Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/100324

An Improved GMM-SVM System based on Distance Metric for Voice Pathology Detection

Fethi AMARA [1,*], *Mohamed FEZARI* [1] *and Hocine BOUROUBA*[2]

[1] LASA Laboratory, Electronic Department,Faculty of Sciences, Badji Mokhtar University, P.O. Box 12, 23000 Annaba, Algeria.
[2] PIMIS Laboratory, Electronic Department, Faculty of Sciences, 08 mai 1945 University, P.O. Box 401, 24000 Guelma, Algeria.

Received: 22 Oct. 2015, Revised: 22 Jan. 2016, Accepted: 23 Jan. 2016
Published online: 1 May 2016

Abstract: As acoustic signal generated from vocal folds is directly affected by vocal tract pathologies, it can be an effective tool for diagnosis purpose. In this work, we present an efficient method for voice pathology detection based on speech signal processing and machine learning techniques. In the proposed method, we used MFCC to represent the signal features, and we chose to combine GMM and SVM classifiers to benefit from their generative and discriminative natures respectively. That is to exploit the similarity function of the RBF kernel to separate the GMM models representing normal and pathological voices. To further improve the separation, we used modified versions of the well known Kullback-leibler and Bhattacharyya distances. The modified distances, unlike the classical ones, do satisfy all metric axioms. As a result, we obtained an improvement of 2 % and 4 % in terms of sensitivity compared to using the classical Kullback-leibler and Bhattacharyya distances respectively. The Receiver Operating Curve (ROC) does illustrate the efficiency of the proposed method.

Keywords: Voice disorders detection, GMM-SVM, Similarity function, Kullback-Leibler divergence, Bhattacharyya distance, Triangle Inequality Violation.

1 Introduction

Assessment voice quality is an important tool for dysphonia evaluation. It is usually based on perceptual analysis [1] or instrumental evaluation which comprises acoustic and aerodynamic measure [2]. However, the first one is subjective because of the variability between listeners, whereas the second evaluation is invasive since it requires instruments, and on the other hand it has a limited reliability. This is why the development of an efficient system for classification is proposed as a complementary tool with the other mentioned techniques. The state of the art is based on two principal approaches: acoustic analysis and statistical methods. The first approach consists in comparing acoustic parameters between normal and abnormal voices such as fundamental frequency, jitter, shimmer, harmonic to noise ratio, intensity [3,4,5,6]. The major disadvantage is that the evaluation of the acoustic parameters depends on the accuracy estimation of the fundamental frequency which is not a trivial task in the case of certain pathologies.

N.Saenz-Lechon *et al.* [7] presented an overview of the previous classification schemes applied to voice disorders on Massachusetts Eye and ear infirmary (MEEI) Database. They described some methodological paradigms to be considered when designing an automatic pathological voice detection system. They used the multilayer perceptron neural network as a classifier with MFCC parameters. The objective of the work was not to improve the performance but to show how to design a detector.

In this task, GMM is considered to be an efficient tool, as mentioned by Godino *et al.* in [8]. They used the GMM to examine the effectiveness of the short term cepstral parameters as features to characterize the vocal folds pathologies, where the best results of 94% of efficiency were obtained using 24 MFCC and a GMM with 6 mixtures. In [9] Ji Yeoun Lee *et al.* compared their results with those obtained in [8], where their proposed technique contains two essential parts, an MFCC based GMM algorithm as primary classifier, and a high order statistics in the second stage. They attended an accuracy of 96,96%.

[*] Corresponding author e-mail: amafethi@gmail.com

David *et al.* in [10], realize a set of experiments around MEEI database and Saarbrcken Voice Database. They proposed MFCC and noise based features to train a generative GMM. The enhancement of the performance are based on the scores calibration and the fusion of different vowels /a/,/i/,/u/ at different intonations. They got 17,67% of improvement for the AUC (Area Under Curve).

Support vector machines (SVM) is also an important classifier which gave very promising results in this domain. In [11] SVM is applied to test the effectiveness and reliability of the short term cepstral and noise parameters. Wenxi Chen *et al.* in [12] confirm the efficiency of the SVM, where 25 acoustic parameters are extracted and transformed via the principal component analysis (PCA). The original dataset reduced into only two features to train SVM via three different kernels. In recent studies, [13,14] Nafise *et al.* investigate different wavelet transforms to train SVM in the context of voice pathologies assessment and voice disorder sorting, and they had obtained good results.

Compared with GMM and SVM, other classifiers such as Hidden Markov Models (HMM) and artificial neural networks (ANN) are less used. In a comparative study, proposed by Jianglin Wang *et al.* in [15], the above mentioned classifiers are evaluated in diagnosis of vocal folds. GMM were very performant since it offers high classification rate in term of TP (true positive) 97,8%. However, SVM and HMM gave small FN (false negative) at 0,5% and 0% respectively. Another comparative study in [16], pattern recognition methods were applied in the classification of respiratory sounds into normal and wheeze. It showed that the more significant results were obtained using MFCC/GMM.

All the mentioned works are concentrated in finding appropriate features, which allow an efficient separation. In this study, we focus our effort on exploiting and improving the capacity of the classifier itself. According to their power, the mentioned classifiers can be divided in two main categories: generative and discriminative. GMM and HMM belong to the first category, and their main advantage is the capacity to represent data which allows us to get optimal model. The second category includes SVM and ANN, which have the ability to separate classes. The development of a hybrid system is a way to exploit the two capacities. As mentioned in the state of the art, GMM and SVM are the more performant and robust combination of classifiers. The hybridization between both classifier is well recommended [11].

Most hybridization of this type used SVM to separate GMM models by the mean of RBF kernel, where Kullback-leibler distance is kernalized. This approach has demonstrated its effectiveness in many Multimedia applications [17,18]. Pathological voice classification is no exceptione, some works focussed on this approach [19, 20]. Evaldas Vaiciukynas *et al.* in [20] developed a hybrid system where the main goal is to exploit the similarity function of the RBF kernel using the Kullback-leibler distance approximated with Monte-Carlo simulation (KL-MCS), and Kullback-Leibler combined with Earth mover's distance (KL-EMD). This study proved that the similarity function is a very powerful tool to measure the similarity/dissimilarity between GMM. However, the embedded distances *i.e* (KL-MCS) and (KL-EMD), are not metrics since they do not satisfy all metric axioms, especially, the triangle inequality. This violation has a negative influence on the detection task where two highly dissimilar models can be both similar to an unknown model.

Recently, Karim.T *et al.* in [21], proposed a modification for the Kullback-Leibler and Bhattacharyya distances in such a way they transform them into distance metrics. In this paper, we are interested to exploit the similarity function of the RBF kernel but by using the modified versions of both distances. This would enable the enhancement the discriminative capacities between GMM.

The rest of the paper is organized as follow: section two describes the different steps used in our method. The experiments are presented in section three. Section four illustrates the results and the performance evaluation. Finally, we conclude the paper and give suggestions for future work in the last section.

2 Methodology

The general block diagram describing the process set up for the detection of voice disorder is presented in fig.1.

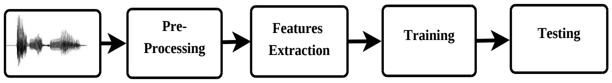

Fig. 1: Block diagram for voice pathology detection

In what follows, we give the description of each step is presented in the following part.

2.1 Voice disorder database

The database represents an essential factor to develop a detector. According to the overview of Nicolas Saenz *et al*. in [7], the use of a standard speech corpora might be necessary to compare the obtained results with those that exist. It allows researchers to test the effectiveness and the reliability of the used methods. They recommend to use Massachusetts Eye and Ear Infirmary (MEEI) database since it is well known in this domain. In the same overview some disadvantages were cited. We quote here the more significant:

– Not all the pathological patients have corresponding recordings nor diagnoses.

- –Normal and pathological voices were recorded at different locations (Kay Elemetrics and MEEI Voice and Speech Lab., respectively), assumedly under the same acoustic conditions, but there is no guarantee that this fact has no influence in an automatic detection system.
- –There is a heterogeneous number of pathologies in the database, with almost 200 different diagnoses, probably because they were included as they were captured in the clinical practice. There are a lot of files labeled with several diagnoses, pertaining sometimes to different categories of voice disorders.

However, our work is built around Saarbrcken Voice Database (SVD). It is a free database developed by Putzer *et al*. at the Institute of Phonetics, University of Saarland (Germany) [22]. It contains healthy and pathological recordings as follow:

- –Sustained vowels /a/, /i/, /u/ pronounced at different intonations (low, normal, high and low-high-low) during 1-3 s.
- –Sentence ”Guten Morgen, wie geht es Ihnen?”. it means: Good morning, how are you?
- –Electroglottogram EGG.

All files are sampled at 50 KHz at 16 bit resolution.
Because of its novelty, it is not used in large works. In [10] David Martinez *et al.* show that SVD is more challenging than MEEI, which motivate us to used it. From this large database, we have selected patients suffering from neurological pathology (spasmodic dysphonia). This disease affects more women than men. This is why we have chosen female voices. All selected files from the database are filtered using one coefficient filter known as the pre-emphasis filter. It is expressed by:

$$h(z) = 1 - az^{-1} \tag{1}$$

Where $a \in [0,1]$, is the coefficient. This filter has the advantage to reduce the effect of the microphone by amplifying high frequencies to create more equal amplitude with low frequencies [20].

2.2 Features extraction

Features selection means finding good parameters which permit to categorize the healthy person from patient, the separation between normal and pathological voices needs efficient features. Spasmodic dysphonia is a disorder of vocal function characterized by larynx muscles spasms that interrupt or impede the regular flow of the voice, those perturbations are clearly audible,especially by qualified speech therapist, this is why we were encouraged to choose the MFCCs parameters. Those parameters are obtained calculating the Discrete Cosine Transform (DCT) over the logarithm of the energy in several frequency band, they are given by:

$$c[n] = \sum_{k=1}^{N} \log\left(E[k]\right) \cos\left(n[k - \frac{1}{2}]\right) \frac{\Pi}{N} \tag{2}$$

Where $n = 0, 1, N$ is the number of desired coefficients.
In order to investigate the proprieties of the dynamic behavior of speech signal, the analysis can be extended to compute the temporal derivatives of the MFCC parameters. The first derivative (Δ) is given by:

$$\Delta c_n[p] = \mu \sum_{k=-K}^{K} kc_n[p+k] \tag{3}$$

Where n is the order of coefficients, p is the time, μ is the normalization constant, and k is number of frames.
The second derivative $(\Delta\Delta)$ are calculated using the same equation. These parameters are extracted using the *melcepst* Matlab function.

2.3 Combining GMM and SVM

2.3.1 Modeling by GMM

Gaussian mixtures models (GMM) is the most popular classifier in speech/speaker recognition. It consists in representing the extracted features by a weighted sum of M Gaussian densities as follow:

$$p(x \setminus \Theta) = \sum_{k=1}^{K} w_k g(x, \mu_k, \Sigma_k) \tag{4}$$

Where x is the features vector, Θ is the model that consists in K components $g(x)$, and w_k is the weights of the k^{th} component, knowing that $\sum_{k=1}^{K} w_k = 1$.
Each component has the following general form:

$$g(x, \mu_k, \Sigma_k) = \frac{1}{(2\Pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\{\frac{-1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\} \tag{5}$$

Where μ_k and Σ_k are respectively the mean and the covariance matrix of the k^{th} densities, and d is dimension of features vector.
Maximum likelihood (ML) is a good way to get the optimal model for representing our data.
ML criteria is given by:

$$p(X \setminus \Theta) = \prod_{i=1}^{M} g(x_i \setminus \Theta) \tag{6}$$

Where $X = (x_1, x_2, \ldots x_M)$.

2.3.2 Discrimination by SVM

Support vector machines (SVM) was introduced by Vapnik [23], and it is used basically in binary classification. It consists in maximizing the margin between the nearest points of the two classes. We receive training examples of the form:

$$\{x_i, y_i\}, \quad x_i \in R^d, \quad y_i \in \{1, -1\}, \quad i = 1...M \tag{7}$$

We call x_i the co-variate or input vectors and y_i the target value or labels. Our task is to predict whether a test sample belongs to one of two classes. We consider a very simple case where the data are linearly separable. We can make a decision according to the following expression:

$$f(x) = w^T x + b \tag{8}$$

We denote w the separating hyperplane, and b the bias term. All data satisfy the following constraints:

$$w^T x + b \geq 0 \quad if \quad y_i = 1 \tag{9}$$

$$w^T x + b \leq 0 \quad if \quad y_i = -1 \tag{10}$$

From (9) and (10) we derive the inequality:

$$y_i(w^T x + b) - 1 \geq 0 \tag{11}$$

We can get the optimal separating hyperplane by maximizing the margin between the two classes. The margin is given by $\frac{2}{\|w\|}$. It is then simple to minimize $\| w^2 \|$, and so the optimization formulation becomes:

$$\begin{cases} minimize \ \frac{1}{2} \| w \|^2 \\ subject\ to \ \ y_i(w^T x + b) \geq 1, \ \ \forall i \end{cases} \tag{12}$$

To solve (12), we need to introduce the Lagrangian formulation to obtain the following dual problem:

$$\begin{cases} L(w,b,\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j (x_i, x_j) \\ \alpha_i \geq 0 \\ \sum_{i=1}^{M} \alpha_i y_i = 0 \end{cases} \tag{13}$$

Finally the decision function has the form:

$$f(x) = \sum_{i \in i_s} \alpha_i y_i (x_i . x) + b \tag{14}$$

Due to the real data nature which are not always linearly separable, the kernel trick appears as a solution to construct a mapping from the vector x into a higher-dimensional feature space, where it is possible to separate the classes linearly. The new decision function takes then the following form:

$$f(x) = \sum_{i \in i_s} \alpha_i y_i k(x_i, x) + b \tag{15}$$

The Mercer condition states that the kernel function $k(.,.)$ must be positive semi-definite to ensure that the margin concept is valid, and the optimization of the SVM is bounded [24]. The $k(.,.)$ can be expressed by:

$$k(x,y) = g(x)^T g(y) \tag{16}$$

Where $g(x)$ is the nonlinear vector function ensuring the mapping to the feature space.
In SVM classification, the radial basis function (RBF) is the most popular kernel function. It is given by:

$$k(x,y) = \exp(\frac{\|x-y\|^2}{2\sigma^2}) \tag{17}$$

Where σ represents the width of the basis function, and $\|x-y\|^2$ is the similarity function.

2.4 Exploiting the similarity function

Similarity plays an essential role in many pattern recognition problems such as clustering, classification and retrieval problems [25]. It consists in giving a quantity that reflects the relationship between objects. It will be then possible to classify new objects in the appropriate group. Kernels can be used to measure the degree of similarity, especially, The RBF kernel. It is based on the similarity function $\|x-y\|^2$, which is a Euclidean distance. However, Euclidean distance fails to measure the distance between distributions. So, to be able to measure the similarity between distributions, other distance measures have been proposed. These new distance measures could be embedded in the kernel's equation as follows:

$$k(x,y) = \exp(\frac{D}{2\sigma^2}) \tag{18}$$

Where D is the distance matrix, and k is the pre-computed kernel.

In this study, we are interested in exploiting the discriminative capacity of the RBF kernel to seperate GMM models using similiarity functions that can actually measure the distance between GMM. As a matter of fact, we are going to use, not even the conventional Kullback-Leibler (KL) and Bhattacharyya (Bh) distance measures, but modified versions of them. We give in the next subsection the description of the conventional distances, and discuss their actual limits. Then we will present the modified distances that we will be using in our method.

2.4.1 Conventional measures

In many applications of pattern recognition, especially, in speech recognition, Kullback-Leibler (KL) and Bhattacharyya (Bh) distance are widely used to measure the distance between distributions [26,27]. A short description of both is presented in the following:

1) Kullback-Leibler distance

The KL distance is one of the important tools for similarity measure between two probability density functions (pdf). The KL distance between the pdf's P and Q is given by:

$$KL(P\|Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \tag{19}$$

This distance is not symmetric, which means $KL(P\|Q) \neq KL(Q\|P)$. It is recommended to use a symmetric version [28] which can be expressed by:

$$KL_s(P\|Q) = \left| \frac{1}{2} \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx + \frac{1}{2} \int_{-\infty}^{\infty} q(x) \ln \frac{q(x)}{p(x)} dx \right| \tag{20}$$

Notice that we cannot compute the KL distance in its present form. Instead, we approximate it using the monte-carlos simulation (MCS) as follows:

$$KL(P\|Q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx \approx \frac{1}{N} \sum_{t=1}^{N} \log \frac{p(x_t)}{q(x_t)} \tag{21}$$

N represents the number of data samples generated from $p(x)$

And so the symmetrized version takes the following formula:

$$\begin{aligned} KL_s(P\|Q) = \Big| &\frac{1}{2N} \sum_{x \to p}^{n} \ln p(x) - \frac{1}{2N} \sum_{x \to p}^{n} \ln q(x) \\ &+ \frac{1}{2N} \sum_{x \to q}^{n} \ln p(x) - \frac{1}{2N} \sum_{x \to q}^{n} \ln q(x) \Big| \end{aligned} \tag{22}$$

2) Bhattacharyya distance

Bhattacharyya (Bh) distance between two Gaussian distributions P and Q is given by:

$$Bh(P\|Q) = \frac{1}{8} \mu^T \Gamma^{-1} \mu + \frac{1}{2} \ln(|\Sigma_1|^{-\frac{1}{2}} |\Sigma_2|^{-\frac{1}{2}} |\Gamma|) \tag{23}$$

Where $\mu = \mu_1 - \mu_2$ and $\Gamma = (\frac{1}{2}\Sigma_1 + \frac{1}{2}\Sigma_2)$.

μ_1, Σ_1 and μ_2, Σ_2 are the means and the covariance matrices of P and Q respectively.

2.4.2 Limitations of conventional measures

First of the all, let us define a metric. A metric $d(x,y)$ is a function that defines a distance between objects x and y. It must verify the following axioms:

- –Separation $d(x,y) \geq 0$.
- –Coincidence $d(x,y) = 0$ if and only if $x = y$.
- –Symmetric $d(x,y) = d(y,x)$.
- –Triangle inequality $d(x,z) \leq d(x,y) + d(x,z)$.

According to the above definition, the conventional distances KL and Bh are not metrics. In particular, they violate the triangle inequality.

Many research works discuss the impact of violating the triangle inequality. Tversky *et al* [29], test the effect of this axiom in the context of the similarity and the separability. It is examined on medoid based clustering of objects [30]. Sometimes, if the distance does not obey the triangle inequality, two highly dissimilar models can be both similar to another third model. In our case, a pathological model can be seen as a normal model and vice versa. This would increase the number of misclassifications and thus decrease the the accuracy of the system. Notice that if the distance used as similarity function preserves the triangle inequality, the exponential mapping (the RBF kernel) will preserve it too. Now in addition to not being a metric, the approximation of the KL distance could also vary in different runs, due to the stochastic nature of the monte-carlos simulation.

2.4.3 Conventional measure modification

As mentioned above, the classical KL and Bh distances do not satisfy all metric axioms. Karim.T *et al.* in [21] proposed a modification of those distances in order to transform them into distance metrics. The effectiveness of the their new metrics was demonstrated in the manifold learning. A short presentation of this modification is expressed in the following lines:

Kullback-Leibler distance can be expressed in its closed form by:

$$KL(P\|Q) = \frac{1}{2} \mu^T \Psi \mu + \frac{1}{2} tr\{\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\} \tag{24}$$

Where $\mu = \mu_1 - \mu_2$, $\Psi = (\Sigma_1^{-1} + \Sigma_2^{-1})$

For Bhattacharyya distance, the closed form is presented in the equation (23).

In the equations (23) and (24), the first terms measure the difference between means weighted with the covariance matrix, and the second terms measure the difference between covariance matrices. In both equations, both terms do not satisfy the triangle inequality, but since the terms are separate, it was possible to make the following modifications. Those consist in taking the square root of the first term, and for the second term, they proposed the Riemannian distance, which is given by:

$$d_R(\Sigma_1, \Sigma_2) = (\sum_{j=1}^{p} \log \lambda_j)^{\frac{1}{2}} \tag{25}$$

Where λ is the eigenvalue

The new distances take the formula:

$$KL_R(P\|Q) = \mu^T \Psi \mu + d_R(\Sigma_1, \Sigma_2) \tag{26}$$

$$Bh_R(P\|Q) = \mu^T\Gamma^{-1}\mu + d_R(\Sigma_1,\Sigma_2) \qquad (27)$$

KL_R, B_R denote the modified distances.
We can use a weighted version expressed as follows:
For kullback-leibler:

$$KL_R(P,Q,\beta) = \beta\mu^T\Psi\mu + (1-\beta)d_R(\Sigma_1,\Sigma_2) \qquad (28)$$

For bhattacharyya:

$$Bh_R(P,Q,\beta) = \beta\mu^T\Gamma^{-1}\mu + (1-\beta)d_R(\Sigma_1,\Sigma_2) \qquad (29)$$

$\beta \in [0,1]$. It weights the importance of each term.

It is worth noting that the modified version of KL, unlike the classical one, can be computed directly without resolving to Monte-Carlo simulation.

2.4.4 Adaptation to GMM

The modified KL and Bh distances, and their modified versions were designed to measure the distance between Guassian distributions. G.Sfikas *et al.* in [31] proposed an adaptation for Bhattacharyya distance to measure the distance between GMM. This is achieved by the following formula:

$$Bh_{GMM} = \sum_{i=1}^{n}\sum_{j=1}^{m}\Pi_i\Pi'_jB(P\|Q) \qquad (30)$$

Bh_{GMM} denotes Bhattacharyya distance adapted for GMM.
Π, n, Π', m are the weights and the number of mixtures of $p(x), g(x)$ respectively.
We can use the same method to adapt the modified versions of KL and Bh, given in equations (28) and (29), to take the following form:

$$KL_{R,GMM} = \sum_{i=1}^{n}\sum_{j=1}^{m}\Pi_i\Pi'_jKL_R(P\|Q) \qquad (31)$$

$$Bh_{R,GMM} = \sum_{i=1}^{n}\sum_{j=1}^{m}\Pi_i\Pi'_jBh_R(P\|Q) \qquad (32)$$

The different steps are detailed in the flowchart presented fig.2.

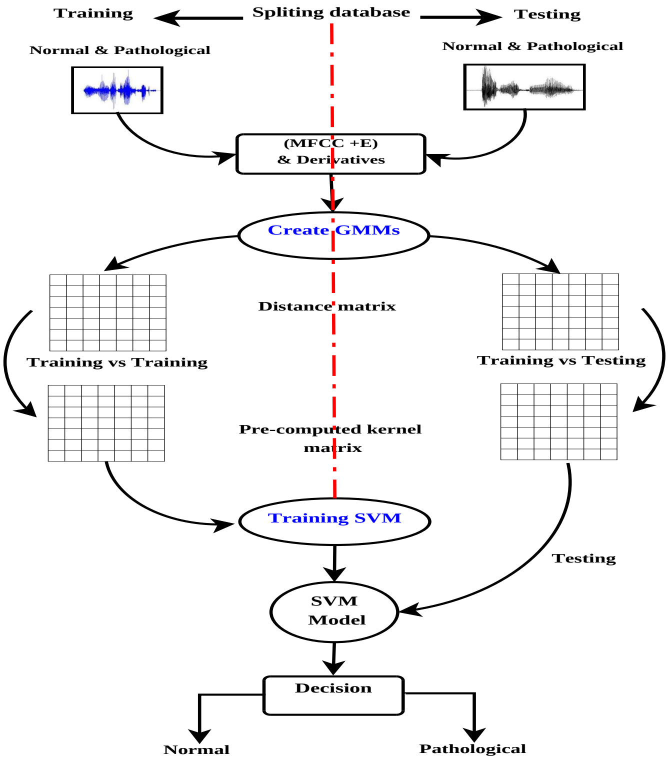

Fig. 2: GMM and SVM fusion flowchart

Table 1: Speech corpus

	Training set	Testing set	Age	Gender
Normal	45	15	30-60	female
Pathological	30	10	30-82	female

3 Experimental scenario

3.1 Data preparing

Voice samples are taken from the German database described above. A total of 100 voices are used, 60 normal and 40 pathological. Patients suffer from spasmodic dysphonia. Each patient phonated a sentence during 3 to 4 seconds. All files are down-sampled at 50Khz.

As mentioned above the speech corpus is divided into two sets, both containing normal and pathological voices. 75% of the data set is used for the training phase, and 25% is reserved to the testing phase. All the details (gender,age...) are described in Table 1. After splitting the database into training set and testing set, and before any processing, all files are down-sampled from 50KHz to 25KHz.
For the pre-emphasis we have applied a finite impulse response high pass filter with a coefficient $a = 0.95$.

3.2 Features extraction

Each signal (in WAV format) is segmented in the time domain into frames, using hamming window of 20 ms. The goal is to ensure the stationarity. Analysis is carried out using 50% of overlapping. From each frame 12

MFCC coefficients, normalized energy parameters, and their first and second derivatives (Δ and $\Delta\Delta$) are extracted. Features vector is a 39 dimensional MFCC. In our previous work [32], those coefficients gave the best accuracy.

3.3 Generative step (GMM)

In this step, we are interested in exploiting the generative capacitiy of the GMM. Each speaker (normal or pathological) is represented by a model. As it is well known, the number of mixtures usually has a major influence on its performance. Many experiments have been done to test this factor. Best results are obtained with 6 mixtures. GMM models are trained using the iterative algorithm expectation Maximization (EM) in order to get the maximum likelihood (ML). 200 iterations are performed to get the convergence. The initialization is ensured by K-means algorithm. This part was carried out using the Matlab toolbox Netlab.

3.4 Discriminative step (SVM)

Once the GMM models are obtained, we compute the distance matrix, noted D, as mentioned in equation (18). For the training phase, we have in total 75 models (40 normal speakers and 35 patients), so the distance matrix for training is 75x75. Noting that it is computed using the training models versus the training models. For testing phase, we have in total 25 models (15 normal speakers and 10 patients). The testing distance matrix is thence 75x25, and it is computed using training models versus testing models.

We note that distance matrices are computed using the weighted versions of both distances as mentioned in equations (28) and (29), where β takes its value between 0.6 and 0.9.

Next, the kernel matrix is pre-computed using the distance matrix as in (18). Then, it is used as the entry data to train SVM.

After the training step, the prediction can be made by comparing the testing kernel matrix with the obtained SVM model, as shown in flowchart of fig.2.

To obtain an accurate detection rate, we have to adjust the parameters of the kernel, the weight σ and the penalty error C. In many studies, grid search shown up to be the best way to determine the optimal pairs (σ, c) [11]. Noting in our case, σ takes its value in $[0.001, 0.9]$, and C takes its value in $[1000, 10000]$.

The cross-validation strategy is used to gauge the generalizability of our system. In other words, we want to test the performance of the learned model versus different testing data set. The experiment is repeated 10 times. The data set is splitted in 10 folds, each fold can be either in training set or testing set. This part has been done using the libsvm Matlab toolbox.

4 Results and discussion

In this section, we present and discuss the experimental results obtained using our proposed method.
The performance of the system could be presented by the confusion matrix given in table 2.

Table 2: Confusion matrix

System's Decision	Actual diagnosis	
	Abnormal	Normal
Abormal	TP	FP
Normal	FN	TN

True positives (TP) are pathological files correctly classified. False negatives (FN) are pathological files wrongly classified. True negatives (TN) are normal files correctly classified. False positive (FP) are normal files wrongly classified.

From the confusion matrix, we present other performance parameters such as sensitivity, specificity, and accuracy. Sensitivity is defined as the ratio between pathological files correctly classified and the total number of pathological files. Specificity is the ratio between normal files correctly classified and the total number of normal files. And accuracy is the ratio between all files correctly classified and the total number of files.Performance parameters are defined as follow:
Sensitivity=TP/(TP+FN)x100.
Specificity=TN/(TN+FP)x100.
Accuracy=TP+TN/(TP+TN+FN+FP)x100.

Table 3 contains the obtained results using the modified KL distance, compared to the classical KL distance approximated with monte carlos simulation (KL-MCS) as in [20]. However, it is worth noting that in [20], the authors used another database and treated a different pathology. To illustrate the fact that the advantage of our method is not due to the difference in the database or the pathology whatsoever, the comparative results in table 3 are based on our implementation of both methods using the same database, the same pathology, and also used the parameters adjusting method. As it can be seen from table 3, the new metric provides a better sensitivity and specificity of 94% and 99% respectively.

Simulations with Bhattacharyya distance are made to demonstrate and reinforce that not only KL distance is influenced by the triangle inequality violation. Results are presented in table 4. Significant enhancement in the performance is obtained by using the new version. It attained 4% in terms of sensitivity, and 2 % in term of specificity.

Receiver operating characteristics (ROC) curve or graphs are another useful tool to visualize the system performance. It evaluates the area under curve (AUC), where the performance is perfect when the area is 1. In

Table 3: GMM-SVM Results using classical and modified KL.

	Distances	
Performance	KL-MCS [20]	Modified KL
Sensitivity	92%	94%
Specificity	96%	99%
Accuracy	94%	96.5%

Table 4: GMM-SVM Results using classical and modified Bhattacharyya.

	Distances	
Performance	Bh	Modified Bh
Sensitivity	89%	93%
Specificity	96%	98%
Accuracy	92.5%	95.5%

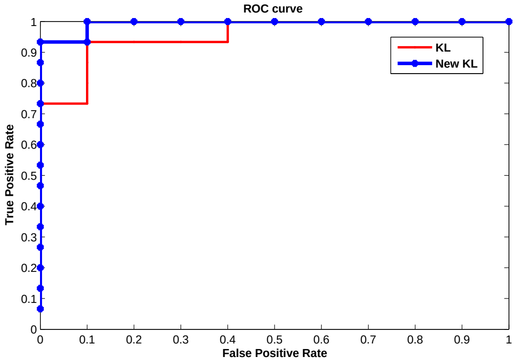

Fig. 3: ROC curve for GMM-SVM using Kullback-Leibler.

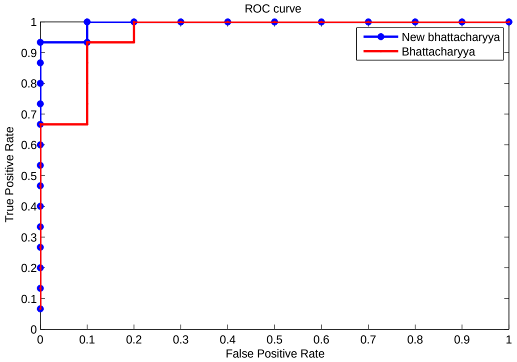

Fig. 4: ROC curve for GMM-SVM using Bhattacharyya distance

this case, the new metric presents the best results with 0.99 of AUC. It is shown in figure 3 and figure 4 by blue curve. Comparing with the old versions, represented in red, there is an improvement of 0.04 and 0.03 obtained respectively for Kullback-Leibler and Bhattacharyya distances.

Knowing that we have used the weighted versions, best detection rate is obtained when β take its value between $[0.6, 0.9]$. This means that the first term (distance between means) is more significant than the second term (distance between covariance matrices). So the Riemannian distance for the covariance matrix did not had a significant impact, and another distance matrix may be more efficient .

5 Conclusion and future work

In this work, we presented a method based on the combination of GMM and SVM classifiers. We focussed in our method on a better choice of distance metric in the RBF kernel. The used distance metrics are modified versions of the kullback leibler and bhattacharyya distances, that do in fact satisfy all metric axioms, unlike their classical counterparts.

The obtained results confirm the efficiency of the RBF kernel as a tool to measure the degree of similarity between objects, and the choice of the distance metric that respects all axioms, especially the triangle inequality, further improve the capacity to distinguish between GMMs models. Specifically, the results show that at least 2% and 4% of improvement in term of sensitivity are achieved when applying the new distance over the use of the classical kullback leibler and bhattacharyya distances respectively.

The promising results motivate us to improve this work. Future work may concern the use of another database in order to assess the independence of our method from the used database. We may also work on the detection and classification of other types of pathologies.

Acknowledgement

We want to thank Mr Manfred putzer and his team for their effort to create the database.

References

[1] A.Ghio,S.Dufour,M.Rouaze,V.Bokanowski,G.Pouchoulin, A.Giovanni, Mise au point et valuation d’un protocole d’apprentissage de jugement perceptif de la svrit de dysphonies sur de la parole naturelle. rev laryngol otol rhinol,**132**,1–9 (2011).

[2] G.Antoine, Y.Pirng, R.Joana, G.Marie-Dominique, T.Bernard, O.Maurice, Analyse objective des dysphonies avec l’appareillage EVA. Fr ORL,90 : 183 (2006)

[3] G.Darcio, Silva, C.Luis. Oliveira and A.Mario, Jitter Estimation Algorithms for Detection of Pathological Voices. Hindawi Publishing Corporation, EURASIP Journal on Advances in Signal Processing Volume 2009, Article ID 567875, 9 pages.

[4] V.Miltiadis, S.Yannis, Voice Pathology Detection Basedeon Short-Term Jitter Estimations in Running Speech. Folia Phoniatr Logop **61**,153–170 (2009).

[5] R.Sonu, K.Sharma, Disease Detection Using Analysis of Voice Parameters. International Journal of Computing Science and Communication Technologies,**4**,(2012)

[6] K.Jacques,P.Manfred, J.Manfred, Correlates of Varying Vocal Fold Adduction Deficiencies in Perception and Production: Methodological and Practical Considerations. Folia Phoniatr Logop **56**,305–320 (2004).

[7] N.Sanez-Lechon, I.Juan,O.Victor,G.Pedro, Methodological issues in the development of automatic systems for voice pathology detection. Biomedical Signal Processing and Control **1**, 120-128 (2006).

[8] I.Juan Member IEEE, G.Pedro Member IEEE, and B.Manuel Member IEEE. Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters.IEEE Transaction on biomedical engineering, **53**, NO. 10, (2006).

[9] L.Ji Yeoun, A two-stage approach using Gaussian mixture models and higher-order statistics for a classification of normal and pathological voices. EURASIP Journal on Advances in Signal Processing,**252**, (2012).

[10] M.David, L.Eduardo, O.Alfonso,M.Antonio, and V.Jesus, Voice Pathology Detection on the Saarbrucken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit.D.T. Toledano et al. (Eds.): IberSPEECH 2012,**328**, 99-109, (2012).

[11] I.Juan,G.Pedro, S.Nicols, B.Manuel, C.Fernando, and A.Miguel, Support Vector Machines Applied to the Detection of Voice Disorders. Springer-Verlag Berlin Heidelberg,219-230, (2005).

[12] C.Wenxi, Member, IEEE. Ce P, Xin Z, Member, IEEE. Baikun W. and Daming W, Member, IEEE. SVM-based Identification of Pathological Voices.Proceedings of the 29th Annual International Conference of the IEEE EMBS Cit Internationale, Lyon, France August 23–26, 2007.

[13] E.Nafise, A.Farshad, T.Farhad, Support vector wavelet adaptation for pathological voice assessment. Computers in Biology and Medicine **41**,822–828 (2011).

[14] E.Nafise, A.Farshad, Wavelet adaptation for automatic voice disorders sorting. Computers in Biologyand Medicine **43**, 699–704 (2013).

[15] W.Jianglin, J.Cheolwoo Member, IEEE, Vocal Folds Disorder Detection using Pattern Recognition Methods. Proceedings of the 29th Annual International Conference of the IEEE EMBS Cit Internationale, Lyon, France August 23–26,(2007).

[16] B.Mohamed,Pattern recognition methods applied to respiratory sounds classification intonormal and wheeze classes.Computers in Biology and Medicine **39**, 824–843, (2009).

[17] X.Zhou, X.Zhuang, S.Yan Chang,M.Hasegawa-Johnson, T.S Huan, Sift-bag kernel for video event analysis. In Proceedings of the 16th ACM international conference on Multimedia 229-238.(2008).

[18] HO, Pedro J. Moreno Purdy P. et VASCONCELOS, Nuno. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. Proc Adv Neural Inf Process Syst, **16**, 1385-1392, (2004).

[19] W.Xiang, Z.Jianping, and Y.Yonghong, Discrimination Between Pathological and Normal Voices Using GMM-SVM Approach. Journal of Voice, **25** 38–43 (2011).

[20] V.Evaldas,V.Antanas, G.Adas, B.Marija, U.Virgilijus, Exploring similarity-based classification of larynx disorders from human voice Speech Communication science direct 601-610,(2012)

[21] T.Karim,P.Frank, A Note on Metric Properties for Some Divergence Measures:The Gaussian Case.JMLR: Workshop and Conference Proceedings **25**,(2012)

[22] http://www.stimmdatenbank.coli.uni-saarland.de.

[23] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.

[24] W.M.Campbell Member IEEE, D.E.Sturim Member IEEE, and D.E.Reynolds Senior Member IEEE. Support Vector Machines Using GMM Supervectors for Speaker Verification. IEEE signal processing letters,**13** (2006).

[25] Julien Ah-Pine, Normalized Kernels as Similarity Indices. PAKDD 2010, Part II, LNAI 6119, pp. 362373, 2010. Springer-Verlag Berlin Heidelberg.

[26] John R. Hershey, Peder A. Olsen, Steven J. Rennie. Variational Kullback-Leibler Divergence for Hidden Markov Models.Automatic Speech Recognition and Understanding, 2007. ASRU. IEEE Workshop on.

[27] John R. Hershey, Peder A. Olsen, Steven J. Rennie. Variational Bhattacharyya Divergence for Hidden Markov Models. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

[28] Don H. Johnson and Sinan Sinanovic. Symmetrizing the Kullback-Leibler Distance. IEEE Transactions on Information Theory 2000.

[29] Tversky A. Gati I. Similarity, Separability, and the Triangle Inequality. Psychological Review,**89**,123-154,1982.

[30] B.Saaid, A.Dan, Simovici and Catalin Z. The Impact of Triangular Inequality Violations on Medoid-Based Clustering, LNAI 6804, 280-289, (2011)

[31] G.Sfikas, C.Constantinopoulos, A.Likas and N.P.Galatsanos, An Analytic Distance Metric for Gaussian Mixture Models with Application in Image Retrieval W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, 835-840,(2005).

[32] I. M. M. El Emary, M. Fezari, and F. Amara, Towards Developing a Voice Pathologies Detection System.Journal of Communications Technology and Electronics, 2014, Vol. 59, No. 11, pp. 12801288.

Fethi AMARA received his MSc degree in advanced telecommunications from Badji Mokhtar Annaba University in 2009. since 2010 he has joined the Automatic and Signals Laboratory of Annaba university in research work in speech recognition,signal processing algorithms and machine learning.

Mohamed Fezari is an associate professor in electronics and computer architecture at the University of Badji Mokhtar Annaba, Algeria. He got a Bachelor degree in electrical engineering from university of Oran, 1983. He got an MSc degree in computer science from University of California Riverside, 1987. He holds a PhD degree in electronics from the University of Badji Mokhtar Annaba, 2006. His research interests include speech processing, DSP, micro-controller, microprocessor, robotics and human machine interaction, and rehabilitation.

Hocine BOUROUBA received his PHD degree in electronics from the University of Badji Mokhtar Annaba, 2009. Actually, he is professor at Guelma university. His research interests include speech processing,biometrics, face recognition.