient feature extraction methods.

BN audio data contains alternating sections of different types, such as speech and music. Xie et al. defined [1] six general acoustic classes of BN audio stream, namely pure speech, speech with environment sound, environment sound, speech with music, music and silence. Each individual class is characterized by unique acoustic properties and random occurrence in audio stream. The most frequently occurring and a dominant audio class is *pure speech*. Reports in studio and field speech in a quiet environment belong to this class. The second biggest class is related to the field speech in noisy environment, defined as *speech with environment sound*, also known as background speech. Many kinds of *environment sounds* such as sound from machines, birds, water, wind, and crowds are simultaneously mixed

with speech. Music at the beginning and at the end of news with anchor speech and commercials with speech are part of the *speech with music* audio class. This category consists of *music sound* like jingles and music during commercials. BN audio stream also contains silent intervals between different speakers and jingles, defined as *silence*.

Particular group of audio events with relatively rare occurrence in BN stream is *telephone speech*. The telephone speech signals are characterized by frequency range 300–3 400 Hz and variable audio content, where pure speech is also mixed with music or environment sound. A simple way for detecting telephone speech involves computing the energies in different frequency bands. The classification of telephone data is beyond the scope of this paper since the presence of telephone speech in TV broadcast news is nearly insignificant (less than 1 %). Some proposals for automatic detection, classification and transcription of telephone speech are examined in [2] and [3].

Different classification methods have been evaluated and compared by employing various features and classifiers. Most of them are aimed on the design of such complex systems that are capable to classify the audio stream into speech and non-speech segments by using novel features and rule-based classification approach. Other works point to the importance of using the rule-based approach in other tasks, like triphone mapping for training the acoustic models [4] and text-to-speech synthesis [5]. There exist also some evaluation projects that have been established with intention to emphasize the difficulty of segmentation task for BN data and examine various feature extraction techniques along with robust machine learning classifiers [6, 7].

In some works, it has been observed that the choice of features seemed to be more important than the choice of classifiers. While the applications can be very different, many studies use similar sets of acoustic features, such as short time energy (STE), zero-crossing rate (ZCR), cepstral coefficients [8], spectral descriptors [1], alongside some novel and robust parameters [9, 10]. Typically some long-term statistics, such as the mean or the variance, and not the features themselves, are used for the discrimination [11].

Song et al. [12] performed deep analysis of audio features in news video, including short time energy, high zero-crossing rate, bandwidth, low short-time energy ratio and noise rate. The aim of their research was to built rule-based classification system for discriminating four general classes, namely music, pure speech, silence and non-pure speech. Experimental evaluation has shown a high classification efficiency, over 93 %, implying that the selected features were rational and effective for addressed task. Dogan et al. [13] developed a general sound classification and segmentation system employing binary decision trees discrimination strategy and set of MPEG-7 low level audio descriptors, each extracted within one second sliding window. They examined classification ability of SVM and HMM classifiers for six predefined classes: silence, pure speech, music, environment sound, speech with music and speech with environment sound. They found out that the combination of several descriptors, specifically audio spectrum centroid (ASC), audio spectrum spread (ASS) and audio spectrum flatness (ASF), can capture the spectral characteristics of mixed type

audio data considerably better then one type of parameter, namely audio spectrum projection (ASP). SVM-based and HMM-based classifiers using ASS+ASC+ASF feature set, extracted from TRECVID 2003 ABC World News Tonight and CNN headline news, yielded approximate accuracy rates to each other, concretely 95.5 %. Castán et al. [14] examined a novel audio segmentation-by-classification approach based on factor analysis for classifying TV news in Catalan language into five different classes: speech, music, speech with music, speech with noise, and others. Proposed technique was compared to a hierarchical system based on GMM modeling utilizing specific acoustic features. The experimental results showed a significant error reduction (29.2 %) using segment-based factor analysis approach in comparison with a hierarchical system based on GMM/HMM modeling utilizing specific acoustic features on each level. Zhang et al. [15] investigated feature integration using universal background gaussian mixture model (UBM) into high-accuracy audio classification algorithm based on SVM-UBM. Experiments performed on radio news showed that UBM-based feature integration can effectively capture the characteristics of an audio stream and have a better classification performance than other segment-based features (MFCC for example).

Other works point to the importance of using appropriate classification architecture rather then parameterization techniques. Theodorou et al. [16] proposed a scheme for automatic categorization of radio broadcast data utilizing a common set of pre-processing and feature extraction modules, involving widely used acoustic parameters. Classification module implemented six different machine learning algorithms, specifically multilayer perceptron neural network (MLP), naive Bayes classifier (NB), SVM, k-nearest neighbour (kNN), C4.5 decision tree learner [17] and boosting algorithm combined with decision trees. The experiments conducted on the Voice of America (VOA) database for the Greek language showed that the best performance, approximately 92 % of accuracy, was achieved by the classification scheme using the boosting technique combined with decision trees. They adopted the manual annotations of the VOA recordings as a ground truth for the sound identity, considering silence, music, pure speech, speech with music and noise.

Chen et al. [18] compared results of four most popular learning classifiers, namely SVM, kNN, Neural Network (NN), and NB in term of classification accuracy. The evaluations showed that the problem of non-linear distribution of classes and small number of training samples can be solved in more appropriate way by SVM classifier.

Other authors also reported superior position of SVM in many classification tasks [19, 20, 21], especially in case of classifying audio stream with various acoustic events [22]. Therefore, we decided to propose a robust classification architecture utilizing SVM-based binary decision architecture and sufficient features with intention to minimize miss-classification error and increase the overall classification accuracy for BN audio data. The proposal of presented classification architecture is based on our previous experience with SVM classifier and various feature extraction techniques we employed in order to classify BN audio data [23, 24]. Considering our earlier research activities, the aim was directed towards proposing the robust binary decision architecture for multiclass classification task, employing the SVM.

Section 2 provides a brief overview about basic components of proposed classification architecture. In Section 3, we analyze used feature extraction techniques. Section 4 gives description about binary decision algorithm, its possible implementation in classification task, and main discrimination principles applied in BDA. Section 5 discusses experimental setup and finally Section 6 gives our conclusions and shows future directions.

2 THE ARCHITECTURE OF PROPOSED CLASSIFICATION SYSTEM

The proposed classification architecture fully exploits basic principles of subsequent binary decision strategy (Figure 1). The first step in processing input audio stream comprises the *segmentation phase*. Segmentation of audio data is in general the task of dividing a continuous audio stream into short audio portions with an equal length, also known as segments, by using rectangular window. Each segment is further divided into the overlapped frames, using Hamming window, in order to avoid spectral distortions. Generated audio frames are usually pre-emphasized by a FIR filter in order to emphasize higher frequencies that are attenuated by the human vocal tract. This procedure is called a *preprocessing*.

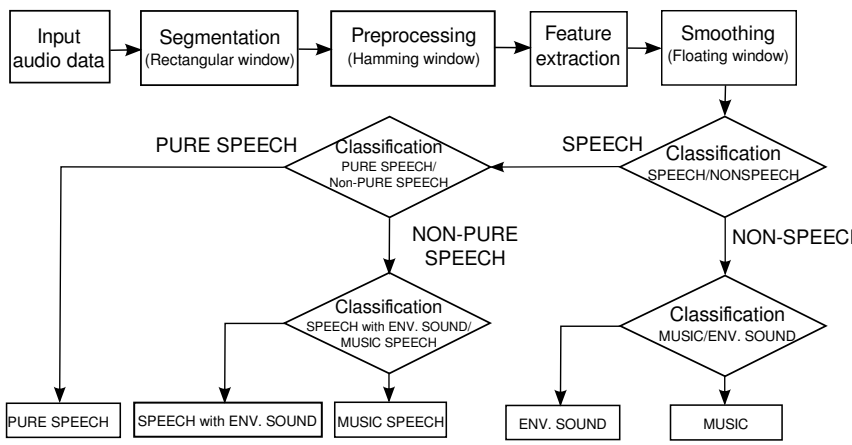


Figure 1. Proposed classification architecture

All the features are calculated within each individual frame in time, frequency or cepstral domain. The output of the *feature extraction phase* is represented by a feature vector matrix, where each column refers to the corresponding coefficient and the number of rows corresponds with the number of frames per file. The variance of feature values is then calculated within each individual segment. Such long-term statistics reduce computational consumptions and the influence of the signal's variability. Such representation of audio signal defines the format of input data for a classifier.

The process of *smoothing* can help to alleviate the influence of the abrupt changes between several adjacent coefficients within the feature vector that represents only one audio class and, as a consequence of that, reduces the miss-classification error. It is a simple technique based on the averaging of values for a particular feature set using floating window.

The overall classification logic of the architecture is based on the assumption that general sounds like speech and non-speech sound differ a lot from each other in time and frequency domain. Therefore, these easy to separate and the most general classes are classified on the first level of topology. The other classes, namely music/environment sound and pure speech/background speech, are classified in the next step, processing the audio data from previous level. The last level performs classification of the two most difficult to discriminate classes: speech with environment sound/speech with music. Each block of classification is represented by one binary discriminator that performs discrimination of input feature vectors using corresponding decision function. Decision function is defined as a set of rules represented by statistical dependencies gained in the process of training or empirical thresholds of observed discrimination parameters. The output value of decision function assigns the final class label for the actual vector. A discrimination parameter can be defined as a feature vector, which components are extracted in the process of feature extraction. It captures statistical or spectral characteristic of an input signal and helps to discriminate between different properties of the input signal, usually by setting a threshold.

3 FEATURE EXTRACTION

Feature analysis primarily investigates the behaviour of audio signals in time and frequency domain utilizing specific parameters. These parameters, also called *features* or *descriptors*, capture temporal and spectral characteristics of the audio signal and help to discriminate acoustically different audio types. The selection of feature extraction method depends on the particular classification task we want to solve. Therefore, the following subsections describe the audio descriptors, we decided to use in our classification architecture, due to their strong discrimination ability between speech and non-speech sounds.

3.1 Mel-Frequency Cepstral Coefficients

Parameterization of acoustic signal by Mel-Frequency Cepstral Coefficients (MFCC) is considered as the most effective feature extraction method for speech and other types of audio signals [25]. Many authors implement this parameterization technique in speech recognition and speech/non-speech discrimination task [26]. MFCCs are derived from the human perceptual system. The calculation procedure consists of filtering acoustic signal by non-linear mel-scale triangular filters and computing the Discrete Cosine Transform (DCT) [27] (Chapter 6: Speech Signal Representations).

Usually, 18–24 mel-filters and only the first 12 cepstral coefficients (excluding the 0^{th} energy coefficient) are used.

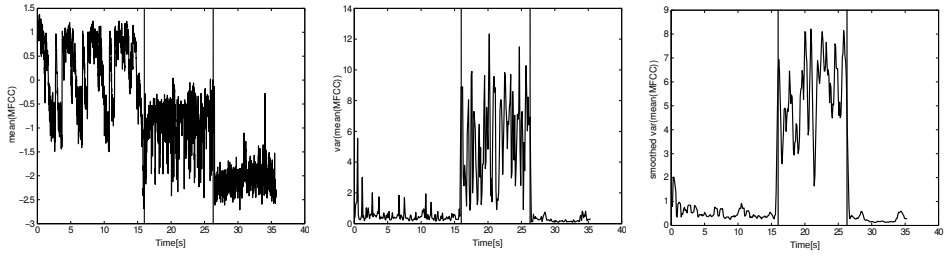


Figure 2. Mel-Frequency Cepstral Coefficients

Figure 2 illustrates the mean of the first 13 MFCCs computed for audio signal with total duration 35.47 s containing music (0–16 s), pure speech (16–26.4 s) and environment sounds (26.4–35.47 s). In the first case, curve on the left, the coefficients were extracted using Hamming window with standard length 25 ms and 10 ms overlapping. The second curve, marked as var(MFCC), represents the variance of MFCCs measured within 200 ms segments with 100 ms overlapping. Smoothed curve of var(MFCC) is depicted on the right. Vertical lines capture the change from one acoustic event to another. The variance of MFCCs is considerably higher for speech signal than for music and environment sounds. It follows relatively high ability of MFCCs to capture all the changes in speech signals. There is also ability of capturing periodicity of music signals, what is illustrated by regular curves in the first characteristic.

3.2 Variance Mean of Filter Bank Energy

Variance Mean of Filter Bank Energy (VMFBE) feature [9] calculates an energy variation in a narrow frequency sub-band of signal's spectrum. It exploits the fact that the energy varies more rapidly and to a greater extent for speech than for non-speech signals. Therefore, an energy variance in such a sub-band is greater for speech than for music or environment sound. Calculation of VMFBE coefficients is similar to the procedure used in case of MFCCs. The difference lies in computation of the energy in each sub-band of triangular mel-filters and not the power spectrum along with DCT as in case of MFCCs. Thus, the number of VMFBE coefficients equals to the number of sub-bands and the final form represents only the single mean value.

The characteristics of VMFBE are depicted in Figure 3. There is a considerable increase of VMFBE values for environment sound on the frame level (curve on left) in comparison with MFCCs (Figure 2). This fact can be explained by the higher variability of the sub-band energies for environment sounds than for speech

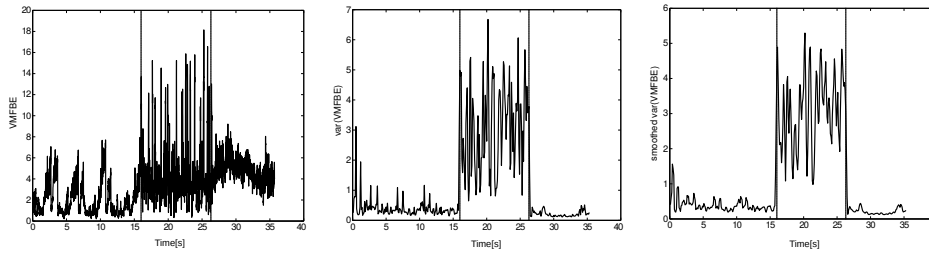


Figure 3. Variance Mean of Filter Bank Energy

or music. On the contrary, more steeper changes of VMFBE parameter are specific for speech signals. The variance of VMFBE (var(VMFBE)) is much more significant for speech than for non-speech signals. That was one of the reasons why the authors in [9] decided to use this parameter for speech/non-speech segmentation.

3.3 Variance of Acceleration MFCCs

Mel-Frequency Cepstral Coefficients do not capture the dynamic behaviour of acoustic signal in longer time span. For that reason, well known Delta MFCCs and Acceleration MFCCs are being used to capture dynamics of acoustic signal in long time span. Based on our original idea, we decided to use a novel parameter for assessing the temporal dynamics of audio signals, namely the Variance of Acceleration MFCCs (VAMFCC). It can be seen, from Figure 4, that the variability of VAMFCCs is much more significant for speech signals than for non-speech signals. Moreover, curves for music signals are much more smoother in comparison with MFC and VMFBE parameters. It gives the assumptions for the efficient use as a speech/non-speech discriminator or silence detector.

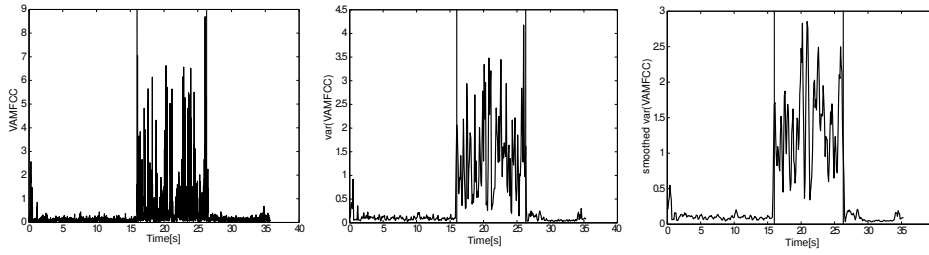


Figure 4. Variance of Acceleration MFCCs

3.4 Band Periodicity

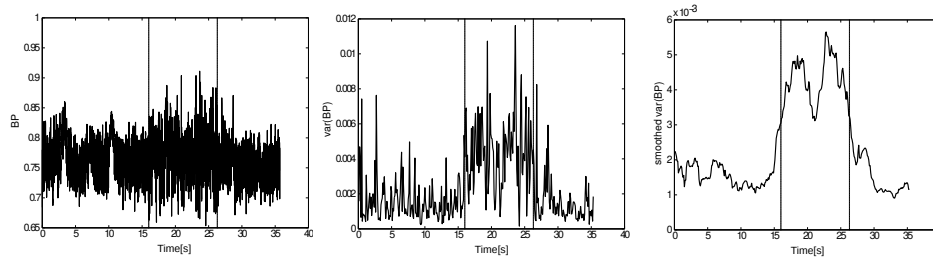


Figure 5. Band Periodicity

Band Periodicity (BP) [28] is defined as the periodicity of audio signal in particular frequency sub-band and can be represented by the maximum local peak of the normalized correlation function computed within adjacent frames in each sub-band. The bandwidth selection of each sub-band is conditioned by the sampling frequency and the frequency range of the audio signal we want to examine. The most energy of frequency components that represent speech signal is concentrated at lower frequency bands, while some frequency components of music signals are much more significant at higher sub-band frequencies. Authors in [28] employed this parameter in music/environment sound discrimination task. Figure 5 shows an example of band periodicity for the frequency band 500–1 000 Hz using 20^{th}-order Butterworth bandpass filter. There can be observed stronger discrimination power between speech and non-speech sounds than between music and environment sound in each characteristic. So, we decided to used mean and not variance of BP parameter computed for four sub-bands. It helped us to increase a discrimination power between non-speech segments.

3.5 Spectral Flux

Spectral Flux (SF) is a measure of how quickly the magnitude of the power spectrum is changing. Lower values of SF parameter are specific for signals with slower changes of magnitude (background noises or music for example) and higher values are characteristic for signals (speech) with significant magnitude variations. (Figure 6). This parameter is therefore often used for speech/music discrimination in automatic audio classification systems [1, 9, 29].

3.6 Spectral Centroid

Spectral Centroid (SC) [9] is defined as a point in the signal's spectrum with dominant frequencies. Determines which frequencies are dominant in observed signal,

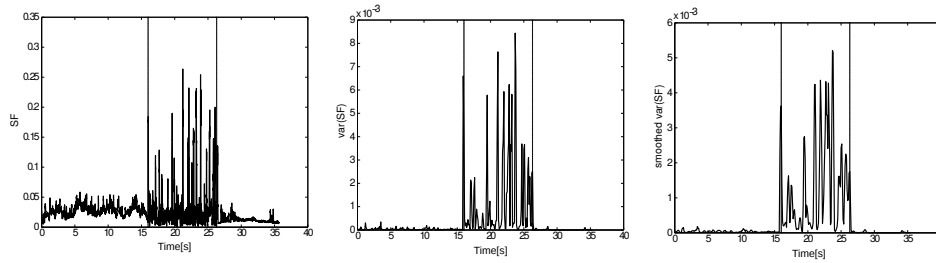


Figure 6. Spectral Flux

low or high. SC parameter has lower values for signals with narrow frequency range (speech for example). On the contrary, low variability but higher values of SC are specific for signals with wide spectrum and insignificant changes in frequency domain (music, environment sounds) (see Figure 7).

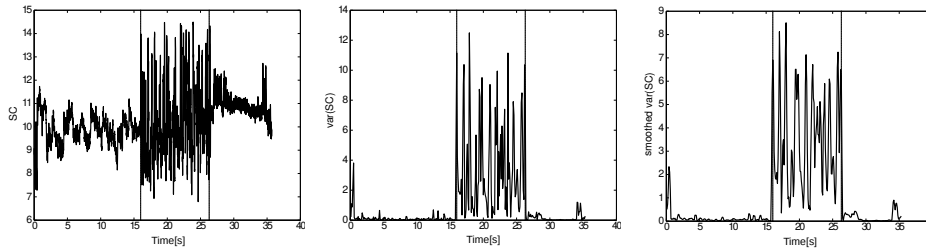


Figure 7. Spectral Centroid

3.7 Spectral Spread

Spectral Spread (SS) is a measure of the bandwidth of the spectrum. It provides an information about the power density of spectral components around the spectral centroid. Low values of this parameter show high concentration of spectral components around the centroid and higher values inform about wide frequency range of a signal (Figure 8). Authors in [30] used this parameter for speech/music segmentation.

3.8 Spectral Roll-Off

Spectral Roll-Off (ROLLOFF) is the measure of skewness of the signals frequency spectrum. It represents the frequency under which usually 95 % of the signals power resides. It is expected that speech has a lower value of spectral roll-off, because it

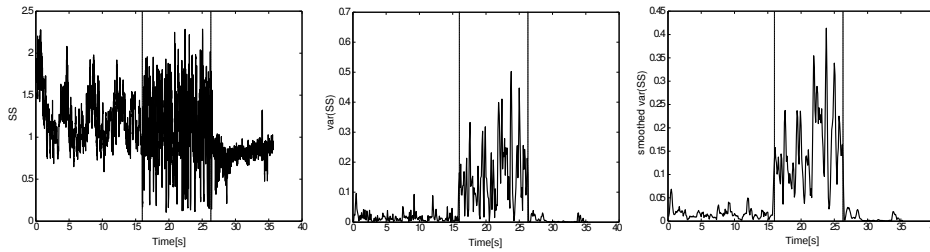


Figure 8. Spectral Spread

has most of the energy concentrated in the lower part of the frequency spectrum. The value of ROLLOFF parameter increases with the bandwidth of the signal's spectrum and is characterized by high variability for speech signals (Figure 9). It is often used as a speech/music discriminator [1, 31].

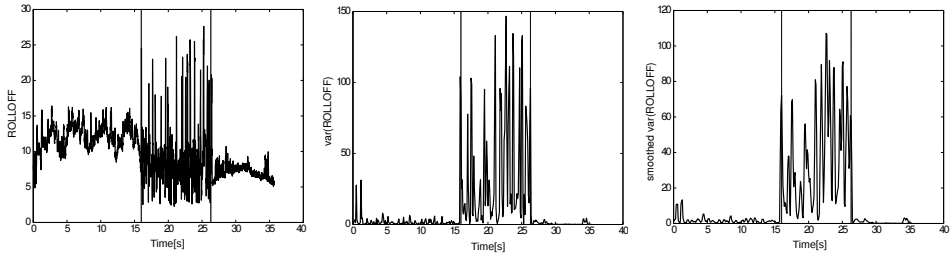


Figure 9. Spectral Roll-Off

4 BINARY DECISION ARCHITECTURE

Feature extraction, performed in the previous section, helped us to find several discrimination parameters for all the examined classification architectures.

Characteristic feature of proposed classification architecture (Figure 1) is smaller number of discrimination blocks for separation N classes in comparison with other architectures like one-against-one in which $(N(N-1)/2)$ classifiers are needed. An advantage of such solution lies in decision function which uses the optimal set of features adapted to actual discrimination problem in each block of classification. The optimal set of features can be defined as the set of such parameters that are able to capture spectral and temporal characteristics of different audio signals, emphasizing differences and keeping the maximum rate of separation between them. A big disadvantage is a miss-classification error which propagates from the top of the architecture. The classification performance of the current block depends on the

previous one. The main effort is therefore aimed on developing an architecture in which the miss-classification error is minimized. Our solution employs three-level binary decision architecture where the influence of the miss-classification error is attenuated by solving one binary classification problem four-times. The fundamental principles are based on a stepwise multi-level classification where each block tries to correct what previous one miss-classified.

The basic element of proposed architecture is the binary decision algorithm which performs binary classification of two classes implementing rule-based decision function and SVM-based non-linear decision function. It follows the naming convention "binary decision" in our proposed BDA. The next reason why we decided to use BDA naming convention is the fact that the overall classification strategy utilizes the same principles that apply in binary decision trees architecture , namely *coarse-to-fine* strategy. It allows to make a discrimination on coarse classes at the top of the decision tree architecture and then make stepwise classification towards the leaves which represent final classes. Therefore, the BDA is kind of modification of the BDT. The novelty of proposed BDA, comparing to BDT with using the SVM classifier Type 1 (see [32], p. 123), lies in successive discrimination amendment that is intended to attenuate the overfitting of each individual SVM classifier in BDA.

4.1 BDA Employing Rule-Based Approach

One possible way how to assess discrimination ability of feature parameters in solving binary classification problem for multiple classes is to build a classification architecture which implements a rule-based binary decision algorithm. Therefore, we proposed BD algorithm that performs binary discrimination of input feature vectors utilizing class-dependent mean distribution for each discrimination parameter and receiver operating characteristics (ROC) [33]. In other words, BD algorithm is intended to set the optimal threshold for each discrimination parameter in order to discriminate each pair of classes: speech/non-speech (S-NS), pure speech/non-pure speech (PS-NPS), music speech/background speech (MS-BS) and music/background (M-B). Decision function is in this case represented by optimal threshold for actual discrimination parameter. Thus, decision function $D(\mathbf{p}_j)$ takes for each component of input feature vector $\mathbf{p}_j : \{p_i\}$, $i \in [1 : M]$, $j \in [1 : P]$ two output values $+1$ and -1 depending on the actual value of discrimination parameter. p_i, $i \in [1 : M]$ relates to the component of each feature vector $\mathbf{p}_j$, $j \in [1 : P]$, where M defines the number of frames (or segments) per input audio file and P relates to the number of discrimination parameters (in our case $P = 7$). Thus, each individual feature vector represents one discrimination parameter with dimension $M \times 1$.

The overall procedure of BD algorithm is stated in Algorithm 1. The input parameters are represented by feature vectors extracted from audio recordings, namely VMFBE, VAMFCC, BP, SF, SC, SS, and ROLLOFF. The MFCC parameter was not considered in case of basic BD algorithm due to the need of computing the mean value over cepstral coefficients and subsequent loss of information about sig-

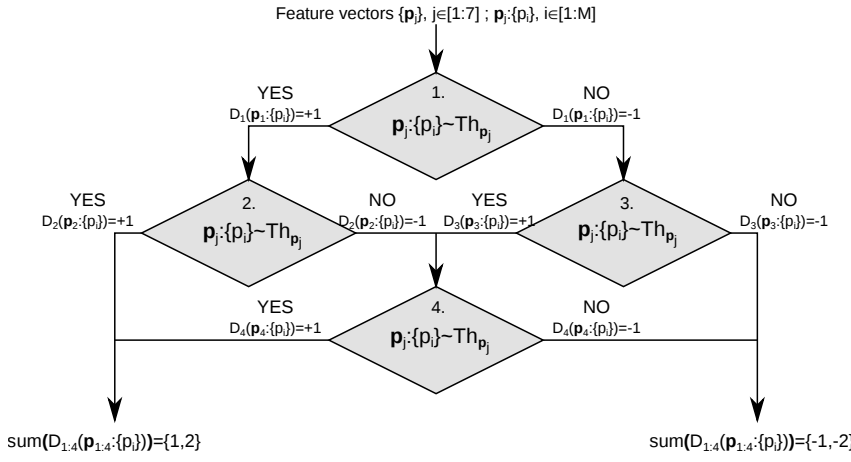
Algorithm 1 BD algorithm

```
1. if (mean(p_{j(+1)}) ≥ mean(p_{j(−1)})) then
2.     for (k = min(p_{j(+1)}); k ≤ max(p_{j(+1)}); k = k + step) do
3.         for (i = 0; i < length(p_j); i++) do
4.             if (p_j[i] >= k) then D[i] = +1;
5.             else D[i] = −1;
6.             end if
7.         end for
8.         diff[k] = abs(SPE(k) − SEN(k));
9.     end for
10.    Th = min(diff);
11. else if (mean(p_{j(+1)}) < mean(p_{j(−1)})) then
12.    for (k = min(p_{j(+1)}); k ≤ max(p_{j(+1)}); k = k + step) do
13.        for (i = 0; i < length(p_j); i++) do
14.            if (p_j[i] < k) then D[i] = +1;
15.            elseD[i] = −1;
16.            end if
17.        end for
18.        diff[k] = abs(SPE(k) − SEN(k));
19.    end for
20.    Th = min(diff);
21. end if
```

nal's cepstrum. The next reason of omitting this parameter was the use of two other single cepstral parameters instead, namely VMFBE and VAMFCC. The output of the algorithm is in a form of the optimal threshold for each input discrimination parameter. The thresholds are set for each parameter individually during a runtime of the algorithm. Thus, the initialization begins with loading the first discrimination parameter (VMFBE) for class S-NS for example. The class label $+1$ or -1 is assigned to each component of the feature vector manually, depending on the class membership. Consequently, a class distribution for each feature vector component is computed, represented by mean values $\text{mean}(\mathbf{p}_{j(+1)})$ and $\text{mean}(\mathbf{p}_{j(-1)})$, where $\mathbf{p}_{j(+1)}$ represents sub-vector containing components belonging to the class $+1$ and $\mathbf{p}_{j(-1)}$ sub-vector with components that belong to the class -1. The threshold varies from $\min(\mathbf{p}_{j(\pm 1)})$ to $\max(\mathbf{p}_{j(\pm 1)})$ with step $(\max(\mathbf{p}_{j(\pm 1)}) - \min(\mathbf{p}_{j(\pm 1)}))/100$ (found empirically) during training phase. Discrimination function is therefore represented by actual threshold value set for parameter $\mathbf{p}_j$. It follows that each component p_i of the actual parameter $\mathbf{p}_j$ takes values $+1$ or -1 according to the threshold value. Based on the predicted labels, specificity (SPE) and sensitivity (SEN) are calculated using the following formula:

$$SPE = \frac{TN}{FP + TN}, \quad SEN = \frac{TP}{TP + FN} \tag{1}$$

where TP relates to true positive values, TN true negative, FP false positive, and FN false negative. The optimal threshold for actual parameter is set by using the minimal value of the difference between SPE and SEN (min(diff)). This procedure is applied for each discrimination parameter and each pair of classes on each discrimination level. After setting the optimal threshold for each discrimination parameter, the one with maximal mean(SPE, SEN) value is selected as a winner for particular pair of classes. There is only one winner for particular discrimination problem (S-SN, PS-NPS, MS-BS, M-B), thus test audio recordings need to be parameterized by selected discrimination parameter (winner).



Components of feature vectors	$D_1(\mathbf{p}_1:\{p_i\})$	$D_2(\mathbf{p}_2:\{p_i\})$	$D_3(\mathbf{p}_3:\{p_i\})$	$D_4(\mathbf{p}_4:\{p_i\})$	sum$(D_{1:4}(\mathbf{p}_{1:4}:\{p_i\}))$
$\mathbf{p}_j:\{p_i\}$	+1	{+1,-1}	0	{0,+1,-1}	{2,+1,-1}
	-1	0	{+1,-1}	{+1,-1,0}	{-2,+1,-1}

Figure 10. Binary decision architecture for rule-based approach

Proposed solution of BDA implements three-level binary discrimination approach, where one classification problem is solved four times through BD algorithm. The proposed architecture is depicted in Figure 10. Binary decision for two-class problem is performed on each discrimination level and the output values are saved into a decision table. A dimension of the decision table is always $M \times d$, where M defines number of feature vector components and d number of decision functions, in our case $d = 4$. Each row of the table represents the actual frame or segment (depends on the level of feature extraction).

All the training vectors $\mathbf{p}_j$, where j defines number of parameters, enter the first block (first level) of discrimination at the beginning of classification. Consequently, BD algorithm is applied on each parameter and the output values of winner's decision function $D_1(\mathbf{p}_1 : \{p_i\})$ are saved into the first column of decision table.

The second level of discrimination is represented by two binary discriminators, marked as the 2. and the 3.. Considering the second discriminator, BD algorithm is applied only on those components of feature vectors which were classified as $+1$ in the previous level of discrimination (the first one). The same procedure as was used on the first level is applied and the output values of winner's decision function $D_2(\mathbf{p}_2 : \{p_i\})$ are saved into the second column of decision table. Decision function on this level takes the values $+1$ or -1 if $D_1(\mathbf{p}_1 : \{p_i\}) = +1$ and 0 if $D_1(\mathbf{p}_1 : \{p_i\}) = -1$. On the contrary, BD algorithm is used within the third discriminator only for those components of feature vectors which were predicted into the class -1 on the first level. Similarly, the output values of winner's decision function $D_3(\mathbf{p}_3 : \{p_i\})$ are saved into the third column of decision table with values $+1$ or -1 if $D_1(\mathbf{p}_1 : \{p_i\}) = -1$ and 0 if $D_1(\mathbf{p}_1 : \{p_i\}) = +1$.

On the third level of discrimination is BD algorithm applied only on those components of feature vectors which were predicted on the first level into the class $+1$ and on the second level into the class -1, or on the first level into the class -1 and on the second level into the class $+1$. Thus, considering decision table elements in a row order, the sum of elements for column 1,2 and 3 is equal to zero: $\mathrm{sum}(D_{1:3}(\mathbf{p}_{1:3} : \{p_i\})) = 0$. Decision function for particular discriminator takes the values $+1$ or -1 if $\mathrm{sum}(D_{1:3}(\mathbf{p}_{1:3} : \{p_i\})) = 0$ and 0 if $\mathrm{sum}(D_{1:3}(\mathbf{p}_{1:3} : \{p_i\})) = \pm 2$.

The final class labels for each row (feature vectors elements) is assigned at the output of BDA. The class label $+1$ is assigned for actual row if $\mathrm{sum}(D_{1:4}(\mathbf{p}_{1:4} : \{p_i\})) = \{1, 2\}$. On the contrary, class label -1 is assigned if $\mathrm{sum}(D_{1:4}(\mathbf{p}_{1:4} : \{p_i\})) = \{-1, -2\}$.

4.2 BDA Employing SVM

For moderately easy tasks like speech/non-speech segmentation in real time the rule-based approach is the right choice due to its efficiency and low computational complexity in solving particular classification task. It is necessary to implement a robust classifier for solving multiclass discrimination problem, with use of learning-machine algorithm, for more difficult tasks like segmentation and classification of broadcast news audio data, where more than two different acoustic events are presented. We decided to use the SVM classifier [32] in our classification scheme due to its generalization ability and superior performance in various pattern classification tasks. The SVM classifier is currently one of the most demanded discriminators, often implemented in various data classification tasks, such as bioinformatics [34], image [35] and audio classification [1, 36, 37, 38].

The input audio data is represented in the form of feature matrix $\mathbf{X} : \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ with dimension $M \times N$, where $\mathbf{x}_m = (x_1, \ldots, x_N)$, $m \in [1 : M]$ are input vectors with dimension N and class label $y_m = \pm 1$. N defines number of coefficients per frame (or segment) and M gives the overall number of frames (segments). Decision function is modeled by separating hyperplane with maximal or soft margin:

$$d(\mathbf{x}_m, \mathbf{w}, b) = \mathbf{w}^T \mathbf{x}_m + b = \sum_{i=1}^{N} w_i x_i + b \tag{2}$$

where $\mathbf{w}^T \mathbf{x}_m$ represents dot product between a weight vector and the input feature vector and the scalar b is called *bias*.

The one that maximizes the margin of hyperplanes between two classes is called optimal hyperplane. The margin is adjusted by distance between nearest point to the hyperplane in case of linearly separable training vectors. If the training data from different classes cannot be linearly separated in the original input space, the SVM at first non-linearly transforms the original input space into the high-dimensional *feature space*. This transformation can be achieved by using various kernel functions such as: linear, polynomial, RBF mappings having as basic function radially symmetric function, i.e. Gaussian function. The margin is in case of non-linear mappings adjusted by penalty parameter C.

After successful training stage the learning machine produces the output $D(\mathbf{x}_m)$, given as:

$$D(\mathbf{x}_m) = \text{sign}(d(\mathbf{x}_m, \mathbf{w}, b)), \tag{3}$$

utilizing weights obtained from the process of learning (training). Decision rule is defined by the following criteria:

- if $d(\mathbf{x}_m, \mathbf{w}, b) \geq 0$, then $D(\mathbf{x}_m) = +1$,
- if $d(\mathbf{x}_m, \mathbf{w}, b) < 0$, then $D(\mathbf{x}_m) = -1$.

The SVM is originally designed to solve binary classification problem and discrimination of multiple classes is realized by combining several binary classifiers. One of the most used architecture for multiclass classification is binary decision trees topology (BDTSVM). Classification procedure of the BDTSVM is based on a coarse-to-fine strategy. The coarse classification that separates two easy to differentiate classes, i.e. speech and non-speech, is performed at the beginning of the classification process. Then the stepwise classification is made, until the other classes are obtained. It follows that the multiclass classification needs to train maximally $(n-1)$ SVMs for n-class problem. The main advantage of BDTSVM lies in choosing feasible feature set for each SVM binary classifier that separates two different classes.

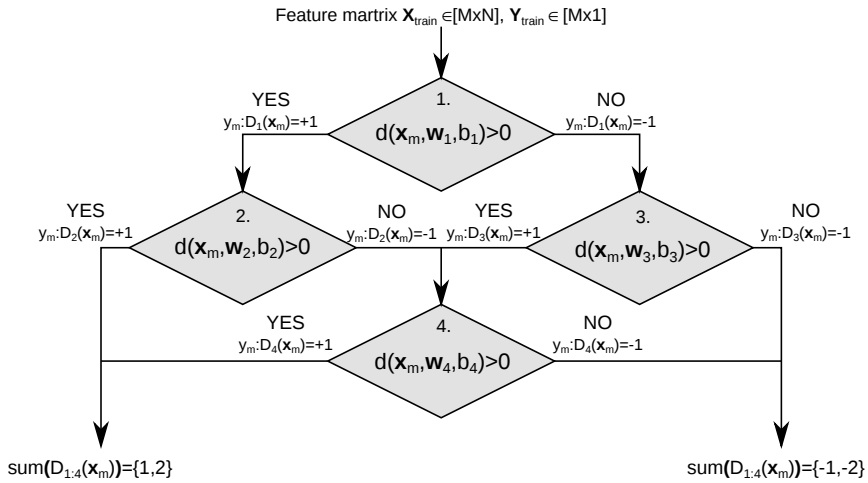
One-Against-One SVM (OAOSVM) or pairwise topology belongs to the one of the most effective architectures implemented in many classification problems. In OAOSVM, there is one SVM for each pair of classes trained to separate each individual class using $n(n-1)/2$ SVM classifiers. The simplest form of classification with pairwise SVMs selects the class chosen by the maximal number of votes from each SVM classifier. Thus each pair of SVMs classify only input data from two classes.

One-against-all SVM (OAASVM) is one of the earliest and simplest multiclass classification approaches. For the n-class classification problem, it constructs n binary SVMs with each one separating one class from all the others. More detailed description about these three and other classification architectures can be found in [1, 32, 39, 40].

Algorithm 2 BDSVM algorithm

1. load($\mathbf{X}_{train}, \mathbf{Y}_{train}$);
2. scaling($\mathbf{X}_{train}$);
3. **for** ($m = 0$; $m <$ length($\mathbf{X}_{train}$); m++) **do** $y_m = +1, y_{m+1} = -1$; **end for**
4. (best_C, best_g, AUC) = crossvalidation($\mathbf{X}'_{train}, \mathbf{Y}'_{train}, \log_2 C$ [0 6 2], $\log_2 g$ [0 6 2], υ 5);
5. (model) = svm_train($\mathbf{X}'_{train}, \mathbf{Y}'_{train}$, best_$C$, best_$g$);

Proposed solution for BD algorithm utilizing SVM is stated in Algorithm 2. The basic principles of BDSVM follow those used in BD algorithm (Algorithm 1). Thus, the main task is to find optimal parameters for binary decision function on each level of classification using training data. The optimal setting is understood as the process of finding parameters for kernel function and penalty parameter in process of training. The input of the BDSVM algorithm is represented as a feature matrix corresponding to training data $\mathbf{X}_{train}$, with dimension $M \times N$. M defines number of frames (segments) within an audio file and N corresponds with the overall number of coefficients analysed in Section 3. Scaling values in the range of 0–1 ensures function scale. It helps to eliminate big differences between coefficients and can be considered as some kind of smoothing. Training feature vectors for each pair of classes S-NS, PS-NPS, MS-BS, M-B, are consequently reordered alternately in succession. Thus, each feature vector takes the values $y_m = +1$ and $y_{m+1} = -1$, while number of vectors for both classes is the same. Reordered feature vectors and labels are assigned as $\mathbf{X}'_{train}$, $\mathbf{Y}'_{train}$. This step helped us to optimize process of cross-validation and alleviate overfitting of classifier. Cross-validation technique, also known as leave-one-out cross-validation) [41], is then applied in order to find optimal parameters of kernel function (g) and penalty parameter (C). After several initial experiments, we decided to use 5-fold cross-validation and RBF kernel function as the best choice. The parameters C and g were adjusted exponentially, taken the values 2^0, 2^2, 2^4 and 2^6. AUC (Area Under the Curve) [42] parameter was used as the main evaluation criterion during cross-validation. This evaluation parameter is similar to the one used in BD algorithm (ROC). Values of AUC vary between 0 and 1. The binary classification is considered as random if the value of AUC is lower than 0.5. The highest value of AUC signifies the most optimal (best) parameters C and γ. The optimal parameters are then used to generate *model* by using *svm_train* function. Model is represented by weighted vector, bias term and by the number of support vectors for each class. Generated model is then used to generate predicted class labels for testing data, in the final phase of classification.



Feature vectors	$D_1(\mathbf{x}_m)$	$D_2(\mathbf{x}_m)$	$D_3(\mathbf{x}_m)$	$D_4(\mathbf{x}_m)$	$\text{sum}(D_{1:4}(\mathbf{x}_m))$
$\mathbf{x}_m$	+1	{+1,-1}	0	{0,+1,-1}	{2,+1,-1}
	-1	0	{+1,-1}	{+1,-1,0}	{-2,+1,-1}

Figure 11. Binary decision architecture for SVM

Proposed BDASVM topology utilizes the same classification principles as in case of BDA (Figure 10). Each block of discrimination is represented by one SVM binary classifier and the same rules are applied for decision table. The overall architecture is depicted in Figure 11. We have observed relatively high rate of overfitting on the second level of discrimination in BDASVM after several initial experiments. The overfitting caused the increase of miss-classification error rate, such that the majority of feature vectors in the second discriminator was classified to the class $+1$ and in the third discriminator to the class -1. The number of input vectors on the third level was therefore insufficient (very small) for training the SVM classifier at this level. This unwanted effect was caused by a high discrimination power on the first level of classification. It follows that the maximum classification accuracy was obtained on the first level of classification in BDASVM, utilizing all available training data. There was a need to decrease discrimination ability on the first level and increase on the third level in order to suppress this effect. Partial solution was to divide training set at the input of BDASVM into two parts $\mathbf{X}_{train\_train}$ and $\mathbf{X}_{train\_test}$ with equal size. Feature vector matrix $\mathbf{X}_{train\_train}$ was used for training the SVM and $\mathbf{X}_{train\_test}$ for testing on the first level of discrimination.

The second level of discrimination enter the feature vectors classified to the class $+1$: $\mathbf{X}_{train\_test_{+1}}$ in case of second discriminator. On the contrary, the third discriminator enter the feature vectors classified to the class -1: $\mathbf{X}_{train\_test_{-1}}$ on the

first level of classification. The whole training set of feature vectors $\mathbf{X}_{train}$ was used for training the SVM on the third level, regardless the testing vectors which enter the classifier from level two.

Generally speaking, following set of feature vectors (matrices) is used within the training phase in BDASVM:

1. discriminator: $\mathbf{X}_{train\_train}$,
2. discriminator: $\mathbf{X}_{train\_test_{+1}}$,
3. discriminator: $\mathbf{X}_{train\_test_{-1}}$,
4. discriminator: $\mathbf{X}_{train}$.

The whole set of testing data $\mathbf{X}_{test}$ enters the BDASVM in testing phase and the values of decision functions are written to the decision table in the same way as was performed in case of BDA (Section 4.1).

The maximum classification accuracy was achieved by adding weighted factor w_C to each discrimination level. We defined it as the value of penalty parameter C, as the product of cross-validation, divided by number of training vectors belonging to particular class. Thus, for class +1: $w_{C_+} = C/num_{+1}$ and for class −1: $w_{C_-} = C/num_{-1}$. In a certain way, weighted factor helps to decrease the influence of overfitting by adding higher weight to the vectors belonging to the minor class and lower weight to the vectors that belong to the major class. More detailed description about the implementation into SVM training algorithm can be found in [43].

5 TEST SYSTEM DESCRIPTION

The classification performance of proposed BDA and BDASVM architectures and comparison with BDT topology was evaluated on KEMT-BN1 database [44], which contains 188 audio recordings in PCM 16 kHz 16 bit mono format from the Slovak TV broadcast audio streams. Total duration of all audio recordings is about 65 hours.

We used only part of the database in our experiments, namely 49 min for training (PS: 10.19 min, MS: 9.26 min, BS: 9.41 min, M: 11.7 min, B: 9.06 min) and 46.2 min for testing (PS: 9.16 min, MS: 9.44 min, BS: 9.25 min, M: 9.04 min, B: 9.31 min). Our aim was to extract maybe smaller amount of audio data but more accurate with equal size for each audio class in order to avoid the overfitting of classifier. Silent parts and other audio events were extracted manually without using any phonetic alignment techniques but only available transcription on word level.

Detailed description about process of feature extraction and smoothing on frame and segment level is illustrated in Figure 12. Each coefficient was at first extracted on frame level within 50 ms Hamming window and 25 ms overlapping. The variance of 7 coefficients is then computed within 200 ms segment with 100 ms overlapping. Coefficients of 8 parameters are used to build feature vector matrix with dimension $M \times N$, where $N = 23$.

Each individual parameter is consequently smoothed by floating window with length 500 ms, thus mean value is computed for 5 adjacent coefficients within one

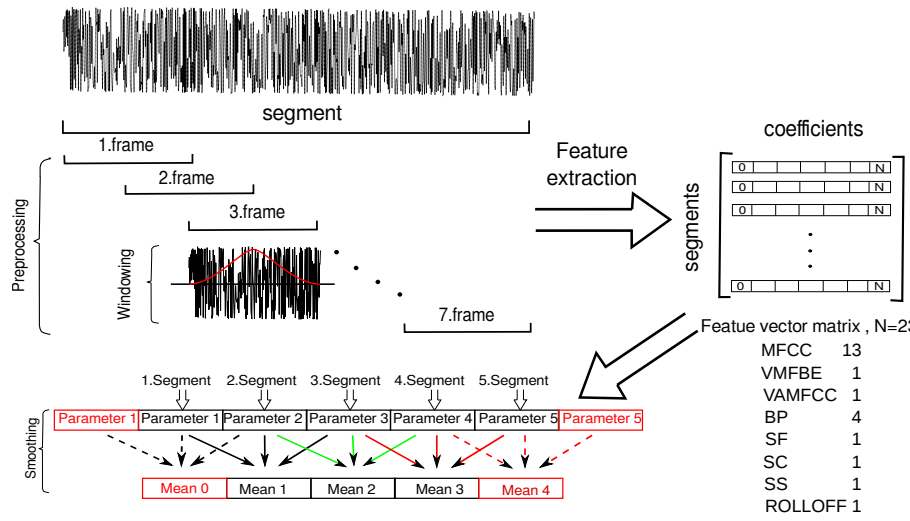
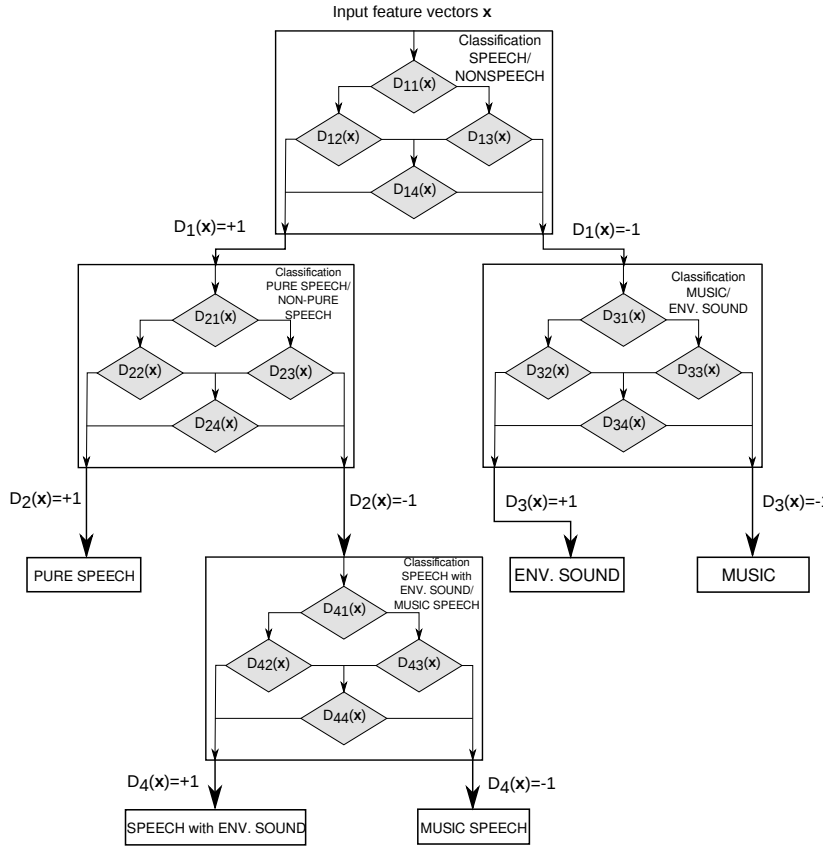


Figure 12. Preprocessing, feature extraction and smoothing of audio stream

floating window. However, there is a need to consider time alignment according to the length of floating window. In other words, there is a need to start smoothing from zero position and not from $(0 + l_{fw}/2)^{\text{th}}$ position, where l_{fw} represents length of floating window. Such alignment is possible with enlargement of each feature vector by length $l_{fw}/2$ at the beginning and at the end of each array. Consequently, coefficients from positions $(0 + l_{fw}/2)$ and $((M-1) - l_{fw}/2)$ are copied to the new positions (extensions at the beginning and at the end of each feature vector).

The last step in our experimental work was to implement proposed BDA architecture into classification topology for BN data. The overall architecture is depicted in Figure 13. Each block of classification is represented by one BDA module for discrimination S-NS on the first level, PS-NPS on the second level, M-B on the third level and finally MS-BS on the fourth level. Input audio data is represented by feature vectors $\mathbf{x}$ with dimension $M \times 1$ in case of rule-based approach used in BDA and $1 \times N$ for BDASVM topology as a part of feature input matrix $\mathbf{X}$. The output gives the information about the particular audio class in audio stream. During the testing phase, depending on the output values of decision functions $D_i(\mathbf{x}), i \in [1:4]$, label $+1$ or -1 is assigned to each input feature vector on each level of discrimination and saved into the decision table. Final class label is assigned according to the following criteria:

- PS: $\text{sum}(D_1(\mathbf{x}), D_2(\mathbf{x})) = 2$, if $D_1(\mathbf{x}) = 1$;
- B: $\text{sum}(D_1(\mathbf{x}), D_3(\mathbf{x})) = 0$, if $D_1(\mathbf{x}) = -1$;
- M: $\text{sum}(D_1(\mathbf{x}), D_3(\mathbf{x})) = -2$, if $D_1(\mathbf{x}) = -1$;



Input vectors	$D_1(\mathbf{x})$	$D_2(\mathbf{x})$	$D_3(\mathbf{x})$	$D_4(\mathbf{x})$	sum($D_{1:4}(\mathbf{x})$)
x	+1	{+1,-1}	0	{+1,-1}	{2,+1,-1}
	-1	0	{+1,-1}	0	{0,-2}

Figure 13. Implementation of BDA into classification architecture

- MS: $\text{sum}(D_1(\mathbf{x}), D_2(\mathbf{x}), D_4(\mathbf{x})) = -1$, if $D_1(\mathbf{x}) = 1$ and also $D_2(\mathbf{x}) = -1$;
- BS: $\text{sum}(D_1(\mathbf{x}), D_2(\mathbf{x}), D_4(\mathbf{x})) = 1$, if $D_1(\mathbf{x}) = 1$ and also $D_2(\mathbf{x}) = -1$;

where PS is for pure speech, B for environment sound (background), M for music, MS is for music speech and BS for background speech. sum() represents a sum of decision function values for one row in decision table.

We used classification accuracy *Acc* as the main evaluation criterion, defined as the ration of correctly classified frames to all predicted frames for each audio class.

All experiments were performed by LIBSVM software[1] and available tools, which cooperate with this software. MATLAB software was used in order to evaluate the characteristics of features from Section 3.

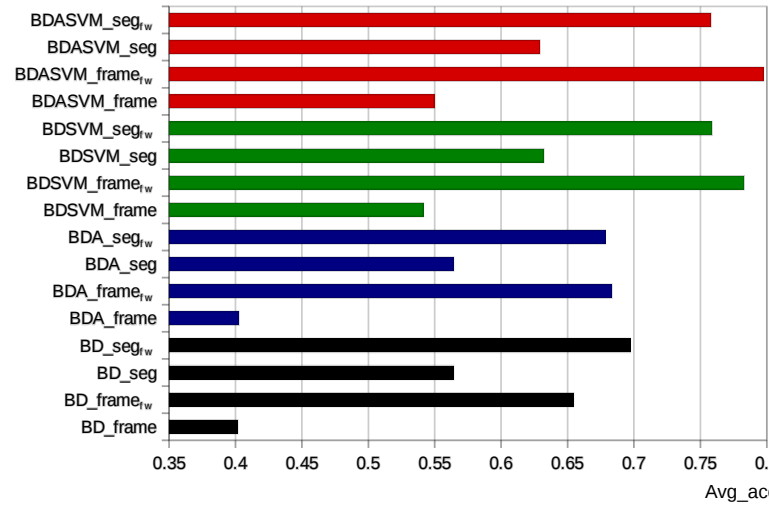


Figure 14. Comparison between classification architecture employing rule-based approach (BD, BDA) and SVM (BDSVAM, BDASVM) and four-level feature extraction technique

System Acc [%]	**PS**	**MS**	**SES**	**M**	**B**	**Avg**	**PT** [*min*]
BD(BDT)	73.11	45.51	20.29	66.08	79.53	**56.90**	*0.74*
BDA	75.2	43.89	30.34	56.29	85.29	**58.20**	*2.54*
BDSVM(BDTSVM)	85.69	54.46	48.63	72.75	77.83	**67.87**	*44.13*
BDASVM	85.94	53.29	48.94	72.85	80.74	**68.35**	*48.37*
OAOSVM	86.92	53.22	46.79	76.49	86.13	**69.91**	*24.56*

Table 1. The overall evaluation of used classification architectures

6 RESULTS AND CONCLUSIONS

The classification performance of basic BD algorithm and the proposed BDA solution utilizing four-level feature extraction was evaluated in the first phase. The results obtained from fundamental BDT architecture and improved BDA topology, employing BD and BDSVM, for parameterization on frame level, segment level with smoothing and without smoothing by floating window (fw) are depicted in Figure 14. Value of *Avg_acc* was obtained by averaging the classification accuracy for each class,

[1] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

namely PS, MS, BS, M, B, on each level of feature extraction. BD algorithm itself, applied for multiclass classification through the basic topology depicted in Figure 1, behaves as a basic binary discriminator in BDT architecture (black and green colour stripes). On the contrary, blue and red colour stripes illustrates the average classification accuracy for proposed BDA solution, representing four discriminators for one binary classification problem. It can be seen that segment-based feature extraction helped to increase the classification accuracy only for basic BD algorithm, comparing BD_{frame} and BD_{seg}, with and without smoothing as well. Segment-based feature extraction showed improvement for other architectures (BDA, BDSVM and BDASVM) only in cases when the smoothing technique was not applied. Smoothing technique helped to increase classification performance only for the same type (level) of parameterization, comparing *frame* with $frame_{fw}$ and *seg* with seg_{fw}. Hereby, the advantage of using segment-based feature extraction and smoothing technique for multiclass classification was proved.

The overall classification performance for each type of architecture is given in Table 1. Each individual value of *Acc* represents the average of four-level feature extraction for particular class. We decided to use only the average value, referring to the classification abilities of each classification scheme, regardless the level of feature extraction. Such interpretation of results helps to minimize redundant information and keeps the substantial information about system's performance. *PT* corresponds with processing time needed for classifying each testing feature vector into the particular class. BD algorithm using rule-based approach and the SVM classifier represents the basic BDT and the BDTSVM architecture.

The lowest classification performance was achieved in case of basic BD algorithm, as we assumed. The expected increase in classification accuracy was ensured by proposed rule-based BDA solution. The average value of *Acc* was about 1.3 % higher than in case of simple BD algorithm, at the expense of almost 4-times higher processing time. However, the differences between *Acc* values for individual acoustic classes are significant, especially for class music (M), environment sound (B) and background speech (BS). More than 10 % increase in classification accuracy can be observed for music class in case of basic BD algorithm. These considerable variations of *Acc* values could be caused by miss-classification error propagated from the first level of discrimination and by BD algorithm itself due to its class-dependent separation using only "optimal" discrimination parameter for making decision.

Almost 11 % enhancement in classification accuracy brought into the results the implementation of SVM decision function to BD algorithm (BDSVM). The cost for that was the huge increase in processing time. The BDASVM topology has the positive effect to overall classification performance. The 0.48 % increase of *Acc* and only 4.24*s* growth of *PT* was achieved in comparison with BDSVM. Only small increase in processing time was caused by loading all the testing data at once in the first step of classifying. We assume that the main cause for relatively low enhancement in classification performance was a high influence of level-dependent miss-classification error in case of MS-BS discrimination problem. Level-dependent error propagates only within one BDA block.

So far, the effectiveness of using BDA architecture for rule-based approach and the SVM was proven by comparing classification performance with basic BDT topology employing BD algorithm. It was necessary to compare it with more effective architecture in order to prove its robustness in solving multiclass classification task.

Considering our previous experience with SVM classifier and BDT architecture, in [45] we proposed solution for BN classification utilizing the combination of SVM-BDT architecture and discrimination principles that apply in case of the OAA topology. The average classification accuracy for all classes was even higher than in case of our currently developed BDASVM architecture (from Table II in [45]: average_Acc(SVM-BDT) = 70.37 %, from Table 1: average_Acc(BDASVM) = 67.87 %). However, the results obtained from the SVM-BDT addressed only discrimination ability of individual SVM classifiers and not all of them, meaning that we reported only results without miss-classification error. It was really hard to use this kind architecture for classification of BN audio stream due to really high miss-classification error that propagated from the very top of the SVM-BDT topology. Therefore we have been making an effort in order to diminish this drawbacks of the SVM-BDT architecture. Such disadvantages were partially eliminated by BDASVM, where the influence of the miss-classification error is attenuated on each level of classification.

One of the most used architectures, alongside BDT and OAA, is OAO classification architecture. The effectiveness of the OAO architecture in solving various classification problems examined authors in [46, 47]. They concentrated the effort on training time, cross-validation accuracy and the overall prediction accuracy within the evaluation phase. They also pointed to ambiguity of choosing classification architecture for solving multiclass classification problem. Finally, they recommended the OAO architecture as the optimal solution due to its high classification rate and the lowest training time. That was the reason why they implemented the OAO into *libsvm* software. It follows our decision to compare proposed BDASVM with OAOSVM.

It is obvious, from Table 1, that BDASVM solution did not overcome OAOSVM approach. Considerable reduction of classification accuracy was observed for music and environment sound. Deeper analysis of miss-classification error rate for both architectures showed the 8.75 % decrease in classification accuracy for BDASVM using frame-level feature vectors, see Table 2.

System Acc [%]	**frame**	**frame$_{\mathbf{fw}}$**	**seg**	**seg$_{\mathbf{fw}}$**
OAOSVM	63.70	77.38	64.41	74.16
BDASVM	54.95	**79.75**	62.91	**75.79**

Table 2. Feature level comparison between BDASVM and OAOSVM

It can be noticed that the use of smoothing technique helped to increase the classification accuracy for frame-based and segment-based feature extraction in case of BDASVM. Thus, if the accuracy on frame-level achieved higher values, the

OAOSVM	**PS**	**MS**	**BS**	**M**	**B**
PS	**38 575**	3 016	12 375	489	326
MS	13 861	**21 446**	15 818	3 191	465
BS	12 849	3 853	**35 186**	1 255	1 638
M	2 331	6 567	11 016	**32 871**	1 996
B	1 572	239	5 649	919	**46 402**
BDASVM	**PS**	**MS**	**BS**	**M**	**B**
PS	**36 850**	3 303	13 559	484	585
MS	11 738	**21 185**	20 697	788	373
BS	11 522	3 569	**37 429**	980	1 281
M	2 375	8 808	19 613	**21 166**	2 819
B	1 380	163	19 141	201	**33 896**

Table 3. Confusion matrix for BDASVM and OAOSVM on frame level feature vectors. The overall number of testing frames is 54 781

BDASVM would overcome the OAOSVM architecture. We decided to make a closer examination of this problem. Therefore, we created confusion matrix for both architectures on frame level (see Table 3).

The diagonal elements represents correctly predicted frames (marked by bold letters), column elements refer to the frames predicted by system and row elements relate to the all tested frames. The results obtained from confusion matrix proved the claim (mentioned above, based on the results obtained from Table 1) that the highest miss-classification error occurred for music and environment sound in case of BDASVM. A huge number of frames from classes M and B was predicted to the class BS, concretely 19 613 for M and 19 141 for B. It is the explanation for considerable drop in classification accuracy for the BDASVM. A big number of feature vectors was also classified into the classes BS and PS and contributed to significant increase of miss-classification error in both architectures. We assume that such increase of error rate can be explained by acoustic characteristics of the class BS. The audio utterances, extracted for this class, contain alongside speech also short pauses between speakers including environment sounds and background noise from other speakers. It follows that the frames, belonging to the class B, were classified as BS and frames belonging to BS predicted as PS in case of short silent speech parts.

The facts mentioned above had the major influence on miss-classification error rate. These drawbacks can be eliminated by more precise extraction of each individual classes from audio files using phone-level alignment. We also suppose, that the feature vector selection algorithm for selecting optimal training and testing set at the beginning of BDSVM algorithm could bring the higher classification accuracy for proposed BDASVM solution. The training of classifiers could be also improved by in parallel training the BDT along with the one-against-all SVM using all the training set with corresponding disjunct subsets, meaning that each SVM classifier would have information about all other classes.

These two aspects will be considered in the near future. The implementation of BDASVM in retrieving system utilizing the ASR will be performed in our future work as well.

The classification algorithms were running on High Performance Computing system with 24 nodes, each one contains computing server IBM Blade System × HS22 with two six-core processor units Intel Xeon L5640 (2.27 GHz) and 48 GB RAM.

Acknowledgements

The research in this paper was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project VEGA 2/0197/15 and the Slovak Research and Development Agency under the project APVV-15-0517.

REFERENCES

[1] Xie, L. Z.—Fu, H.—Feng, W.—Luo, Y.: Pitch-Density-Based Features and an SVM Binary Tree Approach for Multiclass Audio Classification in Broadcast News. Multimedia Systems, Vol. 17, 2011, No. 2, pp. 101–112.

[2] Abad, A.—Meinedo, H.—Neto, J.: Automatic Classification and Transcription of Telephone Speech in Radio Broadcast Data. International Conference on Computational Processing of the Portuguese Language (PROPOR 2008). Springer, LNCS, Vol. 5190, 2008, pp. 172–181, doi: 10.1007/978-3-540-85980-2_18.

[3] Heck, M.—Mohr, C.—Stueker, S.—Mueller, M.—Kilgour, K.—Gehring, J.—Nguyen, Q.—Nguyen, V.—Waibel, A.: Segmentation of Telephone Speech Based on Speech and Non-Speech Models. International Conference on Speech and Computer (SPECOM 2013). Springer, LNAI, Vol. 8113, 2013, pp. 286–293, doi: 10.1007/978-3-319-01931-4_38.

[4] Darjaa, S.—Cernak, M.—Trnka, M.—Rusko, M.—Sabo, R.: Effective Triphone Mapping for Acoustic Modeling in Speech Recognition. 12th Annual Conference of the International Speech Communication Association (INTERSPEECH '11), Florence, Italy, August 2011, pp. 1717–1720.

[5] Panda, S. P.—Nayak, A.: A Rule-Based Concatenative Approach to Speech Synthesis in Indian Language Text-to-Speech Systems. Proceedings of ICCD 2014. Springer, Advances in Intelligent Systems and Computing, Vol. 309, 2015, pp. 523–531, doi: 10.1007/978-81-322-2009-1_59.

[6] Gravier, G. J.—Bonastre, F.—Geoffrois, E.—Galliano, S.—McTait, K.—Choukri, K.: The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, May 2004, pp. 885–888.

[7] Butko, T.—Nadeu, C.: Audio Segmentation of Broadcast News in the Albayzin-2010 Evaluation: Overview, Results, and Discussion. EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2011, 2011, No. 1, p. 10, doi: 10.1186/1687-4722-2011-1.

[8] Valero, X.—Alias, F.: Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification. IEEE Transactions on Multimedia, Vol. 14, 2012, No. 6, pp. 1684–1689, doi: 10.1109/TMM.2012.2199972.

[9] Kos, M.—Grasic, M.—Kacic, Z.: Online Speech/Music Segmentation Based on the Variance Mean of Filter Bank Energy. EURASIP Journal on Advances in Signal Processing, Vol. 2009, 2009, No. 1, p. 13, doi: 10.1155/2009/628570.

[10] Mesgarani, N.—Slaney, M.—Shamma, S.: Discrimination of Speech from Non-speech Based on Multiscale Spectro-Temporal Modulations. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, 2006, No. 3, pp. 920–930, doi: 10.1109/TSA.2005.858055.

[11] Kos, M.—Kacic, Z.—Vlaj, D.: Acoustic Classification and Segmentation Using Modified Spectral Roll-Off and Variance-Based Features. Digital Signal Processing, Vol. 23, 2013, No. 2, pp. 659–674.

[12] Song, Y. W.—Wang, H.—Guo, F. J.: Feature Extraction and Classification for Audio Information in News Video. International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR 2009), Baoding, China, July 2009, pp. 43–46, doi: 10.1109/ICWAPR.2009.5207452.

[13] Dogan, E.—Sert, M.—Yazici, A.: Content-Based Classification and Segmentation of Mixed-Type Audio by Using MPEG-7 Features. International Conference on Advances in Multimedia (MMEDIA 2009), Colmar, France 2009, pp. 152–157, doi: 10.1109/MMEDIA.2009.35.

[14] Castán, D.—Ortega, A.—Miguel, A.—Lleida, E.: Audio Segmentation-by-Classification Approach Based on Factor Analysis in Broadcast News Domain. EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2014, 2014, No. 1.

[15] Zhang, R.—Li, B.—Peng, T.: Audio Classification Based on SVM-UBM. 9th International Conference on Signal Processing (ICSP 2008), Beijing, China, October 2008, pp. 1586–1589.

[16] Theodorou, T.—Mporas, I.—Fakotakis, N.: Automatic Sound Classification of Radio Broadcast News. International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 5, 2012, No. 1.

[17] Cernak, M.: A Comparison of Decision Tree Classifiers for Automatic Diagnosis of Speech Recognition Errors. Computing and Informatics, Vol. 29, 2010, No. 3, pp. 489–501.

[18] Chen, L.—Gunduz, S.—Ozsu, M.: Mixed Type Audio Classification with Support Vector Machine. IEEE International Conference on Multimedia and Expo (ICME 2006), Toronto, Canada, July 2006, pp. 781–784, doi: 10.1109/ICME.2006.262954.

[19] Mukherjee, S.: Classifying Microarray Data Using Support Vector Machines. In: Berrar, D. P., Dubitzky, W., Granzow, M. (Eds.): A Practical Approach to Microarray Data Analysis. Springer, 2003, pp. 166–185, doi: 10.1007/0-306-47815-3_9.

[20] David, A.—Lerner, B.: Support Vector Machine-Based Image Classification for Genetic Syndrome Diagnosis. Pattern Recognition Letters, Vol. 26, 2005, No. 8, pp. 1029–1038.

[21] Wei, L.—Wei, B.—Wang, B.: Text Classification Using Support Vector Machine with Mixture of Kernel. Journal of Software Engineering and Applications, Vol. 5, 2012, No. 12B, pp. 55–58.

[22] Mahale, P.—Rashidi, M.—Faez, K.—Sayadiyan, A.: A New SVM-Based Mix Audio Classification. 40th Southeastern Symposium on System Theory (SSST 2008), New Orleans, USA, March 2008, pp. 198–202.

[23] Vavrek, J.—Vozarikova, E.—Pleva, M.—Juhar, J.: Broadcast News Audio Classification Using SVM Binary Trees. 35th International Conference on Telecommunications and Signal Processing (TSP 2012), Prague, Czech Republic, July 2012, pp. 469–473, doi: 10.1109/TSP.2012.6256338.

[24] Vavrek, J.—Juhar, J.—Cizmar, A.: Audio Classification Utilizing a Rule-Based Approach and the Support Vector Machine Classifier. 36th International Conference on Telecommunications and Signal Processing (TSP 2013), Rome, Italy, July 2013, pp. 512–516, doi: 10.1109/TSP.2013.6613985.

[25] On, C. K.—Pandiyan, P.—Yaacob, S.—Saudi, A.: Mel-Frequency Cepstral Coefficient Analysis in Speech Recognition. International Conference on Computing and Informatics (ICOCI 2006), Kuala Lumpur, Malaysia, June 2006, pp. 1–5, doi: 10.1109/ICOCI.2006.5276486.

[26] Tiwari, V.: MFCC and Its Applications in Speaker Recognition. International Journal on Emerging Technologies, Vol. 1, 2010, No. 1, pp. 19–22.

[27] Huang, X.—Acero, A.—Hon, H. W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR, 2001.

[28] Lu, L. H.—Zhang, J.—Jiang, H.: Content Analysis for Audio Classification and Segmentation. IEEE Transactions on Speech and Audio Processing, Vol. 10, 2002, No. 7, pp. 504–516.

[29] Khan, M.—Al Khatib, W.: Machine-Learning Based Classification of Speech and Music. Multimedia Systems, Vol. 12, 2006, No. 1, pp. 55–67.

[30] Lavner, Y.—Ruinskiy, D.: A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation. EURASIP Journal of Audio, Speech, and Music Processing, Vol. 2009, 2009, No. 2, p. 14, doi: 10.1155/2009/239892.

[31] Schuller, B.—Schmitt, B.—Arsic, D.—Reiter, S.—Lang, M.—Rigoll, G.: Feature Selection and Stacking for Robust Discrimination of Speech, Monophonic Singing, and Polyphonic Music. IEEE International Conference on Multimedia and Expo (ICME 2005), Amsterdam, Netherlands, July 2005, pp. 840–843, doi: 10.1109/ICME.2005.1521554.

[32] Abe, S.: Support Vector Machines for Pattern Classification. Springer-Verlag, 2005.

[33] Davis, J.—Goadrich, M.: The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning (ICM 2006), Pittsburgh, Pennsylvania, USA, June 2006, pp. 233–240, doi: 10.1145/1143844.1143874.

[34] Guyon, I.—Weston, J.—Barnhill, S.—Vapnik, V.: Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning, Vol. 46, 2002, No. 1-3, pp. 389–422.

[35] Anthony, G.—Gregg, H.—Tshilidzi, M.: Image Classification Using SVMs: One-Against-One vs. One-Against-All. Proccedings of the 28th Asian Conference on Remote Sensing (ACRS 2007), Vol. 2, 2007, pp. 801–806.

[36] Dhanalakshmi, P.—Palanivel, S.—Ramalingam, V.: Classification of Audio Signals Using SVM and RBFNN. Expert Systems with Applications, Vol. 36, 2009, No. 3, Part 2, pp. 6069–6075, doi: 10.1016/j.eswa.2008.06.126.

[37] Wang, J.-C.—Wang, J.-F.—He, K.-W.—Hsu, C.-S.: Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor. IEEE International Joint Conference on Neural Network (IJCNN 2006), Vancouver, Canada, July 2006, pp. 1731–1735.

[38] Zhu, Y.—Ming, Z.—Huang, Q.: SVM-Based Audio Classification for Content-Based Multimedia Retrieval. International Workshop on Multimedia Content Analysis and Mining (MCAM 2007), Weihai, China, June 2007, pp. 474–482, doi: 10.1007/978-3-540-73417-8_56.

[39] Lu, L.—Zhang, H.-J.—Li, S.-Z.: Content-Based Audio Classification and Segmentation by Using Support Vector Machines. Multimedia Systems, Vol. 8, 2003, No. 6, pp. 482–492.

[40] Kecman, V.: Learning and Soft Computing, Support Vector Machines, Neural Networks, and Fuzzy Logic Models. MIT Press, Cambridge, MA, USA, 2001.

[41] Hsu, C.-W.—Chang, C.-C.—Lin, C.-J.: A Practical Guide to Support Vector Classification. Availaible on: `http://www.csie.ntu.edu.tw/~cjlin/papers.html`.

[42] Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognition Letters, Vol. 27, 2006, No. 8, pp. 861–874, doi: 10.1016/j.patrec.2005.10.010.

[43] Chang, C.-C.—Lin, C.-J.: LibSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2, 2011, No. 3, p. 27.

[44] Pleva, M.—Juhar, J.: TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation. 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, May 2014.

[45] Vavrek, J.—Cizmar, A.—Juhar, J.: SVM Binary Decision Tree Architecture for Multiclass Audio Classification. Proceedings of ELMAR 2012, Zadar, Croatia, September 2012, pp. 202–206.

[46] Hsu, C.-W.—Lin, C.-J.: A Comparison of Methods for Multiclass Support Vector Machines. IEEE Transactions on Neural Networks, Vol. 13, 2002, No. 2, pp. 415–425.

[47] Duan, K.-B.—Keerthi, S.-S.: Which Is the Best Multiclass SVM Method? An Empirical Study. Proceedings of the 6th International Workshop on Multiple Classifier Systems (MCS 2005), Seaside, CA, USA 2005, pp. 278–285, doi: 10.1007/11494683_28.

Jozef Vavrek graduated M.Sc. (Ing.) at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice in 2010. Four years later, he received his Ph.D. degree in Telecommunications. His research is oriented on audio data classification, retrieving and digital speech and audio processing.

Peter Feciľak graduated in informatics from the Technical University of Košice in 2006. In 2009 he received his Ph.D. degree in informatics with focus on optimization of computer networks. Currently he works as Assisstant Professor at the Department of Computers and Informatics of the Technical University of Košice. He is author and co-author of more than 70 scientific papers. His current teaching and research interests include computer networks, network monitoring, quality of services and smart energy systems.

Jozef Juhár graduated from the Technical University of Košice in 1980. He received his Ph.D. degree in radioelectronics from Technical University of Košice in 1991, where he works as Full Professor and Head of the Department of Electronics and Multimedia Communications. He is author and co-author of more than 300 scientific papers. His research interests include digital speech and audio processing, speech recognition and synthesis, speaker identification and development of spoken dialogue human-computer interactive systems.

Anton Čižmár graduated from the Slovak Technical University in Bratislava in 1980, at the Department of Telecommunications. He received his Ph.D. degree in radioelectronics from the Technical University of Košice in 1986, where he works as Full Professor at the Department of Electronics and Multimedia Communications. He is author and co-author of more than 170 scientific papers. His research areas are broadband information and telecommunication technologies, multimedia systems, telecommunication networks and services and mobile communication systems.