Martin Hric – Michal Chmulík – Roman Jarina *

COMPARISON OF SELECTED CLASSIFICATION METHODS IN AUTOMATIC SPEAKER IDENTIFICATION

This paper presents performance comparison of three different classifiers applied in Automatic SpeakeR Identification: Gaussian Mixture Model (GMM), k Nearest Neighbor algorithm (kNN) and Support Vector Machines (SVM). Each classifier represents different approach to the classification procedure. Mel Frequency Cepstral Coefficients (MFCC) were used as feature vectors in the experiment. Classification precision for each classifier was evaluated on frame and recording level. Experiments were conducted over dataset MobilDat-SK, which was recorded in mobile telecommunication network. Experiment shows promising results for SVM classifier.

Keywords: *kNN, SVM, GMM, MFCC, speaker identification.*

1. Introduction

Nowadays, speaker recognition is one of the basic tasks in various systems for Automatic SpeakeR Identification (ASRI), audio documents retrieval, forensic analysis, etc. Such systems allow recognizing "who is talking" from the speech signal. Identification system consists of various parts working together. In this paper, we deal with three different classification approaches for ASRI system, namely Gaussian Mixture Model (GMM), k Nearest Neighbor (kNN) and Support Vector Machines (SVM). Precision of the classifiers is experimentally evaluated by tests performed on the same dataset. We also focus on ability of the selected classifiers to be trained from limited amount of speech data. Such property is crucial in applications as speaker segmentation and matching in audio stream, or speaker retrieval in digital audio archives using Query-by-Example approach.

The paper is organized as follows. In section 2 each of used classification method is briefly discussed. Section 3 presents results of classification as well as database description, data preparation and parameters of given classifiers.

2. Classification techniques description

In this section, we present three different classification methods. Subsections give a short overview of GMM, kNN and SVM classifiers. The GMM is a typical classification method, which has been successfully used in many applications related to the speech. The SVM method becomes very popular in the present time due to its great classification abilities although it is computational very expensive method. Unlike the model based classification methods as GMM and SVM, kNN represents instance based approach to the classification process. From the set of other available classification methods, the HMM or decision trees can be mentioned.

2.1. GMM classification

In GMM classification, Gaussian mixture model is used for statistical representation of speaker pattern. The distribution of feature vectors, extracted from a speech signal, is modeled by a mixture of Gaussian density functions (Fig.1). For a *D*-dimensional feature vector x, the mixture density for speaker λ_r is defined as [1]:

$$p(x|\lambda_r) = \sum_{i=1}^{M} p_i^r b_i^r(x), \tag{1}$$

where M denotes number of components and p_i^r are mixture weights. Density is weighted linear combination of M component uni-modal Gaussian densities $b_i^r(x)$:

$$b_i^r(x) = \frac{1}{(2\pi)^{D/2}\left|\sum_i^r\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i^r)\left(\sum_i^r\right)^{-1}(x-\mu_i^r)\right\}, \tag{2}$$

each one parameterized by mean vector μ_i^r and covariance matrix Σ_i^r. Mixture weights must satisfy the following constraint:

$$\sum_{i=1}^{M} p_i^r = 1\,. \tag{3}$$

* **Martin Hric, Michal Chmulík, Roman Jarina**
Department of Telecommunication and Multimedia, Faculty of Electrical Engineering, University of Zilina, Slovakia,
E-mail: martin.hric@fel.uniza.sk

Complete GMM is defined by mean vector, covariance matrix and mixture weights (4).

$$\lambda = \left\{ p_i^r, \mu_i^r, \sum_i^r \right\} \tag{4}$$

Every speaker, who should be recognized, has his own model that is used as his representation instead of utterances in identification procedure.

In computation of covariance matrix, we utilized diagonal covariance matrix, which usually gives better results in recognition compared to full covariance matrix. The best results in parameter estimation were achieved by using the iterative Expectation Maximization (EM) algorithm [1], [2]. In this work, we used 100 iteration steps for estimation of the model.

The identification assignment is maximum likelihood classifier. Main task of the system is to make a decision if input utterance belongs to one of the set of speakers, which are represented by its models λ_1, ..., λ_r, index r denotes number of speakers. This decision is based on computation of maximum posterior probability for input feature vector [1]. NETLAB [13] implementation of GMM classifier is applied in the experimental part.

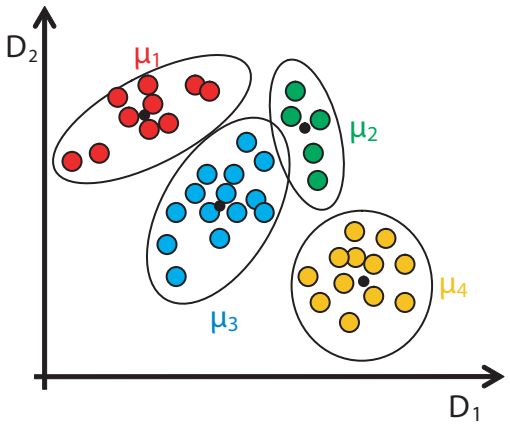

Fig. 1 Example of modeling 2-dimensional data using 4-Gaussian mixtures

2.2. kNN classification

The kNN algorithm (k Nearest Neighbor) can be classed as a nonlinear nonparametric classification method [3]. This algorithm is based on a very simple principle that similar data are close to each other in the searching or data space. In other words, the kNN finds for every object from test data set of k objects in the training data that are closest to the test object (nearest neighbors). The label assignment is usually based on the rule of majority voting, e.g. the most frequent class from the k nearest neighbors for given test object determines the class where this object should belong. A value of k dictates a number of closest objects from training data that are taken into account at the label assignment. If the value is too small, then the result can be sensitive to noise points (objects). If it is too large, then the neighborhood may include too many points from other classes.

Example of k-value impact to classification result is depicted in Fig. 2, where kNN algorithm classifies two dimensional data into two classes. First circle represents region with three neighbors taking into account at decision, where the orange point belongs ($k = 3$). In this example, the classified point belongs to the "red" class. But in the case that six neighbors are considered ($k = 6$) at label assignment, classification result is opposite and unknown point belongs to the "blue" class.

Besides a k-value, the distance metric is important to the kNN algorithm. As can be clearly seen, the distance metric represents the measure of data similarity. The choice of particular distance metric usually depends on the given classification problem. Euclidian (5) or Mahalanobis (6) distance measure are commonly used [3] and the distance between training data vector z and testing vector x are defined as follows:

$$d(x,z) = \sqrt{\sum_{k=1}^{n}(x_k - z_k)^2}\,, \tag{5}$$

$$d(x,z) = \sqrt{\sum_{k=1}^{n}(x_k - z_k)\cdot R^{-1}\cdot(x_k - z_k)'}\,, \tag{6}$$

where n is number of attributes (dimension) and R is the covariance matrix.

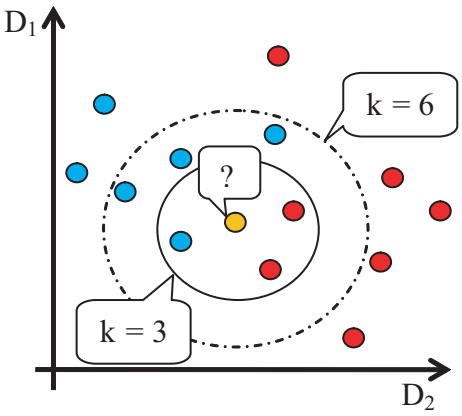

Fig. 2 Example of kNN classification

Regardless the simplicity of kNN, this method is well suitable for multi-modal classes, very flexible and belongs to top 10 data mining algorithms (IEEE Conference on data mining 2007 [3]).

2.3. SVM classification

SVM is a learning procedure based on Vapnik's statistical learning theory [4] proposed in 1979. Classification task includes

a separation of data into two sets – first set consists of data for training process and the second one for testing procedure.

Training set instance-label pairs (x_i, y_i), $i = 1, 2, ..., l$ where $x_i \in R_n$ and $y \in \{1, -1\}$l, the SVM requires the solution of the following optimization problem defined as [5]:

$$\min_{w,b,\xi} \quad \frac{1}{2} w^T w + c \sum_{i=1}^{l} \xi_i , \tag{7}$$

subject to:

$$y_i \left(w^T \varphi(x_i) + b \right) \geq 1 - \xi_i , \xi_i \geq 0 . \tag{8}$$

Each instance in the training set contains features of observed data and class label identifying particular class – in our task it is the index of speaker. The term specified in the following equation

$$K(x_i, x_j) \equiv \varphi(x_i)^T \varphi(x_j) , \tag{9}$$

denotes the kernel function. Training vectors are mapped into higher dimensional feature space by the kernel function. Example of using the kernel function is depicted in Fig. 3. Data from two dimensional feature space are mapped into higher three dimensional feature space by kernel function. There are four basic kernel functions – linear, polynomial, radial basis function (RBF) and sigmoid. RBF kernel function, which was used in our experiment, is defined [4]:

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2 \right) , \tag{10}$$

where $\gamma > 0$.

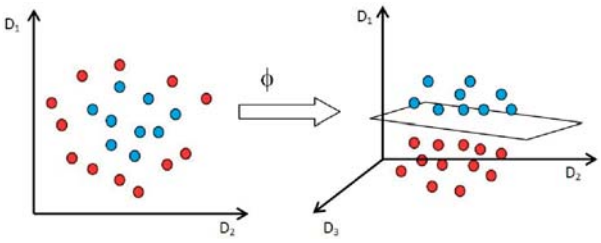

Fig. 3 Example of features mapping using kernel function

Aim of the SVM is to find a linear separating hyperplane with the maximal margin in this higher dimensional space. C is the penalty parameter of the error term. Value of penalty parameter must suffer condition $C > 0$. Not every function can be used as kernel, only those that comply with Mercer's conditions [6].

For SVM classification system, every attribute of the data is scaled to range $[1, -1]$. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric range [5].

SVM classifier requires setting up one or more parameters. In our experiment, we applied C-SVM formulation: included in implementation LIBSVM [7] with RBF kernel function; therefore we searched for two model parameters C and γ. We used Particle Swarm Optimization (PSO) [8] technique for parameter selection task.

3. Experimental results

Selection of classifier and feature vectors are one of the crucial parts of each classification system. Classification task is to correctly identify speakers known to the system based on the previous learning procedure. This learning process could be done by various techniques based on statistical modeling, distance measure or non-probabilistic linear binary classifier. Feature extraction is the process, when feature vectors are extracted from speaker utterances that represent information of identity to system better than the speech signal itself. Fig. 4 depicts a block diagram of classification system that we used in the experiments.

In evaluation process, we used MobilDat-SK database [9], [10]. The MobilDat-SK is corpus of mobile telephone speech recorded over GSM telecommunication network in Slovak language. From the corpus consisting of 1100 speakers, utterances of 20 speakers were randomly selected, while 3 different utterances pronounced by the same speaker were stored for each of the 20 speakers. We decided to use only 20 speakers for each test because of high computational expenses and thus long training procedure of SVM classifier. In many real applications (e.g. speaker separation and indexing in audio documents), this amount of speakers is usually sufficient. In the experimental part, we were also investigated classification ability of particular classification method when only a few training data is available. Each utterance has duration of approx. 8 seconds and is stored as uncompressed PCM WAV file with 16 bits resolution, and 8 kHz sample frequency. From the speaker utterances, 22 MFC coefficients were extracted as the speech features. Since MFC coefficients have great ability to describe a speech signal, we decided to employ these audio features. The frames of 30 ms length and 10 ms overlap were used. Silent frames for each speaker utterances were dropped out using short time energy threshold and simple GMM-based voice activity detector. We used 2 utterances (approx. 16 seconds) as a training data and 1 utterance as a test data for every speaker.

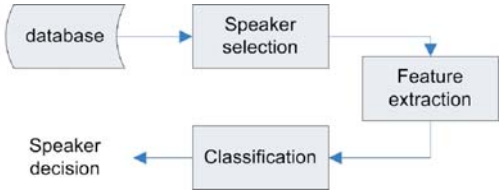

Fig. 4 Classification system

We applied two different approaches during performance evaluation of proposed classifiers. First, we compared classification accuracy on frame level, were each feature vector influences of the overall performance. The second approach of evaluation was per-

formed over whole recordings or utterances and the most frequently class occurred in the utterance were assumed as the class where the classified utterance belongs to. As a classification accuracy measure, F measure based on precision and recall was applied. For every classifier, each test was run 30 times to obtain the statistical credibility of classified data and the final values were averaged. All the tests were run in MATLAB program environment.

We used 3 different classifiers with the following parameters:

- GMM with probability density function (PDF) composed of 8 Gaussians and diagonal covariance matrix. The number of Gaussian components was chosen according to previous studies of training GMM on small amount of data [11], [12].
- SVM with RBF kernel function – model parameters selection were performed over parameters range $C = \{2^{-5}, 2^{-4.9}, ..., 2^{19.9}, 2^{20}\}$ and $\gamma = \{2^{-20}, 2^{-19.9}, ..., 2^{4.9}, 2^{5}\}$, criterion function for model parameters selection was 5-fold cross validation accuracy,
- kNN with $k = 7$ neighbors and Euclidean metric.

Experiment results for classifiers are shown in Tab. 1, Fig. 4.

Classification accuracy results Tab. 1.

Method	GMM	kNN	SVM
Frame level accuracy [%]	16.58	43.21	49.90
Record level accuracy [%]	31.89	92.15	98.11

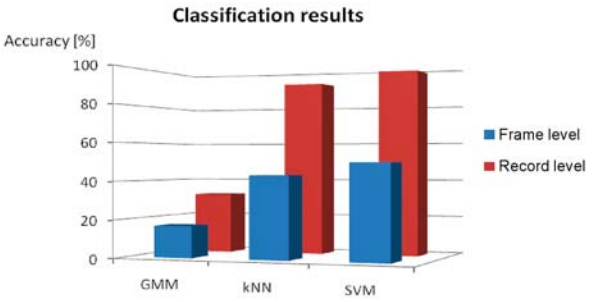

Fig. 4 Classification accuracy results

4. Discussion of the Results and Conclusion

In this paper, we described and evaluated three different classifiers used for speaker identification task. Classification accuracy for the dataset MobilDat-SK was computed for frames as well as for whole recording of each speaker consisting of all frames. The best classification accuracy of 98.11 % was achieved by SVM classifier. Thus the great discrimination properties of SVM as well as its ability to be trained on few examples have been proven by our experiments. Despite of high classification accuracy the disadvantage of SVM are the extremely high computational requirements resulting to very slow training procedure. It is interesting that the KNN classifier scored comparable classification accuracy – 92.15 % despite of its simplicity. The drawback of kNN is increasing computational complexity with large database. GMM classifier achieved significantly worse classification accuracy – 31.89% – than the other two classifiers. Reason of this fact is lack of training data for GMM classifier (Note, less than 16 seconds of speech data were utilized for training of the classifier).

Acknowledgements

This publication is the result of the project implementations:

Creating a new diagnostic algorithm for selected cancer diseases, ITMS 26220220022 supported by the Research & Development Operational Programme funded by the ERDF, and **Centre of excellence for systems and services of intelligent transport II.**, ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

"Podporujeme vyskumne aktivity na Slovensku/Projekt je spolufinancovany zo zdrojov EU

References

[1] REYNOLDS, D. A.: Speaker Identification and Verification Using Gaussian Mixture Speaker Models. *Speech Communication.* Vol 17, No. 1–2, 1995.

[2] BIMBOT, F. et al.: *A Tutorial on Text-Independent Speaker Verification,* EURASIP J. on Applied Signal Processing. Vol. 4, pp. 430–451, 2004.

[3] XINDONG, W., VIPIN, K.: *The Top 10 Algorithms in Data Mining,* Chapman & Hall/CRC, 2009.

[4] VAPNIK, V.: *Statistical Learning Theory,* Wiley, New York, 1998.

[5] HSU, C. W., CHANG, C. C., LIN, C. J.: *A Practical Guide to Support Vector Classification,* http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[6] JUNLI, C., LICHENG, J.: *Classification Mechanism of Support Vector Machines,* Proc. of 5th Intern. Conference on Signal Processing, 2000. WCCC-ICSP 2000, Vol.3, 2000, pp.1556–1559.
[7] CHANG, C. C., LIN, C. J.: LIBSVM: *A library for Support Vector Machines,* 2001. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[8] BLONDIN, J., SAAD, A.: *Metaheuristic Techniques for Support Vector Machines Model Selection.* In 10th Intern. Conference on Hybrid Intelligent Systems. 2010.
[9] RUSKO, M., TRNKA, M., DARJAA, S.: *MobilDat-SK a Mobile Telephone Extension to the SpeechDat-E SK Telephone Speech Database in Slovak.* In Proc. of the 11th International Conference Speech and Computer (SPECOM'2006), St. Petersburg, 2006, pp. 449–454.
[10] JUHAR, J., ONDAS, S., CIZMAR, A., JARINA, R., RUSKO, M., ROZINAJ, G.: *Development of Slovak GALAXY / VoiceXML Based Spoken Language Dialogue System to Retrieve Information from the Internet,* Proc. of the Interspeech 2006 – ICSLP, Pittsburg (USA), 2006, pp. 485–488.
[11] TADJ, C., DUMOUCHEL, P., OUELLET, P.: *GMM Based Speaker Identification Uusing Training-Time-Dependent Number of Mixtures,* Acoustics, Speech and Signal Processing, 1998. Proc. of the 1998 IEEE Intern. Conference on, Vol. 2, 12–15 May 1998, pp. 761–764.
[12] PARALIC, M., JARINA, R.: *Variable Component Approach in GMM – Based Speaker Modelling,* 3rd Intern. Acoustical Conference – EAA Symposium, Acoustics High Tatras 2006, Strbské Pleso, Slovakia, 2006, pp. 164–167.
[13] NABNEY, I. T.: Netlab: *Algorithms for Pattern Recognition. Advances in Pattern Recognition.* Springer, Berlin, 2001.