THE SVM BINARY TREE CLASSIFICATION USING MRMR AND F-SCORE FEATURE SELECTION ALGORITHMS

Jozef VAVREK, Jozef JUHÁR, Anton ČIŽMÁR
Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Park Komenského 13, 042 20 Košice, Slovak Republic, e-mail: {jozef.vavrek,jozef.juhar,anton.cizmar}@tuke.sk

ABSTRACT

The discrimination between various types of speech and non-speech signals in audio data stream is the fundamental step for further indexing and retrieving. This paper considers some of the basic problems in audio content classification which is the key component in automatic audio retrieval system. It illustrates a potential use of statistical learning algorithm called support vector machine (SVM) for broadcast news (BN) audio classification task. The overall classification architecture uses binary tree SVM (BT-SVM) decision scheme in combination with well known audio features such as, MFCCs and low level MPEG-7 audio descriptors. The important step in creating such classification system is to define the optimal features for each binary SVM classifier. There exist various feature selection algorithms that help to create such feature set. Therefore we decided to implement F-score and Minimum Redundancy Maximum Relevance (MRMR) feature selection algorithms, as an effective search algorithms used in many pattern recognition tasks.

Keywords: *Support Vector Machines, F-score, MRMR, MPEG-7 Audio Descriptors*

1. INTRODUCTION

Growing number of audio databases with vast amount of audio data demands for efficient organization and manipulation of this data. Such processing is desirable for applications requiring accurate discrimination of speech and non-speech segments, for instance automatic transcription of broadcast news (BN), speech and speaker recognition, retrieving of audio queries, and so forth. Audio data of BN contain alternating sections of different types of speech and music. Thus, fundamental step in audio stream processing is to automatically classify audio content into appropriate audio classes. We call this separation criterion as audio content classification. Process of classification is often carried out along with the process of audio stream segmentation. These processes are substantial in whole retrieving system and are very useful in many classification task. The overall classification performance is conditioned by the process of feature extraction.

This paper presents possible solution for audio stream classification, utilizing binary tree discrimination technique based on SVM classifier and two effective feature selection algorithms, used for processing and retrieving of BN audio data. Audio content of BN media contains not only single type of acoustic event (pure speech, music), but also mixed sounds (speech and music with background noise). *Pure speech* is one of the most occurred in BN audio stream. Single anchor reports in studio and field speech in a quiet environment represent this class. *Speech with ambient noise* represents events like field speech and telephone conversations in noisy environment. It also contains all types of acoustic events with occurrence in environment, e.g. sound from machines, birds, water, wind, crowd, etc. - *ambient noise*. Jingles at the beginning and at the end of news with anchor speech belong to the *speech with music background* class. Individual jingles and music in commercials represent *pure music* class. BN audio stream also contains silent intervals between different speakers and jingles, defined as *silence*.

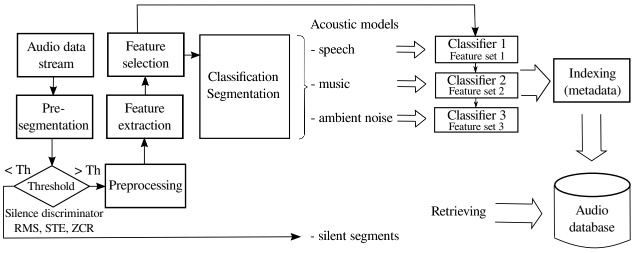

Fig. 1 Automated classification system for broadcast news audio

The most important task in content-based audio classification is to define speech and non-speech audio segments. Zhang and Kuo used hidden Markov models (HMMs) approach, utilizing various types of feature extraction methods, to distinguish speech, music and ambient noise [1]. Proposed approach detected abrupt changes in audio stream by measuring actual values of selected audio features. Consequently, boundary decision points of observed audio classes were set and trained by using HMM classifier. In addition, Han, Gao and Ji focused on the application of Support Vector Machines (SVMs) method for classification and detection of audio signal by new proposed method, namely selective ensemble SVMs with much more used features [2]. Other works are aimed at the design of such complex systems that are able to process broadcast news audio signals, in terms of segmentation, classification, indexing and retrieval [3], [4].

Some authors have developed new and efficient features and segmentation algorithms that can capture various changes in audio stream, thus improve classification accuracy of audio data [5], [6]. Others have developed classifier-dependent (wrapper and embedded), and classifier-independent (filter) feature selection algorithms in order to find optimal feature set for various audio classification problems [7], [8].

Fundamental architecture of the most common used classification system, designed for retrieving and indexing of all audio classes in BN audio stream, is illustrated in Fig. 1.

This paper is organized as follows: The used database is described in section II. Section III deals with segmentation and feature extraction. Used feature selection algorithms are described in Section IV. Section V provides description of the used SVM binary tree topology and section VI discusses realized experiments and obtained results. Finally, the section VII gives our conclusions and shows future directions.

2. DATABASE DESCRIPTION

The Slovak TV broadcast news audio database KEMT-BN1 was used to evaluate the efficiency of proposed classification system [9]. First portion of the database contains the Slovak part of the COST-278 database [10]. It comprises 3 hours of TA3 channel news grabbed from the analogue broadcast stream. The rest of database consists of the STV channel evening and noon news. Broadcast stream was stored using Technisat AirStar 2 PCI internal DVB-T card in MPEG2 format. The audio files were demuxed from grabbed video stream in 48 *kHz* 192 *kbit* stereo MPEG Layer 2 audio format. MPEG files were converted into WAV files by using PCM 16 *kHz* 16 *bit* mono format and used for further process of annotation. Individual audio recordings were annotated using Transcriber[1] tool in XML file format (*.trs) and by the NIST Sclite scoring toolkit[2] in a simple text file format exported from Transcriber (*.stm). Database consists of 188 recordings in total duration 65 hours. One can get 55 hours of data by extracting only speech frames. Clean speech (planned speech, studio environment, no background noise) represents 21 hours and 47 minutes from this database.

3. SEGMENTATION AND FEATURE EXTRACTION

Segmentation of audio data is the task of dividing a continuous audio stream into audio segments. Similar audio segments contain the same type of acoustic signal (speech, music or ambient noise).

Current research in the area of audio signal processing focuses on development an alternative segmentation methods. For example, Huang and Hansen have proposed a novel segmentation and classification algorithm *CompSeg* [11]. It combines three distance metric techniques, namely Bayesian information criterion (BIC), T^2 distance and weighted mean distance (WMD). Siegler at al. used the Symmetric Kullback-Leibler distance for speaker segmentation [12]. Proposed method is effective only for segments with duration longer than 5 *s*. Lin at al. [13] have developed an STMR algorithm for identification dissimilarity boundaries between different speakers, which uses the principles of an unsupervised segmentation by the SVM.

All mentioned works and many others are based on supervised and unsupervised audio change detection [14]. Therefore segmentation ability strongly depends on the assumption, that the acoustic data are composed from different speakers who are either known a priori or unknown.

3.1. Pre-segmentation

Process of pre-segmentation fulfils the task of dividing audio stream into segments with equal length, usually by rectangular window. Time domain statistical parameters such as mean, variance, Root Mean Square (RMS) of Short-time Energy and also Zero-Crossing Rate (ZCR) are computed within these segments. Typical length of one segment is 1 *s*. Pre-segmentation is followed by frame-based segmentation, where each segment is further divided into overlapped frames, using Hamming window, in order to avoid spectral distortions.

3.2. Feature Extraction

Searching for relevant and felicitous representation of audio content is recently most demanded task. Audio content analysis in time, frequency and cepstral domain is therefore inevitable. Vast amount of feature extraction methods have been proposed in order to find appropriate representation of audio signal under different acoustic conditions. Choosing the best one is therefore crucial. Following text describes the basic descriptors and features used in our experiments. These descriptors were chosen according to the sufficient performance in multi-class classification problem, presented in [15], [16].

Zero-Crossing Rate: ZCR of the frame is defined as the number of times the audio waveform changes from positive to negative within the duration of the frame. Zero crossing rate is used as a measure of noisiness and rough estimation of the fundamental frequency of voiced signals. ZCR of unvoiced sounds are usually larger than ZCR of speech signals.

Audio Spectrum Centroid: ASC [17] descriptor gives the information about the shape of the power spectrum. It indicates whether low or high frequencies are dominated in a power spectrum and can be regarded as an approximation of the perceptual sharpness of the signal.

Audio Spectrum Spread: ASS [17] is a measure of the spectral shape and represents the second central moment of the log-frequency spectrum. ASS also gives the information about how the spectrum is distributed around its centroid. A low ASS value means that the spectrum may be concentrated around the centroid, whereas a high value reflects a distribution of a power across a wider range of frequencies.

Audio Spectrum Flatness: ASF [17] reflects the flatness properties of the power spectrum. It can be defined as the deviation of the signal's spectrum from a flat spectrum. Instead of calculating one flatness value for the whole spectrum, a separation in frequency bands is performed, resulting in one vector of flatness values per frame. High values of ASF coefficients reflect noisiness, on the other hand, low values indicate a harmonic structure of the spectrum.

[1] http://trans.sourceforge.net

[2] http://www.itl.nist.gov/iad/mig/tools

Spectral Roll-off: ROF [17] point is defined as the frequency below which 85% of the magnitude distribution of the spectrum is concentrated. It is also a measure of the spectral shape and yields higher values for high frequencies. ROF is used to distinguish voiced speech from unvoiced music, which has a higher roll-off point, because music power is better distributed over the subband range.

Mel-Frequency Cepstral coefficients: MFC analysis has been a popular signal representation method used in many audio classification tasks, especially in speech recognition systems [18]. The basis for the mel-frequency scale is derived from the human perceptual system. Obtaining the MFCCs involve processing of the acoustic signal according to the following steps:

1. Divide the signal into frames and apply the Hamming window function.
2. Get the amplitude spectrum of each frame.
3. Take the log of these spectrums.
4. Convert to mel filter bank.
5. Apply the Discrete Cosine Transform (DCT).

4. FEATURE SELECTION ALGORITHMS

One of the possible way how to minimize the overall classification error is to identify the most characterizing features of the observed data. Number of methods and algorithms have been developed for selecting an optimal subset of features that can help to achieve the highest classification efficiency of the system. Most of them are filter based and the process of defining the sufficient number of features is based on statistical dependencies and distance measure techniques. Therefore, we implemented *F-score* and *Minimum Redundancy Maximum Relevance* (MRMR) feature selection algorithms in order to reduce the overall classifiaction error of our system.

4.1. F-score

F-score [19] is a simple algorithm that measures a degree of separability between two sets of training data. Given training vectors $\mathbf{x}_k, k = 1, ..., m$, if the number of positive and negative instances are n_+ and n_-, respectively, then F-score of the i-th feature is defined as:

$$F(x_i) \equiv \frac{\left(\bar{\mathbf{x}}_i^{(+)} - \bar{\mathbf{x}}_i\right)^2 + \left(\bar{\mathbf{x}}_i^{(-)} - \bar{\mathbf{x}}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{\mathbf{x}}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{\mathbf{x}}_i^{(-)}\right)^2}, \quad (1)$$

where $\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_i^{(+)}, \bar{\mathbf{x}}_i^{(-)}$ represent the mean of the i-th feature within all positive, and negative instances, $x_{k,i}^{(+)}$ represents the i-th feature of the k-th positive instance, and $x_{k,i}^{(-)}$ is the i-th feature of the k-th negative instance. The numerator indicates separability rate between the positive and negative sets, and denominator indicates the one within each of two sets. Higher value of F-score indicates significant discriminative power of feature, in terms of an effective feature selection criterion in combination with the SVM classifier. Therefore, the procedure of finding the optimal feature set firstly selects features with the highest F-score and subsequently implements the SVM for training/testing. The procedure is summarized below:

1. *Calculate F-score for each component in one training vector.*
2. *Select the subset of features with the highest F-scores. Selected feature size is equal to the first half of the feature set with the highest F-score values.*
3. *For each subset, do the following:*
   a) *Randomly split the training data into X_{train} and X_{valid}.*
   b) *Let X_{train} be the new training data. Use the SVM procedure with cross-validation, to obtain SVM parameters and a predictor; use the predictor to predict X_{valid}.*
   c) *Repeat steps a) and b) ν-times, where ν represents ν-fold cross-validation, and then calculate the average validation accuracy.*
4. *If selected feature subset contains greater than or equal to $2^n + 1$ features, for $n = 0, 1$, then go to step 1., otherwise terminate algorithm.*

4.2. MIN-Redundancy and MAX-Relevance

MRMR feature selection algorithm [20] is based on the mutual information difference criterion. If the features of an example are randomly or uniformly distributed in different classes, their mutual information is equal to zero. Otherwise, they should have large mutual information in case of differentially expressed features within different classes. The mutual information of two features $x_i, i = 1, 2$ is defined by their probabilistic density functions $p(x_1)$, $p(x_2)$, and $p(x_1, x_2)$:

$$I(x_1; x_2) = \iint p(x_1, x_2) log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} dx_1 dx_2. \quad (2)$$

Thus the mutual information is used as a measure of relevance. In *MAX-Relevance*, the selected features $x_i, i = 1, ..., S$ are computed individually, in order to have the largest mutual information $I(x_i; c)$ within the target class c, reflecting the largest dependency on the target class. It is defined as mean value of all mutual information values between individual features x_i and class c:

$$maxD(S, c), \quad D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i; c). \quad (3)$$

The class-discriminative power of two highly dependent features will not change much if we remove one of them. Therefore, the main idea of *MIN-Redundancy* is to select the features that are mutually maximally dissimilar, which can be expressed in the following form:

$$maxR(S), \quad R = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i, x_j). \quad (4)$$

The MRMR feature set is obtained by the criterion combining the above two constraints simultaneously:

$$max\Phi(D,R), \quad \Phi = D - R. \tag{5}$$

The final near-optimal feature set is then acquired by incremental search algorithm that optimizes the following condition:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j;c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j;x_i) \right], \tag{6}$$

where $X = \{x_i, i = 1, ..., M\}$ represents input data and S_{m-1} defines feature set with $m-1$ candidate features. The whole process of selecting the optimal number of features n of the candidate set includes the following steps:

1. *Compute incremental selection (6) to select n features from the input X. This leads to n feature sets $S_1 \subset S_2 \subset ...S_{n-1} \subset S_n$.*
2. *Select four subsets of features obtained from step 1., with the feature size S_n; first half between S_n and $S_{n-n/2}$; first half of S_n; and first half between $S_{n/2}$ and S_1. (for $n = 36$ we get S_{36}, S_{27}, S_{18} and S_9)*
3. *Compute cross-validation for each selected subset according to the step 3. defined in the F-score selection algorithm and find the highest classification accuracy.*
4. *Repeat step 2. and 3. for two subset with the highest classification accuracy until the final subset of features S_{n*} is obtained.*

The MRMR algorithm performs well on microarray gene data [21] and abnormal acoustic events occurred in public places, such that gun shot, explosions and breaking glass [22], as well.

5. SVM BINARY TREE

Support vector machine classifier [23] represents a binary discrimination tool, primary designed for separation of two classes. For instance classes with different time, frequency and space characteristics. Currently are SMVs involved in solving many audio classification tasks due to their generalization ability and superior performance in various pattern classification tasks. Moreover, SVMs successfully outperform other types of classifiers in specific audio classification tasks, when lack of sufficient training audio data is significant [24]. The SVM classifier is originally designed for solving binary classification problem and discrimination of multiple classes is realized by combining several binary classifiers. Therefore, we try to create an effective classification binary tree scheme, in order to increase total performance of the SVM technique. In this sense, SVM represents specific type of discriminative function that model class boundary or margin. Discriminative function is modeled by linear separating hyperplane with maximal or soft margin. If the training data from different classes cannot be linearly separated in the original input space, the kernel functions non-linearly transform the original input space into the high-dimensional *feature space*.

Table 1 Kernel functions

Kernel function	Label
$K(\mathbf{x}_i,\mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$	Linear
$K(\mathbf{x}_i,\mathbf{x}_j) = \left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1 \right)^d$	Polynomial of degree d
$K(\mathbf{x}_i,\mathbf{x}_j) = exp(-\gamma \lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert^2)$	Gaussian radial basis function

This transformation can be achieved by various non-linear mappings, such as: linear, polynomial and radial basis symmetric function (RBF). The resulting hyperplane will be optimal in the sense of being a maximal margin discriminative function with respect to training data.

Training data are represented in the form of N-dimensional vectors:

$$(x_1, y_1), ..., (x_l, y_l) \in X \times \{\pm 1\}, \tag{7}$$

where X is some non-empty set of patterns x_i (sometimes called cases, inputs, instances, samples) and labels y_i. Among all the hyperplanes that minimize the training error, classifier have to find the one with the largest margin in the form:

$$d(\mathbf{w},\mathbf{x},b) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = \sum_{i=1}^{l} w_i x_i + b, \tag{8}$$

where l represents all the training data, $\mathbf{w}, \mathbf{x} \in R^n$ and the scalar b is called *bias*. After successful training stage, using obtained weights, the learning machine produces output o, according to an *indication function*, given as:

$$i_F = o = sign(d(\mathbf{w},\mathbf{x},b)), \tag{9}$$

where o is the standard notation for the output from a learning machine.

One basic idea in designing non-linear SVMs is to map input vectors $\mathbf{x} \in R^n$ into the high-dimensional feature space by *kernel functions* $K(\mathbf{x}_i, \mathbf{x}_j)$. The mostly used kernel functions are shown in Tab. 1.

Basic concept of the binary tree architecture is depicted in Fig. 2. Each node represents one binary SVM classifier that realizes separation of two classes +1 and -1. It follows that the multi-class classification needs to train maximally $(K-1)$ SVMs for K-class problem. Therefore, it is more efficient in computation then one-against-one and one-against-rest methods. Classification procedure of the SVM-BT (SVM-Binary Tree) is based on a *coarse-to-fine* strategy that allows us to make a discrimination on coarse classes at the top of the decision tree architecture. 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 one of leaves that represent the fine-gained class is obtained.

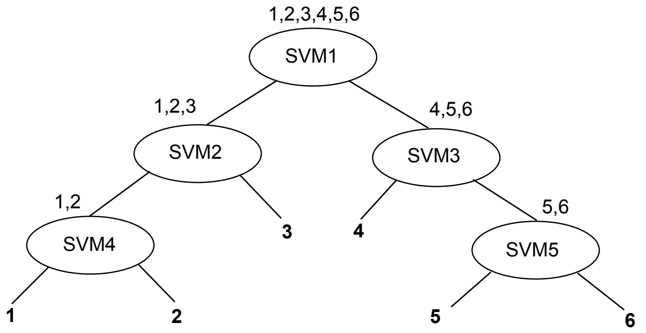

Fig. 2 SVM binary tree architecture

The main advantage of SVM-BT lies in choosing feasible feature set for each SVM binary classifier that separate two different classes. The feature selection is done by MRMR and F-score, as it was presented in paper.

6. EXPERIMENTS AND RESULTS

All the evaluations within the training phase were based on the assumption that the SVM needs only small set of data in order to preserve generalization ability and to avoid the problem of overfitting. At first, we divided each audio recording into non-overlapping 1 *s* long segments by a rectangular window and then time domain features, such as RMS of short time energy and ZCR, were extracted within each segment. Segments were further divided into 40 audio frames, each with 50 *ms* duration and 25 *ms* overlap. This phase of pre-segmenting helped us to remove silent segments by defining threshold of the average short-time energy and ZCR.

Parameterization of the training data was performed by using a set of frame-based audio features extracted from 36 recordings. Consequently, the subsets of features with the same number of frames per each class were selected in random order. It helped us to reduce the huge amount of audio data from highly occurred classes, such as pure speech and background speech, and to prevent the problem of overfitting during the training phase. The final training set contained about 104 minutes of audio data. Testing phase was performed by using 4 recordings, each of which contained TV news, different from those used in training phase. The overall duration of a test set was 79 minutes. Training and testing were evaluated by frame-based features with total dimension 36. MRMR and F-score feature selection algorithms were used for finding an optimal feature set for training data. Found dimension of each feature vector was than used for testing data. Final classification accuracy was evaluated by using testing set only. Tab. 2 presents a summary of used features with corresponding dimensions.

Classification of parameterized audio data was realized by the SVM-BT architecture, listed in Fig. 3. Feature selection algorithms, mentioned in Sec. 4, were used for each binary SVM classifier in order to achieve the best feature set with optimal dimension, listed in Tab. 3. Evaluating of the F-score search algorithm was performed for features with dimensions 36, 18, 9, 4, and 2.

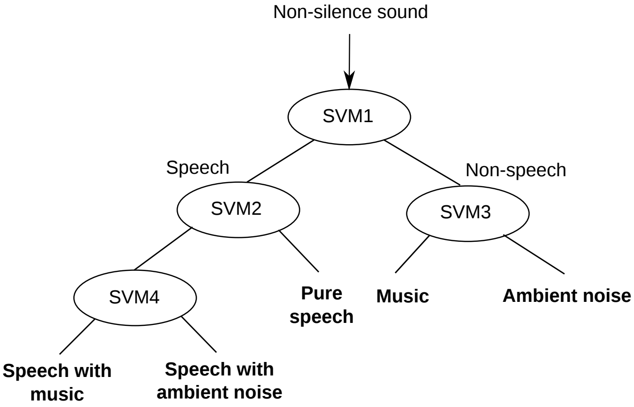

Fig. 3 Adapted SVM binary tree architecture

The optimal feature dimension defined by MRMR algorithm was set to 36, 34, 32, 27, 18, and 9. We used radial basis kernel and 5-fold cross-validation as evaluation functions for each SVM model, in order to get the final F-measure and Accuracy of selected training data. F-measure evaluation parameter was computed as follows:

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{10}$$

where *Precision* represents the proportion of the true positives against all the positive results and *Recall* shows the proportion of the true positives against the positive and false negative results. *Accuracy* was evaluated as the ratio of correctly classified frames to all predicted frames of each audio class. The overall performance of the SVM-BT architecture in comparison with SVM one-against-one multi-class scheme (OAO), which constructs an SVM for each pair of classes and is originally implemented in LIBSVM software[3], is shown in Tab. 4. Testing accuracy of the SVM-BT architecture strongly depends on the discriminative ability of speech/non-speech model. Therefore we implemented testing algorithm that returns only correctly classified frames so that each model predicts only testing data from the corresponding classes. It can help to prevent the misclassification that propagates from the upper level to the lower levels of the architecture. All the experiments were performed by LIBSVM software and available tools, which cooperate with this software.

7. CONCLUSION AND FUTURE WORK

The overall performance of BN audio classification system using SVM-BT architecture and feature selection algorithms was evaluated in this paper. The main goal of this work was to assess the classification ability of proposed SVM-BT scheme in comparison with One-Against-One (SVM-OAO) multi-class classification procedure by using only small set of training data. The effectiveness of F-score and MRMR feature selection algorithms was evaluated as well. Therefore the adapted SVM-BT architecture employing the coarse-to-fine strategy and filter-based search algorithms was built in order to solve the multi-class audio classification problem.

[3] http://www.csie.ntu.edu.tw/ cjlin/libsvm

Table 2 Used features

Feature (descriptor)	*MFCCs*	*ZCR*	*ASF*	*ROF*	*ASC*	*ASS*
Dimension	13	1	19	1	1	1

Table 3 Selected features for training data

Method	**SVM model**	**Selected dim.**	**F-measure [%]**	**Acc [%]**
F-score	SVM1, C=8, g=1	36	93.4	93.31
	SVM2, C=64, g=0.25	36	76.32	76.17
	SVM3, C=8, g=4	36	49.48	98.3
	SVM4, C=64, g=0.25	36	50.16	85.8
	SVM-OAO, C=8, g=1	36	57.32	78.7
MRMR	SVM1, C=8, g=1	34	93.2	93.3
	SVM2, C=64, g=0.25	34	76.5	76.4
	SVM3, C=64, g=4	34	50.12	98.11
	SVM4, C=8, g=1	34	52.32	85.3
	SVM-OAO, C=8, g=4	34	44.15	75.88

Table 4 The overall prediction accuracy of the SVM-BT architecture

Accuracy [%]	**Pure speech**	**Speech & music**	**Speech & ambient noise**	**Music**	**Ambient noise**	**Avg Acc[%]**
F-score						
SVM-OAO	98.31	57.19	53.45	84.23	79.05	**74.44**
SVM-BT	93.47	43.17	45.61	84.89	56.19	**64.66**
MRMR						
SVM-OAO	93.21	1.3	4.95	1.82	0.88	**20.43**
SVM-BT	98.24	10.4	31.92	13.27	8.32	**32.43**

The main reason we employed search algorithms was to find the optimal combination of feature vectors with the maximal possible reduction of their dimensions. Reduction of feature dimension is important step in classification process because of saving the computational complexity and training time. Tab. 3 shows that the higher classification performance of each model was achieved with dimensions 34 and 36. Thus, F-score feature selection algorithm did not accomplish dimension reduction of each feature vector but helped to find an optimal order of each feature element. Experimental results from Tab. 4 show that the use of MRMR algorithm caused rapid decrease in classification accuracy of predicted data for both architectures, namely SVM-OAO and SVM-BT. Therefore, the discrimination ability of the MRMR feature selection algorithm is not suitable for solving the broadcast news audio classification problem. On the contrary, the classification accuracy of SVM-OAO and SVM-BT architectures was much higher in case of the F-score feature selection algorithm. It follows that the SVM classifier is in combination with the F-score feature selection algorithm more sufficient for solving multi-class audio classification problem with restricted amount of training data and does not suffer from overfitting.

The average classification accuracy of proposed SVM-BT architecture achieved higher values than SVM-OAO approach only in case of MRMR feature selection algorithm. SVM-BT approach did not overcome the SVM-OAO approach in case of using F-score feature selection algorithm. Almost 10% reduction of the average classification accuracy was observed in this case. This reduction was caused by the misclassification that propagated from the upper level in the SVM-BT architecture. It explains relatively low values of predicted accuracy for all classes.

Future work will be directed to the design of the SVM-BT architecture using speaker change detection algorithms mentioned in Sec. 3. We will also focus on improving classification ability of SVM-BT architecture by developing classification scheme using misclassification error reduction techniques.

ACKNOWLEDGEMENT

The research presented in this paper was supported by Research and Development Operational Program funded by the ERDF under project ITMS-26220220155 (50%) and ITMS-26220220182 (50%).

REFERENCES

[1] ZHANG, T. – KUO, C. C. : "Hierarchical classification of audio data for archiving and retrieving," *ICASSP '99 Proceedings of the Acoustic, Speech, and Signal Processing*, vol. 06, pp. 3001–3004, 1999.

[2] B. HANG, B. – GAO, X. – JI, H.: "Automatic news audio classification based on selective ensemble SVMs," *Lecture Notes in Computer Science*, vol. 3497, no. II, pp. 363–368, 2005.

[3] ZHANG, T. – JAY KUO, C. C.: "Audio content analysis for online audiovisual data segmentation and classification," in *IEEE Transactions on Speech and Audio Processing*, May 2001, vol. 9, pp. 441–457.

[4] van DINTHER, R. – McKINNEY, M. F. – Li, H. R.: "Real-time segmentation of radio broadcast content in radio devices," *IEEE International Conference on Consumer Electronics*, 2009, art.nr. 5012179.

[5] SCHEIRER, E. – SLANEY, M.: "Construction and evaluation of a robust multifeature speech/music discriminator," in *in Proc. of ICASSP*, vol. 2, 1997, pp. 1331–1334.

[6] M. RAMONA, M. – GICHARD, G.:"Comparison of different strategies for a SVM-based audio segmentation," *17th European Signal Processing Conference (EUSIPCO 2009)*, pp. 20–24, 2009.

[7] GUYON, I. – GUNN, S. – NIKRAVESH, M. – ZADEH, L. – Eds.: *Feature Extraction, Foundations and Applications*. Springer, 2006.

[8] WESTON, J. – MUKHERJEE, S. – CHAPELLE, O. – PONTIL, M. – POGGIO, T. – VAPNIK, V.: "Feature selection for svms," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2000, pp. 668–674.

[9] PLEVA, M. – JUHÁR, J. – ČIŽMÁR, A.:"Slovak broadcast news speech corpus for automatic speech recognition," in *Proc. of RTT '07*, 2007.

[10] VANDECATSEYE, A.: "The COST278 pan-european broadcast news database," *Proc. of LREC'04*, pp. 873–876, 2004.

[11] HUANG, R. – HANSEN, J. H. L.: “Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 907–919, May 2010.

[12] SIEGLER, M. A. – JAIN, U. – RAJ, B. – STERN, R. M.: “Automatic segmentation, classification and clustering of broadcast news audio,” in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.

[13] LIN, P. C. – WANG, J. C. – WANG, J. F. – SUNG, H. C.: “Unsupervised speaker change detection using SVM training misclassification rate,” *IEEE Transactions on Computers*, vol. 56, pp. 1234–1244, 2007.

[14] HUANG, C.-C. – WANG, J.-F. – WU, D. J.: “Automatic scene change detection for composed speech and music sound under low SNR noisy environment,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 13, pp. 689–699, 2005.

[15] XIE, L. – FU, Z. H. – FENG, W. – LUO, Y.: “Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news,” *Multimedia Systems*, vol. 17, pp. 101–112, 2011.

[16] VOZARIKOVA, E. – JUHAR, J. – CIZMAR, A.: “Acoustic events detection using MFCC and MPEG-7 descriptors,” in *Multimedia Communications, Services and Security*, ser. Communications in Computer and Information Science, 2011, vol. 149 CCIS, pp. 191–197.

[17] KIM, H. G. – MOREAU, N. – SIKORA, T.: *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005.

[18] ON, C. K. – PANDIYAN, P. M.: “Mel-frequency cepstral coefficient analysis in speech recognition,” *Computing & Informatics 2006, ICOCI’06*, no. 2, pp. 2–6, 2006.

[19] CHEN, Y. W. – LIN, C. J.: “Combining SVMs with various feature selection strategies,” *Feature Extraction:Studies in Fuzziness and Soft Computing*, vol. 207, pp. 315–324, 2006.

[20] PENG, H. – LONG. F. – DING, C. : “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, 2005.

[21] DING, C. – PENG, H.: “Minimum redundancy feature selection from microarray gene expression data,” in *J Bioinform Comput Biol*, 2003, pp. 523–529.

[22] VOZARIKOVA, E. – LOJKA, M. – JUHAR, J. – CIZMAR, A.: “Performance of basic spectral descriptors and mrmr algorithm to the detection of acoustic events,” in *Multimedia Communications, Services and Security*, ser. Communications in Computer and Information Science, 2012, vol. 287, pp. 350–359.

[23] ABE, S.: *Support Vector Machines for Pattern Classification*, ser. Advances in Pattern Recognition. Springer-Verlang, 2005.

[24] LU, L. – ZHANG, H. J. – Li, S. Z.: “Content-based audio classification and segmentation by using support vector machines,” *Multimedia Systems*, vol. 8, pp. 482–492, 2003.

BIOGRAPHY

Jozef Vavrek was born in Kosice, Slovakia in 1985. In 2010 he 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. He is a Ph.D. student at the same department in the field of Telecommunications. His research is oriented on audio data classification, retrieving and digital speech and audio processing.

Jozef Juhár was born in Poproč, Slovakia in 1956. He graduated from the Technical University of Košice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Košice in 1991, where he works as a full professor and head of the Department of Electronics and Multimedia Communications. He is author and co-author of more than 200 scientific papers. His research interests include digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks.

Anton Čižmár was born in Michalovce, Slovakia in 1956. He graduated from the Slovak Technical University in Bratislava in 1980, at the Department of Telecommunications. He holds a Ph.D. degree in Radioelectronics from the Technical University of Košice in 1986, where he works as a Full Professor at the Department of Electronics and Multimedia Communications. Now he works as the rector of the Technical University of Košice. He is author and co-author of more than 170 scientific papers. His scientific research areas are broadband information and telecommunication technologies, multimedia systems, telecommunication networks and services, 4. generation mobile communication systems and localization algorithms.