Automatic Detection Of Vocal Fold Paralysis and Edema

Maria Marinaki[1], *Constantine Kotropoulos*[1], *Ioannis Pitas*[1],
Nikolaos Maglaveras[2]

[1] Artificial Intelligence & Information Analysis Lab, Department of Informatics
[2] Lab of Medical Informatics, Medical School
Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, GREECE
m_marinaki@yahoo.gr, costas@zeus.csd.auth.gr, pitas@zeus.csd.auth.gr, nicmag@med.auth.gr

Abstract

In this paper we propose a combined scheme of linear prediction analysis for feature extraction along with linear projection methods for feature reduction followed by known pattern recognition methods on the purpose of discriminating between normal and pathological voice samples. Two different cases of speech under vocal fold pathology are examined: vocal fold paralysis and vocal fold edema. Three known classifiers are tested and compared in both cases, namely the Fisher linear discriminant, the K-nearest neighbor classifier, and the nearest mean classifier. The performance of each classifier is evaluated in terms of the probabilities of false alarm and detection or the receiver operating characteristic. The datasets used are part of a database of disordered speech developed by Massachusetts Eye and Ear Infirmary. The experimental results indicate that vocal fold paralysis and edema can easily be detected by any of the aforementioned classifiers.

1. Introduction

Speech processing has proved to be an excellent tool for voice disorder detection. Among the most interesting recent works are those concerned with Parkinson's Disease (PD), multiple sclerosis (MS) and other diseases which belong to a class of neuro-degenerative diseases that affect patients speech, motor, and cognitive capabilities [1, 2]. Such studies are based on the special characteristics of speech of persons who exhibit disorders on voice and/ or speech. They aim at either evaluating the performance of special treatments (i.e. LSVT [2, 3]) or developing accessibility in communication services for all persons [4]. Thus, it would possibly be a matter of great significance to develop systems able to classify the incoming voice samples into normal or pathological ones before other procedures are further applied.

In this paper, we are concerned with vocal fold paralysis and vocal fold edema, which are both associated with communication deficits that affect the perceptual characteristics of pitch, loudness, quality, intonation, voice-voiceless contrast etc, having similar symptoms with PD and other neuro-degenerative diseases. The main causes of vocal fold paralysis are usually either several surgical iatrogenic injuries or a glitch in the recurrent laryngeal nerve or possibly a lung cancer [5], while malfunction at the vocal folds due to edema is usually caused by more trivial reasons such as mild laryngeal injuries, common infectious diseases that affect the respiratory system, or allergies in drugs. We demonstrate that effective classification between normal voice samples and voice samples from persons who suffer from either vocal fold paralysis or vocal fold edema can be achieved, as long as the most significant characteristics of pathological voice are retained. In either case, a two-class pattern recognition problem is essentially studied. Closely related previous works are the detection of vocal fold cancer [6], where a Hidden Markov Model (HMM)-based classifier was employed and the binary classification between normal subjects and subjects suffering from different pathologies in [7], where Mel frequency cepstral coefficients and pitch were used as features for classification that was performed by the linear discriminant classifier, the nearest mean classifier, and classifiers based on Gaussian mixture models or HMMs. The assessment of the classifiers in [7] has also been done on the database of disordered speech developed by Voice and Speech Lab of Massachusetts Eye and Ear Infirmary (MEEI) as in the present work. Three parameters namely the number of discrimination, the level of clustering, and the average clustering were assessed for disease discrimination based on acoustic features in [8].

In this paper, we assess the performance of Fisher's linear classifier, the K-nearest neighbor classifier and the nearest mean one. We are not interested in the detection of pathological speech as in [7], but in the assessment of the discriminatory capability of the aforementioned classifiers for particular vocal fold pathologies. The main contribution of the paper is in the appropriate feature extraction and reduction methods, the design of the aforementioned classifiers, and the thorough assessment of their classification ability by employing the probabilities of false alarm and detection or the receiver operating characteristic (ROC) curve.

The outline of the paper is as follows. The datasets used in the experiments are presented in Section 2. The feature extraction and reduction techniques as well as the design of classifiers employed are described in Section 3. Experimental results are demonstrated in Section 4, and conclusions are drawn in Section 5.

2. Subjects-Datasets

In the first experiment, the dataset contains recordings from 21 males aged 26 to 60 years who were medically diagnosed as normals and 21 males aged 20 to 75 years who where medically diagnosed with vocal fold paralysis. In the second experiment 21 females aged 22 to 52 years who were medically diagnosed as normals and 21 females aged 18 to 57 years who where medically diagnosed with vocal fold edema served as subjects. The subjects might suffer from other diseases too, such as hyperfunction, ventricular compression, atrophy, teflon granuloma, etc. All subjects were assessed among other patients and normals at the MEEI [9] in different periods between 1992 and 1994. Two different kinds of recordings were made in each session: in the first recording the patients were called to articulate the sustained vowel "Ah" (/a/) and in the second one to read the "Rainbow Passage". The former is the one concerned with the present work. Therefore, all procedures were applied to voiced speech frames far away from transition periods. The recordings were made at a sampling rate of 25 Hz in the pathological case, while at 50 Hz in the normal case. In the latter case, the sampling rate was normalized to 25 Hz by subsampling.

3. Method description

First a 16 msec Hamming moving window is employed to obtain frames from each voice recording for short-term voice analysis purpose [10]. From each recording, two central frames are selected among the ones that belong to the most stationary portion of the sustained speech signal as is proposed in [8, 11]. This selection yields 42 frames of pathological voice and another 42 frames of normal voice. Thus, we obtain 84 frames in total for each experiment. The frames exhibit an overlap of 50%.

The feature vector extraction is performed via short-term linear prediction of order 14 [10]. The LP model of order 14 is regarded as a good choice. It has been reported that the use of more than 14 LPCs does not improve significantly the discrimination of laryngeal diseases [8, 12]. Let $\mathbf{x}_j$, $j = 1, 2, \ldots, 84$, be the jth (14×1) feature vector, whose elements are the linear prediction coefficients. The dimensionality of the feature space is then reduced by using principal component analysis (PCA) [13]. This is done as follows. Let $\tilde{\mathbf{x}}_j$ be the normalized feature vector independent of class:

$$\tilde{\mathbf{x}}_j = \mathbf{x}_j - \mathbf{m} \tag{1}$$

where $\mathbf{m}$ is the class-independent mean feature vector. Let also N be the number of vectors of the whole feature space (i.e., $N = 84$ for each experiment). The covariance matrix $\mathbf{S}$ of the overall feature space is:

$$\mathbf{S} = \frac{1}{N} \sum_{j=1}^{N} \tilde{\mathbf{x}}_j \tilde{\mathbf{x}}_j^T. \tag{2}$$

In PCA, the p eigenvectors $\mathbf{e}_k$, $k = 1, 2, \ldots, p$ with $p \leq 14$ that correspond to the p largest eigenvalues of $\mathbf{S}$ are computed. By experiments we found that more than two principal components did not improve significantly the performance of the classifiers employed afterwards. Thus, a 2×14 projection matrix $\mathbf{P}$ is built, which is used to project the 14th dimensional feature vectors to a two-dimensional plane as follows:

$$\hat{\mathbf{x}}_j = P\, \tilde{\mathbf{x}}_j \tag{3}$$

where

$$\mathbf{P} = \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{bmatrix}. \tag{4}$$

For both experiments the aforementioned pre-processing step enables a visualization of the projected 2-D feature vectors on the two-dimensional plane, as can be seen in Figure 1(a) for the first experiment, and (b) for the second one.

Subsequently three two-class classifiers are described. The first classifier is designed as follows. The available data are split into a training set and a disjoint test set. 72 feature vectors (36 from normal voice samples and 36 from the pathological voice samples) constitute the training set. The remaining 12 feature vectors (6 from each class of voice samples) are being retained to serve as test feature vectors. We compute first the within-class covariance matrix $\mathbf{S}_w$

$$\mathbf{S}_w = \sum_{i=1}^{2} p_i \mathbf{S}_{w\,i} \tag{5}$$

where p_i is the a-priory probability of class C_i, $i = 1, 2$. We have assumed equal a-priori probabilities. The class dependent covariance matrix $\mathbf{S}_{w\,i}$ is computed by

$$\mathbf{S}_{w\,i} = \frac{1}{N_i} \sum_{\mathbf{x}_j \epsilon C_i} \hat{\mathbf{x}}_j \hat{\mathbf{x}}_j^T \tag{6}$$

where N_i $(i = 1, 2)$ is the number of training feature vectors that belong to class C_i. (Here, $N_1 = N_2 = 36$.) Secondly, we compute the mean vector for each class, $\hat{\mathbf{m}}_1$ and $\hat{\mathbf{m}}_2$, respectively. Fisher's classifier is defined by the following vector $\mathbf{w}$ [14, 13]:

$$\mathbf{w} = \mathbf{S}_w^{-1} \left(\hat{\mathbf{m}}_1 - \hat{\mathbf{m}}_2\right). \tag{7}$$

The projection of a 2-D feature vector onto $\mathbf{w}$ guarantees the best separation between the two classes. The classifier is optimal if the feature vectors in the two classes are Gaussian distributed [13].

Next, two other classifiers are discussed. The first classifier is based on the K-nearest neighbor (K-NN) method applied as follows: for each feature vector of the test set we peak the feature vectors of the training set within a circle around it, whose radius is increased until at least K training feature vectors are enclosed, the K-nearest ones. The test sample is assigned to the class where the majority of the training feature vectors belongs to. The second classifier depends on the class-dependent mean vector computed from the training samples, employs the distance of each test feature vector from the mean vector of each class and assigns the test sample to the class of the nearest mean vector.

4. Results

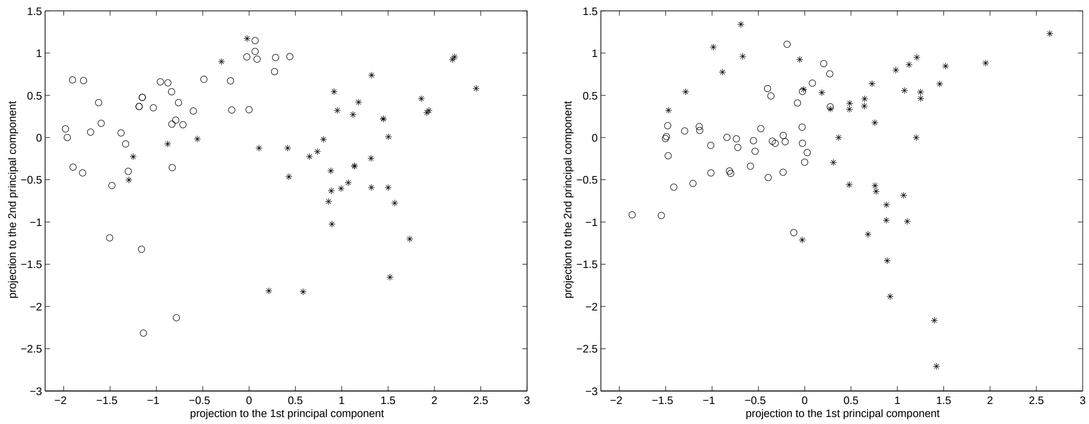
The assessment of Fisher's classifier was done via the ROC curve [13]. While a threshold value t is moving from a minimum value towards a maximum one across the projection axis, the pathological samples that are successfully detected are counted and the probability of detection $P_d(t)$ is computed for each threshold value t as

$$P_d(t) = \frac{\text{\# correctly classified pathological samples}}{\text{\# pathological samples}} \tag{8}$$

where # stands for number. The probability of false alarm $P_f(t)$ is estimated by

$$P_f(t) = \frac{\text{\# normal samples misclassified as pathological ones}}{\text{\# normal samples}}. \tag{9}$$

By plotting the various pairs $(P_f(t),\, P_d(t))$ the ROC curve is obtained. We measured the aforementioned probabilities both in the training and the test set. To cope with the lack of data we repeated the procedure 75 times by randomly selecting a training set of 72 samples (36 samples from each class) and a test set of 12 samples (6 samples from each class). The average ROC curves for the detection of vocal fold paralysis are plotted in Figure 2(a) and (b) for the training set and the test set, respectively. The average ROC curves for the detection of vocal fold



(a) (b)

Figure 1: The whole 2-D feature space for (a) the first experiment concerned with vocal fold paralysis and (b) the second experiment concerning vocal fold edema. (Each normal feature vector is represented with an ‘o’, while each pathological feature vector is represented by a ‘*’.)

edema are shown in Figure 3(a) and (b) for the training set and the test set, respectively. It can be seen that a probability of detection close to 85% can be achieved for a probability of false alarm equal to 10% in the case of vocal fold paralysis. At the same probability of false alarm the probability of detection for vocal fold edema is approximately 73 %.

For the two other classification algorithms cross-validation was applied, with 756 different training sets of 72 (36 normal and 36 pathological) randomly selected samples. The remaining 12 samples in each of the 756 cases served as test set. The P_f and P_d were computed in each repetition of the experiment and finally the mean P_f and P_d values were computed. Tables 1 and 2 summarize the results for the K-NN classifier (for K=3,5,7) and the nearest mean classifier for the first and the second experiment, respectively.

Table 1: Mean P_f and P_d for the K-NN algorithm (for $K = 3, 5, 7$) and the nearest mean classifier for the detection of vocal fold paralysis.

Method	**mean P_f**	**mean P_d**
3-NN	0.0851	0.8402
5-NN	0.0353	0.8331
7-NN	0.0295	0.8329
Nearest mean	0.0745	0.8571

From the inspection of Table 1 and 2 it is seen that the nearest mean classifier attains a better performance than the K-NN classifiers considered. More significant gains have been obtained by the nearest mean classifier in the case of vocal fold edema detection. The performance deterioration of all classifiers in the case of vocal fold edema could be attributed to the fact that linear prediction analysis is less effective for low pitched voice (e.g. women, children) [10].

Table 2: Mean P_f and P_d for the K-NN algorithm (for $K = 3, 5, 7$) and the nearest mean classifier for the detection of vocal fold edema.

Method	**mean P_f**	**mean P_d**
3-NN	0.0679	0.7590
5-NN	0.0580	0.7286
7-NN	0.0545	0.7354
Nearest mean	0.1175	0.8325

5. Conclusions

It has been demonstrated by experiments, that efficient detection of voice disorders can be achieved by Fisher’s linear discriminant, K-NN, and the nearest mean classifier for vocal fold paralysis. Slightly worse results have been reported for vocal fold edema detection. The spectral characteristics extracted by linear prediction analysis of order 14 combined with principal component analysis of order 2 for feature reduction have been proved to be very efficient for the aforementioned classification tasks.

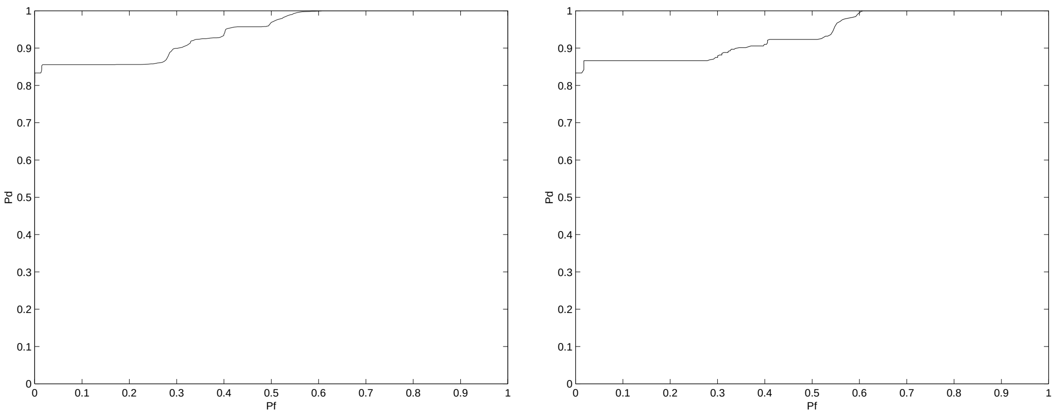
6. Acknowledgement

This work has been supported by the FP6 European Union Network of Excellence “Multimedia Understanding through Semantics, Computation and LEarning” (IST-2002-2.3.1.7).

7. References

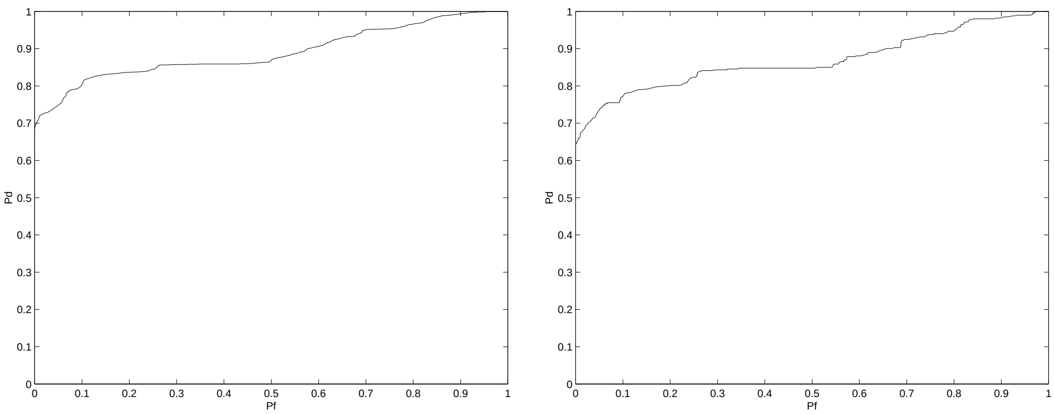
[1] F. Quek, M. Harper, Y. Haciahmetoglou, L. Chen, and L. O. Ramig, “Speech pauses and gestural holds in Parkinson ’s Disease,” in *Proc. 2002 Int. Conf. Spoken Language Processing*, 2002, pp. 2485–2488.

[2] L. Will, L. O. Ramig, and J. L. Spielman, “Application of Lee Silverman Voice Treatment (LSVT) to individuals with multiple sclerosis, ataxic dysarthria and stroke,” in



(a) (b)

Figure 2: The average ROC curves for the detection of vocal fold paralysis from pathological voice samples in (a) the training set and (b) the test set.



(a) (b)

Figure 3: The average ROC curves for the detection of vocal fold edema from pathological voice samples in (a) the training set and (b) the test set.

Proc. 2002 Int. Conf. Spoken Language Processing, 2002, pp. 2497–2500.

[3] J. L. Spielman, L. O. Ramig, and J. C. Borod, "Oro-facial changes in Parkinson's Disease following intensive voice therapy (LSVT)," in *Proc. 2002 Int. Conf. Spoken Language Processing*, 2002, pp. 2489–2492.

[4] V. Parsa and D. G. Jamieson, "Interactions between speech coders and disordered speech," *Speech Communication*, vol. 40, no. 7, pp. 365–385, 2003.

[5] "http://www.emedicine.com/ent/byname/vocal-fold-paralysis-unilateral.htm," .

[6] L. Gavidia-Ceballos and J. H. L. Hansen, "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection," *IEEE Transactions on Biomedical Engineering*, vol. 43, pp. 373–383, 1996.

[7] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proc. Engineering Medicine and Biology Symposium 02*, 2002, vol. 1, pp. 182–183.

[8] M. O. Rosa, J. C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," *IEEE Transactions of Biomedical Engineering*, vol. 47, pp. 96–104, 2000.

[9] *Voice Disorders Database Version 1.03*, Voice and Speech Laboratory, Massachusetts Eye and Ear Infirmary, Boston MA, Kay Elemetrics Corp. 1994.

[10] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete Time Processing of Speech Signals*, MacMillan Publishing Company, New York, N. Y., 1993.

[11] R. A. Prosek, A. A. Montgomery, B. E. Walden, and D. B. Hawkins, "An evaluation of residue features as correlates of voice disorders," *Communication Disorders*, vol. 20, pp. 105–107, 1987.

[12] S. B. Davis, "Acoustic characteristics of normal and pathological voices," *Speech and Language: Advances in Basic Research and Practice*, vol. 1, pp. 271–335, 1979.

[13] K. Fugunaka, *Introduction in Statistical Pattern Recognition*, Academic Press, N.Y., 1990.

[14] R. J. Schalkoff, *Pattern Recognition: Statistical, Structural and Neural Approaches*, MacMillan Publishing Company, N. Y., 1993.