EGG SIGNAL CLASSIFICATION USING PRINCIPAL COMPONENT ANALYSIS

Subramanya Nayak G.[1], Dr. Puttamadappa C.[2]
[1]Manipal Institute of Technology, Manipal- 576104,
[2]Sri Jagadguru Balagangadharanatha Institute of Technology, Bangalore
E-mail: [1]gs_nayak@rediffmail.com

Abstract

The method is used to register the laryngeal behavior indirectly by measuring change in the electrical impedance across the throat during speak or voice. In this Electroglottography (EGG) signal acquisition, the electrodes are made of steel. They have the form of rectangles covering an area of 10.75 cm2. It is designed as a ring electrode encircling each of the two other electrodes. The electrodes are mounted on a flexible band whose length is adjusted to hold the electrodes in a steady position and to still allow the subject to comfortably speak and breathe naturally. The electrodes are mounted on a small holder which is pressed against the throat by hand. A signal generator supplies an AC sinusoidal current usually ranging from 2 MHz. The RF carrier signal is amplitude modulated by the modulating speech/ voice signal and the demodulated signal is extracted. The variations in the signal correspond to the vocal fold abduction/laryngeal movement. For normal and pathology conditions, the results are recorded. These values form a feature vector, which reveals information regarding pathology. Principal Component Analysis technique (PCA) is used for classification, giving successful results for the specific data set considered.

Keywords: Electroglottography, Principal Component Analysis.

I. INTRODUCTION

The voice pathology is very common in all over the world. In the current study the vocal fold abduction /laryngeal movement of normal and pathology patients have been recorded in terms of demodulated signals and MATLAB @6.1 supported Principal Component Analysis technique (PCA) is used for classification, giving successful results for the specific data set considered.

II. INSTRUMENTATION

The electrodes are made of steel. They have the form of rectangles covering an area of 10.75 cm2. It may be designed as a separate electrode or as a ring electrode encircling each of the two other electrodes. The electrodes are mounted on a flexible band whose length may be adjusted to hold the electrodes in a steady position and to still allow the subject to comfortably speak and breathe naturally. The electrodes are mounted on a small holder, which is pressed against the throat by hand. A signal generator supplies an AC sinusoidal current usually ranging from 300 KHz to 5 MHz as shown in Figure.1.

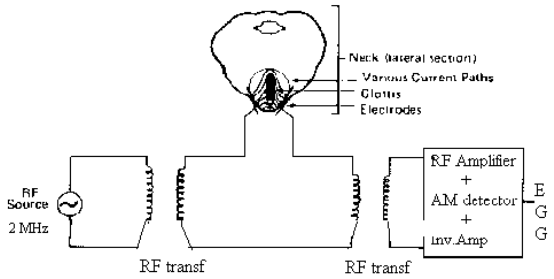

Fig. 1. Instrumentation for EGG

The frequency selected for the above test is 2 MHz. This frequency is sufficiently high, so that the current capacitancively bypasses the less conductive skin layer without the use of additional conductive paste [1]. The generator may produce constant voltage or constitute constant current

source [1]. The supplied current is different for each particular device, but is not stronger than several milliamperes. The voltage between the electrodes depends on the tissue impedance [1-3]. The power dissipation of only several microwatts occurs at the level the subject's vocal folds. An integral part of the electroglottographic signal is the varying component generated by the vertical movement of the whole larynx. Therefore, the signal of rapid movements of the vocal folds is superimposed on the signal produced by the slower movements of the other structures. Fourcin & Abberton proposed the name Gx for the waveform of larynx movement and the name Lx for the vibration component. The Gx component originates, for example can be observed in swallowing, but it is caused by the vertical movement of the larynx, which is related to the voice quality setting of the raised/ lowered larynx. Gx to calculate vocal fold abduction [1] The DC offset changes (Gx) can be evened out because, the effects of the varying larynx height are compensated by the use of additional electrodes or high pass filtering of the registered signal. The sensing electrode detects the current as it passes through the skin and the throat .The percentage of amplitude modulation of the received signal reflects the percentage change in 3 tissue impedance in the current's path. The output from the second RF transformer is then amplified using the above RF amplifier circuit. The output

is demodulated using a diode detector circuit (Gx + Lx) .The output is then amplified using a OPAMP inverting amplifier (Gx) as shown in Fig. 2. The output spectra were recorded.

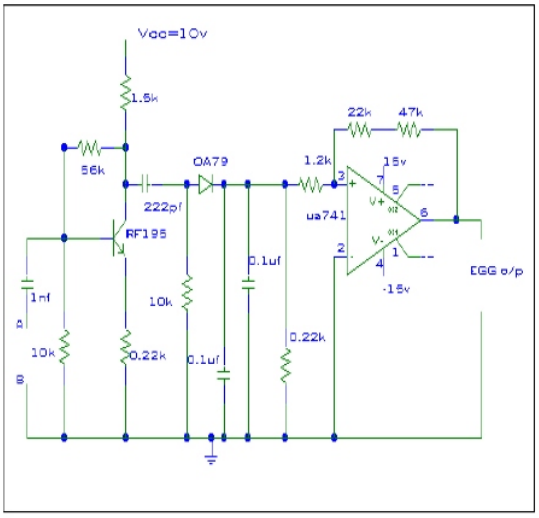

Fig. 2. RF Amplifier and Detector

III. PCA ANALYSIS

In the development of diagnostic or analytical methods for routine applications, only a small amount of data from a very large quantity of data is made use of from the point of view of expediency. This may lead to wrong interpretation and consequent faulty decision making, especially in clinical applications where personal judgment of the clinician may influence the decision. This kind of subjective evaluation of data can be avoided when we have enough data (e.g. a large number of spectra, each consisting of several data points) by appropriate mathematical / statistical analysis. Almost always, the enormous amount of data could be understood in terms of a much smaller number of components, called principal components or factors. This is equivalent to the situation, where any number of vectors in 3-dimentional spaces can be expressed in terms of 3 unit vectors and characteristic numbers for each vector. The n number of spectra with p data points each may be expressed in terms of a much smaller number of components or factors each with p data points. The identification of these unique factors is known as Principal Component Analysis (PCA) [4,5].

In real samples, there are usually many different variations that make up a spectrum: the constituents in the sample mixture, inter-constituent interactions, instrument variations such as detector noise, changing environmental conditions that affect absorbance, and differences in sample handling. Yet, even with all these complex changes occurring, there should be some finite number of independent variations occurring in the spectral data. Hopefully, the largest variations in the calibration set would be changes in the spectrum due to different concentrations of the constituent of the mixtures. If it were possible to calculate set of "variation spectra" that represented the changes in the intensities at all the wavelengths in the spectra, then this data could be used instead of raw spectral data for building the calibration model. These should be fewer common variations than the number of calibration spectra (in most cases), and thus, the number of calculations for the calibration equations will be reduced as well.

Presumably, the "variation spectra" could be used to reconstruct the spectrum of a sample by multiplying each one by an appropriate constant scaling factor and adding the results together until the reconstructed spectrum closely matches the sample spectrum. Obviously, each spectrum in the calibration set would have a different set of scaling constants for each variation since the concentrations of the constituent is different. Therefore, the scaling constant of each "spectrum" that must be added to reconstruct the unknown data should be related to the concentration of the constituents.

The "variation spectra" are often called "eigenvectors" (also called spectral loading, loading vectors, principal components or factors), from the methods used to calculate them. The scaling constants used to reconstruct the spectra are generally known as "scores".

Since the calculated eigenvectors came from the original calibration data, they must be somehow relating to the concentrations of the constituents that make up the samples. The same loading vectors can be used to predict "unknown" samples; thus only difference between spectra of samples with different constituent concentrations is the fraction of each loading added (scores).

A. Signal Preprocessing

In the present study, we have used MATLAB @6.1 software tool is used to carryout smoothing, mathematical and statistical analysis.

i) Smoothing

Highly noisy spectrum can be smoothened to a great extent using various smoothing functions. These include Fourier-domain smoothing, binomial smoothing etc. Fourier transforming the data, applying a filter function and then inverse Fourier transforming the data, accomplishes Fourier smoothing.

IV. CLASSIFICATION

This method can classify samples into well-defined groups or categories based on a training set of similar samples without prior knowledge of the actual composition of group of training samples. The aim of this analysis is to

identify unknown sample. The spectrum of sample is compared against the model to determine if it matches the training data for the model [6]. If the training set was constructed from spectra of samples that were of known quality, the model can accurately predict if the sample is of same quality by matching the spectrum and giving a "YES" or "NO" answer.

A. Classification Parameters

i) Eigenvectors And Scores

Eigenvectors are the spectral equivalents of principle components of the sample and scores corresponding to contribution of each principle component to a given sample. Multiplying the eigenvectors with the scores for that sample and adding the product for all scores can reconstruct each sample spectrum.

ii) Residual Errors Or Spectral Residual

When each sample is predicted, a set of scores is found that best fits the model loading vectors to the sample spectrum. By using the calculated scores and calibration loading vectors, a new model reconstructed spectrum can be calculated. This new spectrum is what the PCA model thinks the sample spectrum look like. The residual errors or spectral residual is the difference between this spectrum and the actual prediction spectrum.

iii) Mahalanobis Distance

It is very sensitive to inter variable changes in calibration data. The distance is measured in terms of standard deviations from the mean of the training samples The values give a statistical measure of how well the spectrum of unknown sample matches the original training spectra. Typical discrimination model is as shown in Figure 3.

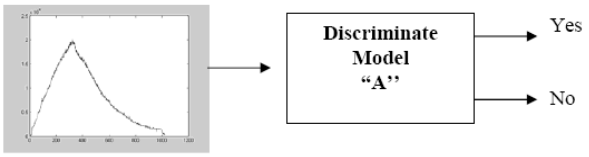

Fig. 3. Typical Discrimination Model

In PCA analysis, twenty spectra each from certified normal and pathology samples (details of spectra used are listed in table 1) were combined to see the best approach to prepare calibration sets in the two classes. It is seen that the eigenvalues decrease very rapidly and are almost zero after 7 to 8 factors, and about 99% of total spectral contribution come from these factors only as shown in Table 2. From the Table 2 and Figure 4, it can be seen that the eigenvalues decrease very rapidly and are almost zero after eight factors and also 100% of the total spectral contribution come from these eight factors. This can be further confirmed by using an appropriate number of factors from the model set and regenerating the spectrum of any sample. The difference between the observed and regenerated spectrum, expressed as residual errors squared sum can be used as a measure of desired number of factors, as well accuracy of the model. In the present analysis it is found that four factors contributed to about 98% of total variance, and these four factors completely describe the spectra. The higher factors were found only to account for variations in day-to-day runs, noise etc., and did not improve sum of squared spectral residuals, or other parameters like average predicted Mahalanobis distance [7]. All final calculations were thus carried out with using only four factors. In this case, the statistical parameters like spectral residuals, Mahalanobis distances etc. are used for discrimination between normal and pathology cases.

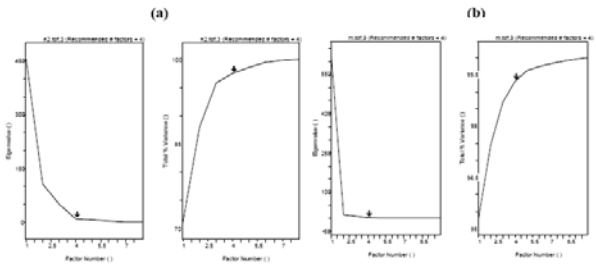

Fig. 4. PCA Eigen values and total % variance for a model set of (a) 20 normal spectra (b) 20 pathology spectra.

Table 1. Spectral Details

Spectral No	Sample Type	Mean age	Histopathology	Signal
1- 20	Normal Standard Set	48 ± 6.1	Uninfected area - Normal	-
21-40	Pathology Standard set	48 ± 6.1	Laryngeal movement / vocal fold abduction	-
41-80	Normal Test Set	44 ± 7.5	Uninfected area - Normal	Normal
81-117	Pathology Test Set	44 ± 7.5	Laryngeal movement / vocal fold abduction	Pathology

Table 2. Factor number with corresponding eigen values and total percentage variance for 20 normal and 20 pathology calibration spectra

Factor Number	Eigen value	Total % Variance
1	453.980101	71.1774449
2	107.09752	88.0199317
3	50.1468976	95.841953
4	10.7986678	97.5396094
5	7.3732443	98.7601175
6	5.16731475	99.5704086
7	2.36394707	99.8509443
8	0.01421524	100

The Mahalanobis distance is normally expressed in units of standard deviation. For classification of oral tissues, we have employed the Mahalanobis distance (M-distance) and spectral residual (the residual error squared sum) as the criteria. The M-distance can be represented by:

D2 = (S test) M –1(S test) 1,

Where S test is the vector of scores and sum of squared spectral residuals for a given test sample, and M is given by S¢S/(n-1), where S contains the corresponding parameters for the calibration set (n standards).

Since the Mahalanobis distance is a standard deviation, a distance of »2 for a sample corresponds to a 5% probability of the sample belonging to the standard set, and higher distances will have still less probability. There are two main advantages in using D2 as a discriminating parameter. As seen from the equation, D2 explicitly accounts for any correlations between the variables, namely scores of factors. By fixing an upper limit for inclusion in any class represented by the standard calibration set for that class, we can possibly achieve any desired level of discrimination for staging.

We have made match mismatch tables of calibration set as well as test set samples (normal and malignant spectra) by comparing these with the normal calibration set considering the Mahalanobis distance (M-distance) of »3. The results are listed in tables 3, 4, 5 and 6.

Table 3. Retrospective test of normal calibration set samples against calibration set of normal samples. Mean M distance for normal calibrated set is 1.0602±0.51 and Mean Spectral residual is 0.7569±0.028. Acceptance value is fixed to twice the mean M distance of normal calibration set.

Spectral number	Match	M. Distance	Limit test ***	Spectral residual
1	YES	0.7692	PASS (PPP)	0.5074
2	YES	1.0833	PASS (PPP)	0.9164
3	YES	0.8911	PASS (PPP)	0.7391
4	YES	1.0007	PASS (PPP)	0.1847
5	YES	0.8405	PASS (PPP)	0.5694
6	YES	1.7137	PASS (PPP)	1.5398
7	YES	0.8633	PASS (PPP)	0.2917
8	YES	1.3845	PASS (PPP)	0.2797
9	YES	1.0032	PASS (PPP)	0.5828
10	YES	1.6049	PASS (PPP)	1.3779
11	YES	1.7517	PASS (PPP)	1.5069
12	YES	1.7636	PASS (PP?)	1.3078
13	YES	0.7862	PASS (PPP)	0.9804
14	YES	0.5342	PASS (PPP)	0.7848
15	YES	0.5356	PASS (PPP)	0.6027
16	YES	0.7404	PASS (PPP)	0.6673
17	YES	1.2598	PASS (PPP)	0.5892
18	YES	0.8114	PASS (PPP)	0.5102
19	YES	1.0734	PASS (PPP)	0.9164
20	YES	1.2451	PASS (PPP)	0.5988

Table 4. Retrospective test of pathology calibration set samples against calibration set of normal samples. Mean M distance for normal calibrated set is 1.0602±0.51 and Mean Spectral residual is 0.7569±0.028. Acceptance value is fixed to twice the mean M distance of normal calibration set.

Spectral no	Match	M distance	Limit test **	Spectral residual
21	NO	7.8591	FAIL (FFF)	5.0169
22	NO	8.5018	FAIL (FFF)	5.3739
23	NO	7.3529	FAIL (FFF)	4.6739
24	NO	4.9425	FAIL (FFF)	3.3747
25	NO	8.5206	FAIL (FFF)	5.4821
26	NO	7.8294	FAIL (FFF)	5.0665
27	NO	5.4197	FAIL (FFF)	3.6616
28	NO	5.5963	FAIL (FFF)	3.6608
29	NO	5.1404	FAIL (FFF)	3.6272
30	NO	4.7382	FAIL (FFF)	5.0196
31	NO	5.4372	FAIL (FFF)	3.5981
32	NO	10.099	FAIL (FFF)	6.3894
33	NO	4.9682	FAIL (FFF)	2.4129
34	NO	4.9958	FAIL (FFF)	2.4899
35	NO	16.0364	FAIL (FFF)	9.5869
36	NO	4.5523	FAIL (FFF)	2.5229
37	NO	4.7859	FAIL (FFF)	2.8999
38	NO	4.6783	FAIL (FFF)	5.0518
39	NO	7.8393	FAIL (FFF)	5.0665
40	NO	8.9971	FAIL (FFF)	2.5192

Table 5. PCA of test normal samples against calibrated set of standard normal samples. Mean M distance for normal calibrated set is 1.0161 and Mean Spectral residual is 0.7112. Acceptance value of M - distance =3.0.

Spectral number	Match	M. Distance	Limit test ***	Spectral residual
41	YES	0.8593	PASS (PPP)	0.6065
42	YES	1.1834	PASS (PPP)	0.9234
43	YES	0.8912	PASS (PPP)	0.8382
44	YES	1.1125	PASS (PPP)	0.0958
45	YES	0.7503	PASS (PPP)	0.6585
46	YES	1.8137	PASS (PPP)	1.6498
47	YES	0.7833	PASS (PPP)	0.3917
48	YES	1.3955	PASS (PPP)	0.2897
49	YES	1.1232	PASS (PPP)	0.6837
50	YES	1.6959	PASS (PPP)	1.4888
51	YES	1.5617	PASS (PPP)	1.6069
52	YES	1.5547	PASS (PPP)	1.4129
53	YES	0.8763	PASS (PPP)	0.8815
54	YES	0.6343	PASS (PPP)	0.7849
55	YES	0.8993	PASS (PPP)	0.7066
56	YES	1.0844	PASS (PPP)	0.9232
57	YES	0.7813	PASS (PPP)	0.7482
58	YES	1.0108	PASS (PPP)	0.0769
59	YES	0.8513	PASS (PPP)	0.5786
60	YES	1.7228	PASS (PPP)	1.5434
61	YES	0.5569	PASS (PPP)	0.8116
62	YES	0.9366	PASS (PPP)	0.3452
63	YES	0.8719	PASS (PPP)	0.4259
64	YES	1.0687	PASS (PPP)	0.2527
65	YES	0.9893	PASS (PPP)	0.2553
66	YES	0.9319	PASS (PPP)	1.0016
67	YES	0.7813	PASS (PPP)	0.2299
68	YES	0.9303	PASS (PPP)	0.5785
69	YES	1.0133	PASS (PPP)	0.5928
70	YES	1.0258	PASS (PPP)	0.6294
71	YES	0.7763	PASS (PPP)	0.6962
72	YES	0.6243	PASS (PPP)	0.5848
73	YES	0.7357	PASS (PPP)	0.7027
74	YES	0.7405	PASS (PPP)	0.6773
75	YES	1.2699	PASS (PPP)	0.5872
76	YES	0.8345	PASS (PPP)	0.4802
77	YES	1.2344	PASS (PPP)	0.9464
78	YES	1.2436	PASS (PPP)	0.5999
79	YES	0.5555	PASS (PPP)	0.8145
80	YES	0.9457	PASS (PPP)	0.3543

V. DISCUSSION

In PCA analysis, we have used twenty spectra each from certified normal and malignant oral tissue samples and calibration set is built. As mentioned earlier , we have

used four factors for all final calculations and M-distance and spectral residual as discrimination parameters. Figure 5, shows a plot of the M–distance against residual errors squared sum for a new set of 77 samples (40 normal + 37 pathology), compared to a standard set of normal spectra. It is clearly seen from the plot that all samples diagnosed as normal by pathological examination in the new set fall in the lower left-hand corner of the plot. If we take a M-distance of 1 as acceptance, then almost all the samples classified as normal samples fall within 3 times this value, while all samples classified as pathology lie far outside. The specificity and sensitivity of this technique is thus quite good, 100% and 83.8% respectively (shown in table 7). A closer observation of Figure 5 shows a very small number of samples outside the acceptable range of either normal or pathology species. i.e. the overlap between two sets are negligible up to mean +-2 standard deviations, which shows the probability of samples being in the respective clusters to be about 97% and finding them out of the cluster is less than 3%. All the pathology samples have a Mdistance >>2, indicating the probability of their belonging to the normal group, practically zero. All the normal samples have a M-distance much lower than 2, showing the probability of these being out of the group negligible. Once the validity of the standard calibration sets was established, we carried out a limited test on the predictive value of the method. For this 77 additional spectra (40 normal and 37 pathology spectra) were predicted with the standard calibration set prepared earlier.

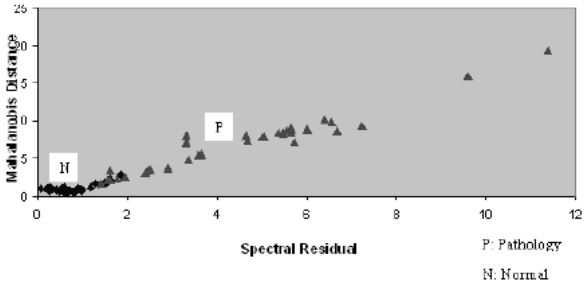

Fig. 5. Plot of spectral residual v/s M.distance- normal and pathology spectra against normal calibration set

For better discrimination of normal and pathology spectra, we have used match mismatch criteria by comparing calibration set as well as test set samples (normal and pathology spectra) with the normal calibration set considering the Mahalanobis distance (M-distance) of »3. According to this criterion, all spectra that fall within the limits are labeled as 'match' and others are labeled as 'no match'. When a set of 20 spectra of normal samples is used as calibration standard, all normal spectra of the calibration set were shown 'match' and all pathology calibration set spectra were shown 'no match' as shown in Table 3 and 4 respectively. In this case all the normal spectra are tested retrospectively by rotating out each spectrum from the calibration set, while all the pathology spectra are tested prospectively.

PCA of this region was then repeated in the prediction mode for testing match or mismatch of test samples with the normal calibration set. As expected when the set of 40 spectra of normal samples were shown 'match' and 31 out of 37 pathology spectra were shown 'no match' as shown in Table 5 and 6. Six pathology test spectra which shown match when normal calibration set was used for prediction of 37 pathology spectra may be due to the recording of the spectra from normal site of the pathology tissue. From the tables it can be seen that the results are very satisfactory and the PCA using match mismatch can be used for the discrimination between normal and pathology cases.

We have also plotted M-distance versus sample number for 117 spectra (20+40 normal and 20+ 37 pathology) as shown in Figure 6.

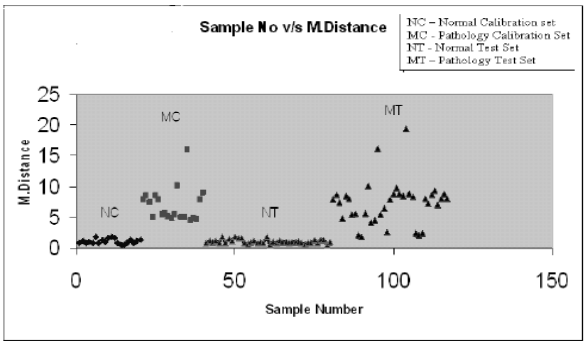

Fig. 6. Classification of 117 spectra (20+40 normal, 20+37 pathology). Sample number against M. distance for normal and pathology calibration and test spectra

Table 6. PCA of test pathology samples against calibrated set of standard normal samples. Mean M distance for normal calibrated set is 1.0161 and Mean Spectral residual is 0.7112. Acceptance value of M - distance =3.0

Spectral number	Match	M. Distance	Limit test ***	Spectral residual
81	NO	7.8899	FAIL (FFF)	5.1179
82	NO	8.7124	FAIL (FFF)	5.4739
83	NO	7.3679	FAIL (FFF)	4.6924
84	NO	4.7835	FAIL (FFF)	3.4764
85	NO	8.5322	FAIL (FFF)	5.4921
86	NO	7.9394	FAIL (FFF)	5.1772
87	NO	5.4124	FAIL (FFF)	3.6544
88	NO	5.5765	FAIL (FFF)	3.5643
89	YES	2.1505	PASS(PPP)	1.6373
90	YES	1.7433	PASS(PPP)	1.4396
91	NO	5.5472	FAIL (FFF)	3.5675
92	NO	10.103	FAIL (FFF)	6.4322
93	NO	4.1235	FAIL (FFF)	2.4266
94	NO	4.5434	FAIL (FFF)	2.4933
95	NO	16.1346	FAIL (FFF)	8.6433
96	NO	5.4443	FAIL (FFF)	2.5434
97	NO	6.4636	FAIL (FFF)	2.9968
98	YES	2.5643	PASS(PPP)	1.9455
99	NO	7.8454	FAIL (FFF)	6.0943
100	NO	8.8766	FAIL (FFF)	6.1278
101	NO	9.8162	FAIL (FFF)	6.5458
102	NO	8.7603	FAIL (FFF)	5.6559
103	NO	8.4583	FAIL (FFF)	5.6559
104	NO	19.3782	FAIL (FFF)	11.3654
105	NO	8.8594	FAIL (FFF)	6.0079
106	NO	8.4018	FAIL (FFF)	5.4739
107	YES	2.3404	PASS(PPP)	1.6269
108	YES	2.1404	PASS(PPP)	1.5672
109	YES	2.3304	PASS(PPP)	1.7872
110	NO	8.0493	FAIL (FFF)	4.6424
111	NO	7.2592	FAIL (FFF)	5.7323
112	NO	8.6743	FAIL (FFF)	6.6666
113	NO	9.3213	FAIL (FFF)	7.2223
114	NO	7.1233	FAIL (FFF)	3.3246
115	NO	8.0428	FAIL (FFF)	3.3202
116	NO	8.8162	FAIL (FFF)	5.5459
117	NO	8.0428	FAIL (FFF)	3.3204

Table 7. Performance of PCA (77 test signals)

Classifier	Specificity(%)	Sensitivity(%)	Accuracy (%)
PCA	100	83.8	92.2

As seen from the performance tables of PCA analysis, it is found that the sensitivity is 83.8% in case of PCA. PCA analysis play important role when biochemical composition of subject is considered [8].

In case of pathology test spectra used for prediction against normal calibration set, it is seen that there were few pathology spectra, which were classified, as normal. This deviation may be due to the recording of the spectra from normal site of patient.

VI. ACKNOWLEDGMENT

We are thankful to Prof. Jagdish Nayak for his valuable suggestions during PCA analysis.

REFERENCES

[1] G.Subramanya Nayak, Jagdish Nayak, 2006,, Electroglottographic Signal Acquisition and Neural Network based Classification for Pathology, IFMBE Proceedings, Vol 15,pp 54 –57.

[2] Krishnamurthy, Childers, 1986, Two Channel Speech Analysis. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-34: pp.730-743.

[3] Robert F.Orlikoff, 1991, Assessment of Dynamics of vocal Fold contact From the Electroglottogram, Journal of Speech and Hearing Reaserch, 34:1066 –1072.

[4] Rudra Pratap, Getting started with MATLAB5, Oxford University Press, 2000

[5] William Palm J, Introduction to MATLAB 6 for Engineers.

[6] B.K.Manjunath, J.Kurien, C.Muralikrishna, V.B.Kartha, 2006, Autofluorescence of Oral tissue for optical pathology in Oral malignancy, Journal of Photochemestry and Photobiology Biology, 73, pp. 49 -58.

[7] Mahalanobis, P.C., Proc. Natl. Inst. Sci. India, 1936, 2, 49.

[8] G.S.Nayak, Sudha Kamath et al. 2006, Principal Component Analysis and Artificial Neural Network Analysis of oral Tissue Fluorescence Spectra: Classification of Normal,Premelignant and Malignant Pathological Condition, Biopolymers,82:152-166.