EXPLORING KERNELS IN SVM-BASED CLASSIFICATION OF LARYNX PATHOLOGY FROM HUMAN VOICE

Evaldas VAIČIUKYNAS*, Adas GELŽINIS*, Marija BAČAUSKIENĖ*, Antanas VERIKAS*', Aurelija VEGIENĖ°
*Department of Electrical & Control Instrumentation, Kaunas University of Technology, Lithuania; 'Intelligent Systems Laboratory, Halmstad University, Sweden; °Department of Otolaryngology, Kaunas University of Medicine, Lithuania

Abstract: In this paper identification of laryngeal disorders using cepstral parameters of human voice is investigated. Mel-frequency cepstral coefficients (MFCC), extracted from audio recordings, are further approximated, using 3 strategies: sampling, averaging, and estimation. SVM and LS-SVM categorize pre-processed data into normal, nodular, and diffuse classes. Since it is a three-class problem, various combination schemes are explored. Constructed custom kernels outperformed a popular non-linear RBF kernel. Features, estimated with GMM, and SVM kernels, designed to exploit this information, is an interesting fusion of probabilistic and discriminative models for human voice-based classification of larynx pathology.

Keywords: Laryngeal disorder, Pathological voice, Voice processing, Mel-frequency cepstral coefficients, Sequence kernel, Principal canonical correlation, Monte-Carlo sampling, Kullback-Leibler divergence, Earth mover's distance, GMM, SVM

1. Introduction

Identification of laryngeal diseases in clinical practice is a rather complex diagnostic procedure, involving evaluation of patient's complaints, case-record, and data of instrumental as well as histological examination. Patient's complaints are usually summarized as questionnaire data, while the instrumental examination results into a sequence of laryngeal images and voice records. Both questionnaire data and voice records can be attributed to the category of non-invasive measurements, which can be used for early detection of potential diseases and therefore be of great value in preventive care. Such noninvasive techniques are not restricted to the medical area alone, as they may also be of special interest in voice quality control for voice professionals such as singers, speakers, etc.

Data in this study are categorized into 1 normal class and 2 classes of laryngeal disorders, namely, nodular (nodules, polyps, and cysts) and diffuse (papillomata, hyperplastic laryngitis with keratosis, and carcinoma). Categorization into nodular and diffuse classes is based on visual appearance of vocal fold mass lesions, evaluated under direct microlaryngoscopy. Nodular lesions (localized thickenings) visually appear as single lesions of various sizes with a smooth, regular surface and distinct margins surrounded by a normal tissue of the vocal fold. Respectively, diffuse lesions visually appear as irregular, rough, multiple thickenings without distinct margins, often surrounded by an inflamed tissue and have a tendency to become cancerous. Final diagnosis was confirmed by histological examination of laryngeal specimens taken during endolaryngeal microsurgical intervention.

Pathological voice is induced by mass increase, a lack of closure, or elasticity change of the vocal folds. The result is that the movement of the vocal folds is not balanced and an incomplete closure of the vocal folds may appear in glottal cycles. This is the reason of changes in the whole harmonic structure (increasing the inter-harmonic energy and the fundamental frequency perturbation). Energy increases at higher components are from aerial turbulence induced by an incomplete closure of the glottal clef. Alterations related to the mucosal waveform due to an increase of mass emerge in low bands, whereas higher bands tend to reflect noisy components due to a lack of closure. Both alterations manifest themselves as noise with poor outstanding components and wide band spectra [2].

2. Voice database

Voice samples were recorded at the Department of Otolaryngology, Kaunas University of Medicine, Lithuania in a sound-proof booth on a digitized Sony Mini Disc Recorder MDS-101 (Tokyo, Japan) through a D60S Dynamic Vocal (AKG Acoustics, Vienna, Austria) microphone (with freq. range from 70 Hz to 20 kHz). Distance from the mouth was ~10 cm. Audio was saved in wav format (mono-channel PCM, 16 bit samples at 11 kHz rate), Nyquist frequency F_{max} = 5,5 kHz. Sustained phonation of vowel sound /a/ was considered.

In this study from mixed gender database of 810 subjects (130 normal / 212 diffuse / 459 nodular) we selected 410 subjects (130 normal / 140 diffuse / 140 nodular) and this balanced dataset was used to train and test the SVM-based classifiers. During preprocessing, silent parts, especially at the beginning and the end of recording, were eliminated. Each patient has 2 – 4 such recordings of various lengths (0.5 – 3 s) and associated clinical diagnosis – normal, diffuse, or nodular.

A previous study on the same database selected 104 subjects (25 normal / 25 diffuse / 54 nodular) where each subject had exactly 3 recordings (with average length of 2.4 s). A correct classification rate of 84.6% was achieved when using a 4-member voting committee of SVMs, with each member trained on different feature sets, to classify one voice record and averaging to aggregate decisions about one-subject data. Correct classification rate dropped to 67.31% (aggregated by averaging) and 68.27% (weighted averaging) when only single feature set of 10 MFCCs was used [1]. Note that,

since subjects we chose are not the same, these classification rates should not be compared to our results.

3. Feature extraction

Before windowing, the voice signal is pre-emphasized by forward differencing to reduce the effects of drifting amplitude. Since voice has low frequencies higher in amplitude than high frequencies, the 6dB/octave (naturally occurring attenuation) pre-emphasis finite impulse response high-pass filter is used to flatten the spectrum of signal by creating more equal amplitude of lows and highs (emphasizing higher formant components), which results in louder and sharper signal:

$$s(t) = 1 - 0.97 \cdot s(t-1). \quad (1)$$

Hamming windowing ensure smooth frame to frame transitions. Frame rate was 33 fps (~30ms size window). MFCC are extracted from preprocessed audio recordings. Data dimensionality is then further reduced by different strategies: selecting some frames as samples, squeezing closest frames by taking average of them or estimating statistical model from all the frames.

3.1. Cepstral coefficients

MFCCs are widely used features to characterize a voice signal and can be estimated by using a parametric approach derived from linear prediction coefficients (LPC), or by the non-parametric discrete fast Fourier transform (FFT), which typically encodes more information than the LPC method.
Signal is windowed in the time domain and converted into the frequency domain by FFT, which gives the amount of energy present within particular frequency range for each of 256 bins. With an 11 kHz sampling rate, the total frequency range is from 0 to 5,5 kHz (Nyquist frequency) and by splitting it into 256 equal intervals, the ~21,5 Hz range (frequency resolution) is covered with each bin. Frequency resolution tells how many Hz are represented by a single bin or how narrow the band filter of each bin is. Triangular Mel-frequency filters are then applied to reduce the amount of data by summing filtered FFT bin values to get the Mel filter bank outputs. Mel-scaling is performed to get higher resolution at low frequencies and lower resolution at high frequencies. This is based on the human perception, where relationship between the real frequency scale (Hz) and the perceptual frequency scale (Mel) is logarithmic above 1000 Hz and linear below.
Finally, MFCCs are obtained by applying discrete cosine transform (DCT) to the logarithm of Mel filter bank outputs (or energies). DCT represents signal in terms of the first basis function (constant component) and the remaining basis functions (components of successively increasing frequency), which are uncorrelated. First 13 components of DCT represent a compacted MFCC vector of the corresponding frame. Since sometimes better results can be achieved with just 12 components (without constant component, which reflects fundamental frequency), this version of MFCC features was also tested.
The Matlab code to calculate MFCC features was adapted from the Computer Audition Toolbox [9], where 40 (13 linearly spaced + 27 logarithmically spaced) triangular Mel-frequency filters are used, covering the frequency range from 133 Hz to 6854 Hz.

3.2. Sampling and averaging

After converting an audio signal to cepstral coefficients we have a vector of 13 MFCCs for each frame (window). The number of frames depends on the duration of single recording and, in our case, ranges from 21 to 98. Sampling then means selecting one or several frames, i.e. to get 1 sample we select the center frame and to get more samples we select other frames, spaced evenly. Such selection of equally spaced sample frames reminds putting centered 'comb' on the whole recording.
For example, to get 4 sample frames from 52, frame indices are calculated by these Matlab expressions:

```
cRadius = ceil( 52 / 4 / 2 ); mfccIndices = uint8( 1 + cRadius + ( (52-2*cRadius-1) / (4-1) ) .* [0:4-1] );    (2)
```

Indices resulting from equation (2) are 8, 20, 33 and 45.
Since voice recordings are of different length and number of frames is not the same across them, to get a fixed number of frames (which is a pre-requisite for SVM-based classification), simple time scaling can be implemented by averaging closest frames, instead of sampling. When a recording is shorter than the predefined number of frames, an extra frame is added between two neighboring frames that are closest in the Euclidean sense (have a smallest distance between their MFCC vectors). The inserted frame is the mean vector of the closest frames. When a recording is longer, such an averaged frame is placed instead of two neighboring frames. The process is repeated until the desired length is reached (recording is stretched or squeezed enough), where the number of iterations is equal to the absolute difference between the number of frames in the recording and the predefined number of frames to be left in the result [5].

3.3. Estimation with GMM

When applying estimation, we can use a description of fixed size to represent all frames. Higher number of frames here becomes an advantage, since it results in more exact representation of statistical information. One possible solution is to represent each recording (or all recordings of single subject) with a statistical model and use it's signature as features or to apply some parametric or non-parametric measure to assess distance between estimated models and use the distance to calculate kernel (Gram matrix) for classification.
Gaussian mixture modeling (GMM) can be regarded as a way of clustering and represents the data (in our case MFCCs') distribution as a probability density function p, which is a weighted sum of K components (Gaussians):

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \cdot G(\mathbf{x}; \theta_k) \quad (3)$$

where $\mathbf{x}$ is an M-dimensional feature vector, K is the number of components, w_k is a weight ($w_1 + .. + w_k + .. + w_K = 1$), and $G(\mathbf{x}; \theta_k)$ is an M-variate Gaussian density with it's parameters θ_k (mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$).

Given a set of feature vectors (all frames), the model parameters are estimated using an iterative expectation–maximization (EM) algorithm. Though the EM algorithm converges to a maximum likelihood it may converge to the local maximum. GMMs produced by the EM algorithm are, consequently, sensitive to initialization. Usually parameters are initialized by the K-means algorithm.

4. Classification

A support vector machine (SVM) and its least squares simplification (LS-SVM), are used for classification. SVM was originally created for binary classification problems. Multiclass classification (when the number of classes $C \geq 3$) usually combines several binary SVMs.

The minimum output coding (MOC) requires $L \leq C$ classifiers, where $L = \log_2 C$. The one-vs-one (1vs1) scheme constructs a separate binary classifier for every pair of classes and yields $C \cdot (C-1)/2$ classifiers, while the one-vs-rest (1vsR) scheme constructs a binary classifier for each class by separating observations of this class from the rest and yields C classifiers. Decision is implemented by the voting (1vs1) or winner-takes-all strategy (1vsR).

Single optimization by Weston & Watkins (SOW) attempts to directly solve a multiclass problem. This is achieved by modifying the binary class objective function and adding a constraint to it for every class [8].

4.1. SVM

SVM is a large margin classifier and determines the optimal hyperplane by maximizing the margin. The generalization error of SVM decreases with increasing margin. Some important advantages of the SVM compared to other AI techniques are good generalization properties, robustness in high dimensions, convexity of objective function, and a well-defined learning theory.

Suppose we have a set of N training samples, each represented as ($\mathbf{x}_i$, y_i), where $\mathbf{x}_i$ is the feature vector in the input space and y_i is the class label, which can be positive (+1) or negative (-1).

Let $\mathbf{z}_i = \Phi(\mathbf{x}_i)$ denote the corresponding feature space vector with a mapping function Φ from the input space to a high-dimensional feature space. The hyperplane can then be defined as:

$$\mathbf{w} \cdot \mathbf{z} + b = 0 \quad (4)$$

where $\mathbf{w}$ is the vector defining the orientation of the hyperplane and b is the bias parameter. Data samples are said to be linearly separable if there exists ($\mathbf{w}$, b), such that

$$\mathbf{w} \cdot \mathbf{z}_i + b \geq +1 \Rightarrow y_i = +1 \quad (5)$$

$$\mathbf{w} \cdot \mathbf{z}_i + b \leq -1 \Rightarrow y_i = -1 \quad (6)$$

are valid for all data samples. To deal with samples that are not linearly separable, (5) and (6) can be generalized by introducing the non-negative slack variables ξ

$$y_i(\mathbf{w} \cdot \mathbf{z}_i + b) \geq 1 - \xi_i \quad (7)$$

where ξ_i are non-zero for those $\mathbf{z}_i$, which do not satisfy (5) or (6).

To construct an optimal hyperplane, SVM uses an iterative training algorithm, which minimizes the error function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \quad (8)$$

subject to constraints (7), where C is the capacity constant (or regularization parameter).

The error function (8) is minimized by introducing Lagrange multipliers and using Kuhn-Tucker theorem of optimization theory. Non-zero coefficients in Lagrange expansion are the so called support vectors.

4.2. LS-SVM

LS-SVM is a modified version of SVM, with equality in (7) constraints instead of inequality [3], which results in a set of linear equations instead of quadratic programming. The solution of the linear system can be calculated efficiently using a conjugate gradient method. In the LS-SVM case all data points are relevant and used as support vectors.

5. Kernel trick and kernel functions

Kernel trick is a method of using a linear classifier to solve a non-linear problem by nonlinearly mapping the original observations into a higher-dimensional space, where a linear classifier is subsequently used. This makes linear classification in the new feature space equivalent to non-linear classification in the original input space. Instead of dealing with samples in the input space, one works with their mappings in the feature space without explicitly calculating them, since a kernel function returns a dot-product between vectors there.

The kernel function measures similarity or distance between a pair of variables. Once the kernel is chosen, the feature space (subspace or space spanned by all the training samples) is automatically determined and can be used for classification.

For all feature extraction strategies, besides the RBF kernel, we also explored a version of sequence kernel, based on kernelized principal angles (KPA). In the

estimation case, the Gaussian mixture means (centers) alone were used as features for classification with the RBF and KPA kernels. To exploit information, present in the covariance matrices, new kernels were created by calculating the distance between two GMMs. The similarity metrics used here were: the distance approximation from the Monte-Carlo sampling (MCS), and the Kullback-Leibler divergence combined with the earth mover's distance (EMD).
GMM models were estimated using the Matlab toolbox Netlab, while the similarity measures (MCS and EMD) between them were calculated using the MA Toolbox [10]. The resulting distance matrix **D** was further processed by the rbf_of_dist function (from the Spider toolbox [11]), to get a well-formed kernel matrix **K** :

$$\mathbf{K}_{ij} = \exp\left(-\frac{\mathbf{D}_{ij}}{2\cdot\sigma^2}\right) \tag{9}$$

5.1. Radial basis function

Radial basis function (RBF) is by far the most popular choice of kernel types used in SVM classification. The RBF kernel function is

$$K(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-\frac{\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2}{\sigma^2}\right) \tag{10}$$

where σ is the width (variance) of the basis functions.

5.2. Kernel principal angles

Ensemble matching methods generally consider a task of obtaining a similarity function which operates on pairs of sets of feature vectors (matrices) or ensembles.
The sequence kernel, we explored, is defined over a pair of matrices, rather than over a pair of vectors, and calculates the kernelized principal angle (KPA) between subspaces. The principal angle is the angle between two linear subspaces of two matrices, each matrix composed of feature vectors as columns. Kernelizing this angle via the kernel trick, allows it to be calculated between non-linear subspaces. The degree of alignment of two subspaces spanned by the elements of the two ensembles is used here as a measure of similarity. Larynx pathology recognition is done on the premises, that different disorders generate different subspaces and principal angles between them can be measured.
Positive-definite kernel (similarity metric) is given by:

$$K(\mathbf{X}_i,\mathbf{X}_j) = \prod_{k=1}^{K}\cos^2(\varphi_k) \tag{11}$$

where $\cos(\varphi_k)$ are often referred to as principal correlations or canonical correlations of the matrix pair. If angle $\varphi_k = 0$, then $\cos(\varphi_k) = 1$ and the vectors are said to be parallel. If angle $\varphi_k = 0$, then $\cos(\varphi_k) = 1$ and the vectors are said to be orthogonal.
The kernel trick is performed here for correlation, to compute principal angles in the feature space induced by a minor Gaussian (RBF) kernel [6]. Using a linear minor kernel (polynomial degree 1) is equivalent to computing principal angles in the input space (between linear subspaces).
The modified version of the kernel Gram-Schmidt (MKGS) orthogonalization was used to compute principal correlations. The MKGS algorithm [7] for QR decomposition in the feature space, used in this work, is much more numerically stable than the classical kernel Gram-Schmidt (KGS) version [6].

5.3. Kullback-Leibler divergence

The Kullback-Leibler divergence, also known as mutual information, relative entropy or, simply, information divergence, is a classic information gain measure of the asymmetric difference between two distributions, i.e. it measures the divergence from one probability distribution to another. The symmetric KL-divergence between two distributions p and q (two GMMs, for example) may be expressed as

$$D(p,q) = \int p(\mathbf{x})\cdot\log\frac{p(\mathbf{x})}{q(\mathbf{x})} + \int q(\mathbf{x})\cdot\log\frac{q(\mathbf{x})}{p(\mathbf{x})} \tag{12}$$

Its value ranges between 0 and infinity, and is 0 if and only if the two distributions are identical.

5.4. Monte-Carlo sampling

A closed form expression for KL divergence only exists when the number of Gaussian mixtures is 1. We can use Monte-Carlo simulations to approximate the KL-divergence between two non-single Gaussians p and q as follows:

$$D(p,q) \cong \sum_{t=1}^{T}\log\frac{p(\mathbf{x}_{pt})}{q(\mathbf{x}_{qt})} + \sum_{t=1}^{T}\log\frac{q(\mathbf{x}_{qt})}{p(\mathbf{x}_{pt})} \tag{13}$$

where $\mathbf{x}_p$ and $\mathbf{x}_q$ are either the real data observations that were used to estimate the parameters of p and q or they are synthetic samples, i.e. randomly generated from the estimated probability distributions p and q, and T is the number of observations or samples. The above approximation of KL-divergence is, exactly, the distance measure based on the cross likelihood ratio test. The drawback of the Monte-Carlo approach is that even though we have a compact probabilistic representation of data in the GMM form, we still have to refer back to the original data, and because of the stochastic nature of the Monte-Carlo method, approximations could vary in different runs.

5.5. Earth mover's distance

We use the Earth Mover's Distance (EMD) to calculate the distance between the probability distributions in each dimension and in such a way compute the distance between 2 recordings. The EMD computation is based on a simplified solution to the transportation problem where the total supply equals the total demand (sum of

priors of source and target GMMs are equal). Instead of comparing the values of each GMM mean, the minimum amount of work needed to transform one distribution (hills) into the other (valleys) is calculated.
EMD is conceptually equivalent to the Mallows or Wasserstein distance between probability distributions and in the case of two distributions with equal masses, they are exactly the same [8].

6. Experiments

6.1. Experimental setup

MFCCs were normalized to zero mean and unit variance. To evaluate the generalization error of SVM, stratified 10-fold cross validation was used. The appropriate values of SVM parameters C and σ were found experimentally. Comparison of results obtained for two different models was done with the help of the right tailed two-sample T-test (with Behrens-Fisher's problem when variances were found to be unequal).

6.2. Results

Fig. 1 – Fig. 7 present the test set data classification accuracy obtained using the different coding schemes for the pure SVM (SOW, 1vs1, 1vsR) and the LS-SVM (MOC, 1vs1, 1vsR). As can be seen from the figures, the MOC performed significantly worse than the other techniques, since it used the least number (only 2) of binary classifiers. The ordinary SVM (left side) was superior to the LS-SVM (right side), in all the tests. The 95% confidence interval is also shown in the figures.
On average, the sequence kernel (KPA) has shown a more stable and slightly better accuracy than the Gaussian (RBF) kernel, see Fig. 1 – Fig. 3.
When using RBF and KPA kernels, it was found that it is better to concatenate GMM means of 3 recordings, rather than using a single GMM from all recordings of a subject, see Fig. 3 and Fig. 4.
As can be seen in Fig. 5 – Fig. 7, the MCS and EMD kernels outperformed the RBF and KPA ones. In many tests the difference in accuracy was statistically significant.

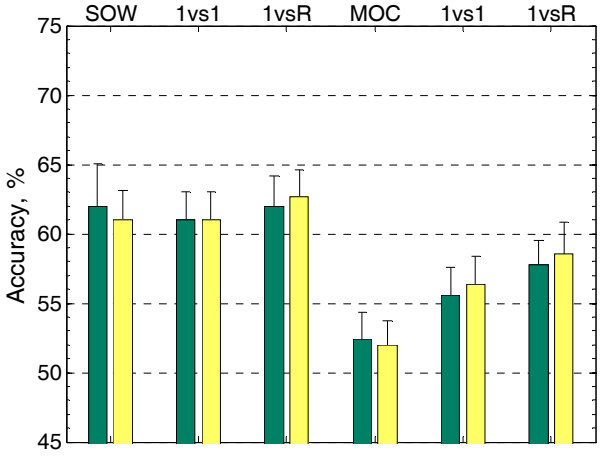

Fig. 1. The test set classification accuracy, obtained by MFCC sampling and using RBF (dark) or KPA (light) kernels.

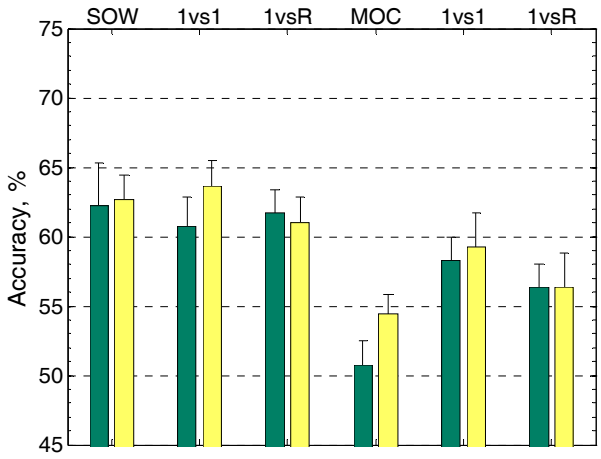

Fig. 2. The test set classification accuracy, obtained by MFCC averaging and using RBF (dark) or KPA (light) kernels.

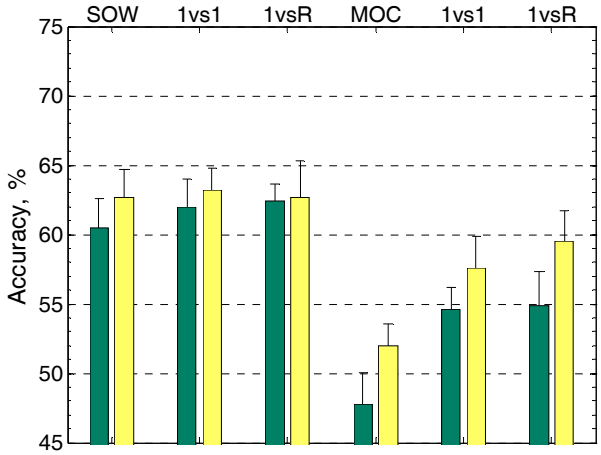

Fig. 3. The classification accuracy, obtained by MFCC estimation with 3 GMMs and RBF (dark) or KPA (light) kernels.

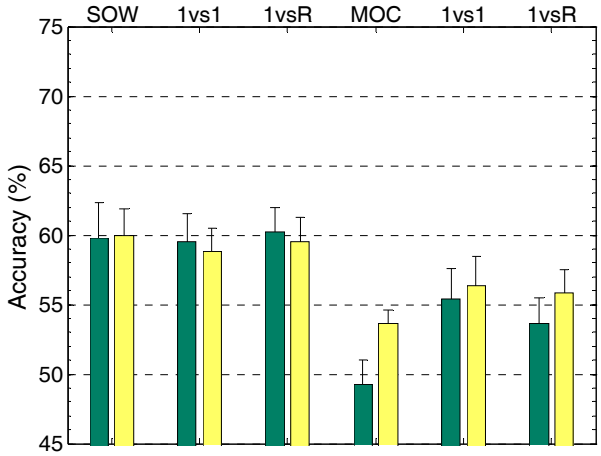

Fig. 4. The classification accuracy, obtained by MFCC estimation with 1 GMM and RBF (dark) or KPA (light) kernels.

The MCS kernel performed better when GMM used the full covariance matrix, $\boldsymbol{\Sigma}_{full}$, rather than a diagonal one, $\boldsymbol{\Sigma}_{diag}$, see Fig. 5. However, the EMD kernel has shown the opposite behavior, see Fig. 6. The EMD kernel, using GMM with diagonal covariance matrix, $\boldsymbol{\Sigma}_{diag}$, provided the best overall performance, see Fig. 7.

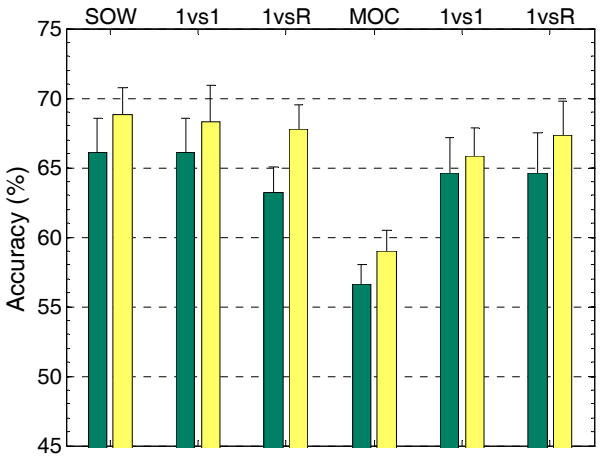

Fig. 5. The test set classification accuracy, obtained using the MCS kernel and GMM, with Σ_{diag} (dark) or Σ_{full} (light).

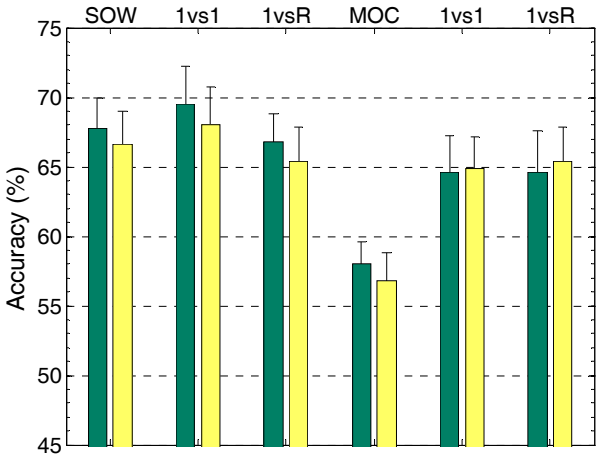

Fig. 6. The test set classification accuracy obtained using the EMD kernel and GMM, with Σ_{diag} (dark) or Σ_{full} (light).

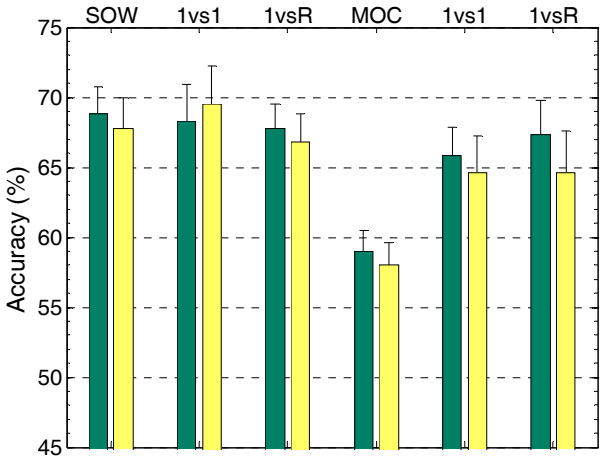

Fig. 7. Best classification accuracy, obtained using the MCS kernel with Σ_{full} (dark) and EMD kernel with Σ_{diag} (light).

7. Discussion and conclusions

It was shown experimentally that the KPA kernel slightly outperforms the RBF one. However, the difference is not statistically significant.

While using RBF and KPA kernels, it was found that it is better to concatenate GMM means of 3 recordings, rather than use a single GMM representation of all subject's recordings. For MCS and KPA kernels, when each patient is represented with several GMMs, as opposed to a single GMM from all individual recordings, his data should be aggregated on the decision level. By automatically detecting the most representative number of Gaussian mixtures for each patient (or recording) and, therefore, reducing the model order (i.e. variable size GMMs instead of fixed), one could probably achieve a better accuracy while significantly lowering the computational cost.

The sequence kernel (KPA) and the distance kernels (MCS and EMD) outperformed the popular Gaussian (RBF) kernel, but the difference is statistically significant only in the distance kernels case. The MCS kernel, using GMM with full covariance matrices, and the EMD kernel, using GMM with diagonal covariance matrices, provided the best results.

8. References

1. Gelzinis A., Verikas A, Bacauskiene M. Automated speech analysis applied to laryngeal disease categorization. Computer Methods and Programs in Biomedicine 1, Vol. 91, 2008, p. 36-47
2. Godino-Llorente J. I., Gómez-Vilda P., Blanco M. Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters. IEEE Transactions on Biomedical Engineering 10, Vol. 53, 2006, p. 1943-1953
3. Suykens J. A. K., Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters 9, 1999, p. 293-300
4. Weston J., Watkins C. Multi-class Support Vector Machines. 7th European Symposium on Artificial Neural Networks (ESANN), 1999, Brussels
5. Doremalen J. Hierarchical Temporal Memory Networks for Spoken Digit Recognition. Dept. of Language & Speech, Radboud, 2007, p. 17-18
6. Wolf L., Shashua A. Kernel principal angles for classification machines with applications to image sequence interpretation. 10th IEEE Conference on Computer Vision and Pattern Recognition, Madison, 2003, p. 635-642
7. Zheng W. Class-Incremental Generalized Discriminant Analysis. Neural Computation 18, 2006, p. 979–1006
8. Levina E., Bickel P. The Earth Mover's Distance is the Mallows Distance: Some Insights from Statistics. 8th IEEE International Conference on Computer Vision (ICCV), 2001, p. 251-256
9. Dubnov S., Yazdani M. Computer Audition Toolbox (CATbox), UCSD's Computational Statistics and Machine Learning group (CoSMaL), http://cosmal.ucsd.edu/cal/projects/CATbox, 2008
10. Pampalk E. A Matlab Toolbox to Compute Music Similarity from Audio. 5th Int. Symposium on Music Information Retrieval (ISMIR), 2004
11. Matlab Toolbox for Kernel Methods: Spider. Max Planck Institute for Biological Cybernetics, 2006