Performance of Gaussian Mixture Models as a Classifier for Pathological Voice

Jianglin Wang, Cheolwoo Jo

SASPL, School of Mechatronics
Changwon National University
Changwon, Gyeongnam 641-773, Republic of Korea
xiaowangyc@hotmail.com, cwjo@sarim.changwon.ac.kr

Abstract

This study focuses on the classification of pathological voice using GMM (Gaussian Mixture Model) and compares the results to the previous work which was done by ANN (Artificial Neural Network). Speech data from normal people and patients were collected, then diagnosed and classified into two different categories. Six characteristic parameters (Jitter, Shimmer, NHR, SPI, APQ and RAP) were chosen. Then the classification method based on the artificial neural network and Gaussian mixture method was employed to discriminate the data into normal and pathological speech. The GMM method attained 98.4% average correct classification rate with training data and 95.2% average correct classification rate with test data. The different mixture number (3 to 15) of GMM was used in order to obtain an optimal condition for classification. We also compared the average classification rate based on GMM, ANN and HMM.

1. Introduction

The diagnosis of pathological voice is a hot topic that has been received considerable attention. There are several medical diseases that adversely affect our human voice (Jo & Kim, 1998; Dibazar & Narayanan, 2002). The doctor can use the available apparatus for detection of pathological voice. However, it is invasive and requires an expert analysis of numerous human speech signal parameters. Automatic voice analysis for pathological speech has its advantages, such as having its quantitative and non-invasive nature, allowing the identification and monitoring of vocal system diseases and reducing analysis cost and time (Dibazar & Narayanan, 2002). In the pathological voice classification, based on the voice of a patient, the goal is to make a decision whether it is normal or pathological. Successful pathological voice classification will enable an automatic non-invasive device to diagnose and analyze the voice of the patient.

In the recent approaches to pathological voice classification, various pattern classification methods have been used. Several researches, such as classification of pathological voice including severely noisy cases (Li, Jo & Wang, 2004), pathological voice quality assessment using ANN (Ritchings, Mcgillion & Moore, 2002) and using short-term cepstral parameters and neural network based detectors for automatic detection of voice impairments (Godino-Llorente & Gomez-Vilda, 2004), have recently been applied to various kinds of pathological classification tasks. Generally, ANN has been widely used because there is no need to think about the details of the mathematical models of the data and relatively easy to train and has produced a good pathological recognition performance. The major drawback to ANN method is that it depend on the data set and the ANN method can not guarantee a good performance in such cases when the total size of the data is small and when the new data is added to the original data set.

In previous study, HMM-based method has also been conducted to automatically detect disordered speech. Dibazar et al. applied HMM to classify the pathological speech (Dibazar& Narayanan, 2002; Wang & Jo, 2006).

The GMM method is based on a finite mixture probability distribution model. And the method was successfully applied on robust speaker recognition system. GMM provides a robust speaker representation for the difficult task of speaker identification using corrupted, unconstrained speech (Reynolds, Rose & Smith, 1992). The models are computationally inexpensive and easily implemented on a real-time platform. Furthermore, their probabilistic framework allows direct integration with speech recognition systems and incorporation of newly developed speech robustness techniques. The success of speaker identification provides us insight to apply the GMM method into pathological voice identification.

In this paper the GMM method was used to classify the mixed voiced data set (pathological and normal voice) into normal and pathological voices. Six characteristic parameters were chosen. The Gaussian mixture model was used to classify the pathological voice. Finally, the results of GMM method were compared to the previous results done by ANN (Li et al., 2004) and HMM (Wang & Jo, 2006).

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

2. Database

To collect voice data, collection system was installed in a room of the ENT department of hospital. The recording process was executed semi-automatically with the intervention of operator to control the quality and procedure. Also the voice materials from the different male speakers were collected using DAT (Digital Audio Tape) (Jo, Kim, Kim, Wang & Jeon, 2001). The sampling rate was 50 KHz and the resolution 16 bits.

The collection was conducted in the soundproof room of a hospital. All the subjects were asked to pronounce a sustained vowel /a/. Total voice data included 41 normal cases and 111 pathological cases (108 relatively less noisy voices and 3 severely noisy voices) after removing invalid data from the raw data sets. The vocal diseases considered in the relatively less noisy pathological cases consisted of Vocal Polyp, Vocal Cord Palsy, Vocal Nodule, Vocal Cyst, Vocal Edema and Laryngitis. The database of vocal fold diseases is shown in Table 1.

Table 1. Vocal Fold Diseases

Disease	No. of Case
Cyst	5
Edema	10
Laryngitis	5
Nodule	10
Palsy	10
Polyp	10
Glottic Cancer	58
Total	108

3. Methodology

In this section, the used analysis procedures and methods are described and our study focuses on the classification between normal and pathological cases. The analysis method for classification of the severely noisy pathological voice was introduced in previous work (Li et al.).

The MDVP (Multi-Dimensional Voice Program) was used as an analyzer to calculate the 6 different kinds of numerical parameters (Operations Manual, 1993). The 6 numerical parameters are Jitter, RAP, Shimmer, APQ, NHR and SPI. The 6 parameters are divided into 3 categories according to their characteristics. There are 2 pitch related parameters (Jitter, RAP), 2 amplitude related parameters (Shimmer, APQ) and 2 noise related parameters (NHR, SPI). They can represent several aspects of the speech characteristic information. So the six parameters were chosen as the input in our experiments. Each parameter is defined as follows.

3.1. Parameters

3.1.1. Jitter

Jitter is relative evaluation of the period-to-period (very shot-term) variability of the pitch within the analyzed voice sample. Voice break areas are excluded.

$$\text{Jitt} = \frac{\frac{1}{\text{N-1}}\sum_{i=1}^{N-1}\left|To^{(i)} - To^{(i+1)}\right|}{\frac{1}{\text{N}}\sum_{i=1}^{N} To^{(i)}} \qquad (1)$$

where $To^{(i)},\ i=1,\ 2\ ...\ N$ is the extracted pitch period data and $N = PER$ is the number of extracted pitch periods.

3.1.2. RAP

RAP (Relative Average Perturbation) is the relative evaluation of the period-to-period variability of the pitch within the analyzed voice sample with smoothing factor of 3 periods. Voice break areas are excluded.

$$\text{RAP} = \frac{\frac{1}{\text{N-2}}\sum_{i=2}^{N-1}\left|\frac{To^{(i-1)} + To^{(i)} + To^{(i+1)}}{3} - To^{(i)}\right|}{\frac{1}{\text{N}}\sum_{i=1}^{N} To^{(i)}} \qquad (2)$$

where $To^{(i)},\ i=1,\ 2\ ...\ N$ is the extracted pitch period data and $N = PER$ is the number of extracted pitch periods.

3.1.3. Shimmer

Shimmer Percent /%/ is relative evaluation of the period-to-period (very short term) variability of the peak-to-peak amplitude within the analyzed voice sample. Voice break areas are excluded.

$$\text{Shim} = \frac{\frac{1}{\text{N-1}}\sum_{i=1}^{N-1}\left|A^{(i)} - A^{(i+1)}\right|}{\frac{1}{\text{N}}\sum_{i=1}^{N} A^{(i)}} \qquad (3)$$

where $A^{(i)},\ i=1,\ 2...N$ is the extracted peak-to-peak amplitude data and $N = PER$ is the number of extracted impulses.

3.1.4. APQ

APQ (Amplitude Perturbation Quotient) /%/ is relative evaluation of the period-to-period variability of the peak-to-peak amplitude within the analyzed voice sample at smoothing of 11 periods. Voice break areas are excluded.

$$\text{APQ} = \frac{\frac{1}{\text{N-4}}\sum_{i=1}^{N-4}\left|\frac{1}{5}\sum_{r=0}^{4} A^{(i+r)} - A^{(i+2)}\right|}{\frac{1}{\text{N}}\sum_{i=1}^{N} A^{(i)}} \qquad (4)$$

where $A^{(i)},\ i=1,\ 2...N$ is the extracted peak-to-peak amplitude data and $N = PER$ is the number of extracted impulses.

3.1.5. NHR

NHR (Noise-to-Harmonic Ratio) is the average ratio of the inharmonic spectral energy in the frequency range 1500-4500 Hz to the harmonic spectral energy in the frequency range 70-4500 Hz. This is a general evaluation of noise present in the analyzed signal.

3.1.6. SPI

SPI (Soft Phonation Index) is the average ratio of the lower-frequency harmonic energy in the range 70-1600 Hz to the higher-frequency harmonic energy in the range 1600-4500 Hz.

3.1.7. GMM

The classification of pathological voice used in this study was based on GMM-based approach, which consists of 3 steps. First, the characteristic parameters of pathological voice data were calculated. In our study, there were 6 characteristic parameters, which had been introduced above sections. Second, each GMM model with different Gaussian mixtures was trained. The GMM model was set with different Gaussian mixtures to obtain an optimal classification performance. Third, trained GMM models were used as a classifier to perform the classification tasks. The classification method of GMM is shown in Fig. 1.

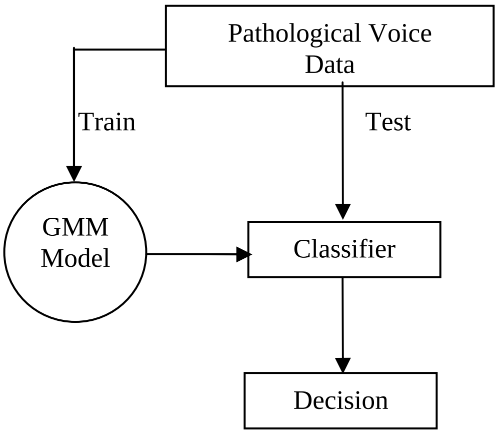

*Figure 1.*Block diagram of the method.

4. Experimental Results

The experimental results for the classification of pathological voice are shown in this section. And the purpose of our experiment is to classify the mixed voice data set into normal and pathological voice and to compare the classification performance to the results of artificial neural network in previous work. So the same pre-condition was configured, such as the same data set, the same characteristic parameters and the same training times.

Since the total number of data was small, we tried to train and test the ANN and GMM model by splitting total data sets into two parts. Two thirds of the data were used for training and the remaining one third of data was used for test. In each training stage, the artificial neural network and Gaussian mixture model were trained and tested separately using different combination of data sets, which is especially useful when few speech data are available for the model training and testing. It is able to compensate the small size of the data sets. And each stage for training and test was performed 5 times.

The GMM based classifier was constructed to classify the normal voices and pathological voices after the computation of the 6 parameters. Total data combination sets were performed 5 times. Different Gaussian mixtures (3, 4 and 5) are set to compare which Gaussian mixture model can attain a high classification performance. Table 2 shows the classification rate from the GMM training and testing. The "1st Run" means to run the first set of data and the others own the same meaning. From the Table 2, the highest classification is 5 Gaussian mixtures model. The average classification rate for training data is 98.4% and for test data is 95.2%.

*Table 2.*The classification rate (%) in GMM.

GMM Mixtures	Performance Times	Training Data	Test Data
3 Mixtures	1st Run	98.0	96.0
	2nd Run	99.0	92.0
	3rd Run	98.0	92.0
	4th Run	95.0	90.0
	5th Run	95.0	94.0
4 Mixtures	1st Run	97.0	94.0
	2nd Run	98.0	84.0
	3rd Run	96.0	88.0
	4th Run	90.9	94.0
	5th Run	97.0	92.0
5 Mixtures	1st Run	98.0	96.0
	2nd Run	99.0	96.0
	3rd Run	100.0	94.0
	4th Run	97.0	96.0
	5th Run	98.0	94.0

5. Discussion

In this section, we discuss some experimental results obtained from the proposed analysis methods.

In our experiment, we also need to know which number of mixture can give us the optimal classification rate. So the different mixture number (3 to 15) of GMM has been trained to find the optimal number. In Fig. 2, it shows the classification rate for different mixtures (3 to 15). From the Fig. 2, the best mixture for test data is 5 and train data is 11. Due to the limited number of data, each time we used the different data set for train and test. In Fig. 2, the average classification rate for each mixture was calculated.

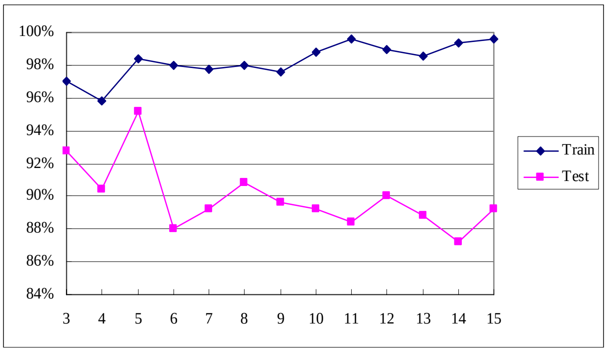

*Figure 2.*The classification rate for different mixtures.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

The confusion matrix can show us how many correct classification rates have been identified. The best result (5 GMM mixtures in 2nd run) is presented through confusion matrices. In the Table 3 and 4, True positive (TP) is the ratio between normal voice correctly classified and the total number of normal voices. False negative (FN) is the ratio between wrongly classified pathological voices and the total number of normal voices. True negative (TN) is the ratio between pathological voices correctly classified and the total number of the pathological voices. False positive (FP) is the ratio between normal voices wrongly classified and the total number of pathological voices. From Table 4, GMM classification for test data shows a highly correct classification rate for normal and pathological voices. The FN=2.8% means that some pathological speech data were misclassified as normal case, however, the TN=97.2% demonstrates that most of the pathological speech data were correctly classified. The former case is serious for practical condition. Comparing Table 3 to 4, the difference of FP between train data and test data is 7.1% and TN between train data and test data is 1.4%.

*Table 3.*The confusion matrix for Train data.

		Decision	
		Normal	Pathological
In	Normal	TP=100%	FP=0
	Pathological	FN=1.4%	TN=98.6%

*Table 4.*The confusion matrix for Test data.

		Decision	
		Normal	Pathological
In	Normal	TP=92.9%	FP=7.1%
	Pathological	FN=2.8%	TN=97.2%

From the result, we can reason that the pathological speech voice owns so distinctive characteristics that GMM can recognize the data according to its characteristic parameters.

In this part, the comparison among three methods (ANN, HMM and GMM) will be discussed. The ANN based approach had been introduced in the previous work (Li et al.). The different hidden layers (6, 9 and 12) were set to obtain the much better classification rate. The best performance of ANN is 12 hidden layers. The average classification rate for training data is 98.0% and for test data is 94.2%.

From the experiment using six different sorts of parameters as an input and with the different structures model of GMM and ANN, the classification rate for discriminating the voice into the normal and pathological cases were obtained as shown in Table 2. The classification rates of test data were high so that the accurate classification could be realized. And it was clearly observed that the classification rate rose with the number of hidden layers and Gaussian mixtures increased and the model of ANN and GMM is more sophisticated.

In GMM for each number of mixtures, the average classification rate for training data and test data was obtained as shown in Fig. 3. In ANN for each hidden layers, the average classification rate for training data and test data was calculated as shown in Fig. 4. In Fig. 5 it shows our previous result using HMM-based method to classify the pathological voice (Wang & Jo, 2006). Comparing the results between Fig. 3 and Fig. 4, the best GMM configuration showed slightly better classification compared to the best performance of ANN (0.4% in training data, 1% in test data). Although the absolute amount of the rate difference is small, it can present that GMM can be used as more robust classifier in pathological

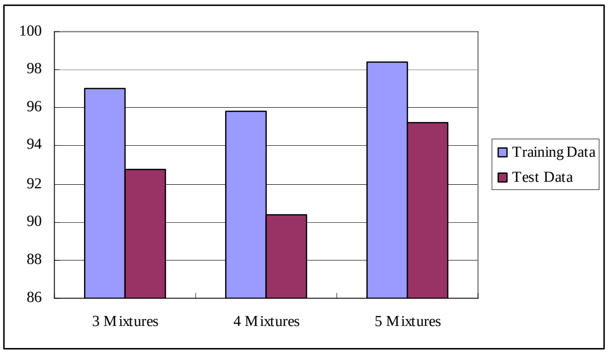

*Figure 3.*The average classification rate (%) in GMM.

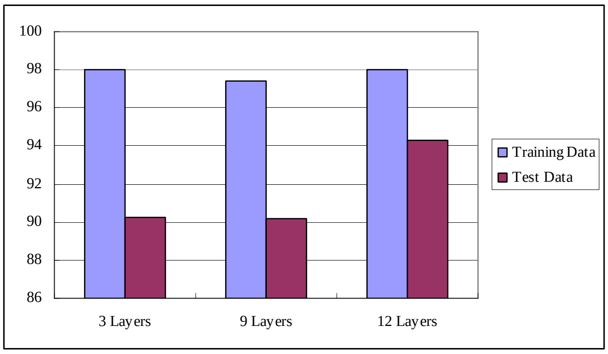

*Figure 4.*The average classification rate (%) in ANN.

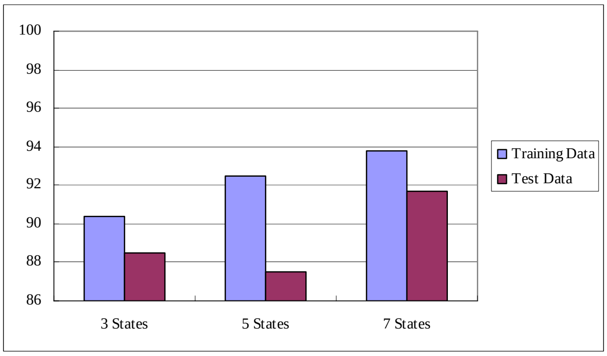

*Figure 5.*The average classification rate (%) in HMM.

*Table 5.*The best case for each classification method

Methods	Train Data	Test Data
GMM	98.4	95.2
ANN	98.0	94.2
HMM*	93.8	91.7

**different characteristic parameter sets*

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

voice classification giving us a comparable classification rate to ANN. In Fig. 3 and Fig. 5, GMM-based method for the best cases of train data and test data are 4.6% and 3.5% more than HMM-based method. However, HMM-based method used different characteristic parameters sets. In Table 5, it presents the best case of the average classification rate for each classification method.

6. Conclusions

This paper has performed a classification of the pathological voice using GMM method as one of the trial to obtain a better classification performance than ANN method, which was previously performed. The pathological classification method based on GMM shows 0.4 % improvement with training data and 1% improvement with test data on the average. Although the amount of the difference is small, it is proved that GMM can be used effectively for the classification method of the pathological voice giving us more robustness in practical applications. Comparing HMM-based method to GMM-method, the latter one can show us good results for the same pathological data. However, characteristic parameters are different in each experiment. So the direct comparison is not proper.

In the future work, to improve the performance with real data, more investigations are required on the proper number of mixtures on Gaussian model and on the proper parameter sets. Also it is highly required to extend the size of the pathological voice database.

7. References

Dibazar, A. A. & Narayanan, S. (2002). A System for Automatic Detection of Pathological Speech, *Proceedings of 36th Asilomar Conf. Signals, Systems & Computers*, Pacific Grove, CA, USA.

Godino-Llorente, J.I. & Gomez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors, *IEEE Transactions on Biomedical Engineering*, 51, 2, 380-384.

Jo, C. & Kim, D. (1998). Diagnosis of Pathological Speech Signals Using Wavelet Transform, *Proceedings of International Technical Conference on Circuits/Systems, Computers and Communications*, 657-660, Sokcho, Korea.

Jo, C., Kim, K., Kim, D., Wang, S., & Jeon, G. (2001). Comparisons of Acoustical Characteristics between ARS and DAT Voice, *International Conference on Speech Processing*, 949-953, Taejon, Korea.

Li, T., Jo, C. & Wang, S. (2004). Classification of pathological voice including severely noisy cases, *Proceedings of the 8th International Conference on Spoken Language Processing*, I, 77-80, Jeju, Korea.

Operations Manual (1993). Multi-Dimensional Voice Program (MDVP) Model 4305, 93-131, Kay Elemetrics Corp.

Reynolds, D. A., Rose, R. C. & Smith, M. J. T. (1992). PC-based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system, *Proceedings of Int. Conf Signal Processing Appl., Technol.*, 967-973, Boston, MA, USA.

Ritchings, R.T., Mcgillion, M.A., & Moore, C.J. (2002). Pathological voice quality assessment using artificial neural network, *Medical Engineering & Physics*, 24, 8, 561-564, ELSEVIER.

Wang, J. & Jo, C. (2006). Classification of Vocal Fold Disorder using Hidden Markov Models, *Proceedings of the 23rd Korean Speech Communication & Signal Processing Conference*, 229-232, Ansan, Korea.