MULTIVARIATE MODELING AND DIAGNOSTIC CLASSIFICATION OF PULMONARY SOUNDS

by

İpek Şen

B.S., Electrical and Electronics Engineering, Boğaziçi University, 2002

M.S., Electrical and Electronics Engineering, Boğaziçi University, 2005

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University
2013

MULTIVARIATE MODELING AND DIAGNOSTIC CLASSIFICATION OF PULMONARY SOUNDS

APPROVED BY:

Prof. Z. Yasemin Kahya .....
(Thesis Supervisor)

Assoc. Prof. Murat Saraçlar.....
(Thesis Co-supervisor)

Prof. Ahmet Ademoğlu .....

Prof. Emin Anarım .....

Prof. Nizamettin Aydın .....

Prof. Ayşın Ertüzün .....

DATE OF APPROVAL: 17.06.2013

To myself...
The darkest enemy
and the dearest ally,
the well hidden destination
of my lifelong journey.

ACKNOWLEDGEMENTS

I can not express my gratitude to my thesis supervisor, Prof. Yasemin Kahya. Her warm support in many aspects of my life, in my academic and non-academic efforts and interests, rendered it possible for me to complete the study which began with a senior project and grew through two graduate periods. Her practical solutions to problems, strong intuition and experience, and also her encouragement and understanding, helped me to align with the larger picture when I lost direction. I am also deeply grateful to my co-supervisor, Assoc. Prof. Murat Saraçlar, for his invaluable contribution to this thesis study. His deep knowledge and valuable suggestions helped this thesis to shape better towards its aim. I would also like to thank the thesis committee members for their valuable comments and suggestions at various steps of the study throughout the path.

I would like to thank Sibel Yurt, MD, from Yedikule Teaching Hospital for Chest Diseases and Thoracic Surgery, for her guidance on data acquisition and diagnosis of pulmonary disorders of pathological subjects. I am grateful for her kind hospitality during our visits and for her supportive efforts in the formation of the large data bank of pulmonary sounds. I also thank Prof. Filiz Koşar, the chief of the sixth clinic of the hospital, the associate physicians who collaborated with Sibel Yurt, and all the hospital staff that helped the data acquisition procedure in their own ways. I would also like to thank the pathological and healthy subjects who accepted to take part in the study. I wish them all to have healthy lives.

Mete Yeğiner and R. Koray Çiftçi were the two members of LAL when I joined them as a senior student, and we worked together for years until they handed over Usagi to me and left for the subsequent phases of their lives. I am grateful to them for their contributions to my maturing process from a senior student to a researcher. They provided me to understand the importance of harmony and friendship between colleagues, putting aside the academic perspectives and direct contributions they offered. I would also like to express my special thanks to my former student - present

buddy İsmail Kara. It was a big chance for me to be accompanied by him in the last two years of weekly visits to the hospital.

I would like to thank my friends, Oya Çeliktutan Dikici, Doğaç Başaran, A. Çisel Aras and Erinç Dikici, for the the loving, friendly, and supportive atmosphere that any kind of product requires to flourish. I am deeply grateful to Oya for offering a fresh air at the times I stacked at a local minimum, and for reminding me the ways to make life easier and more balanced, joyful, fruitful and meaningful. I thank Doğaç a lot for the peaceful and loving atmosphere that we shared in the laboratory, and for the kind support he offered during the hard times of this thesis study. I am also grateful to my dear friend A. Yasin Çitkaya, who was magically present at critical times. I would not be able to solve many problems, both technical and non-technical, without the competent and kind help offered by him. My millions of thanks go to Leyla Çeken and Aynur İşler, the two angels of our department, for their warm and skillful support.

I would also like to thank my precious friends Tuğba Davran Candan and Başak Ay Saylam for their warm presence and support at all times, and for showing me that it is possible to build a family from strangers. Their contribution to this study can not be neglected, although indirectly felt.

My deepest gratitude goes to my parents Perihan and Alâeddin Şen for being my first and lifelong teachers and for offering more than I dare to wish. I thank my little sister İdil for her surprising wisdom at times as if she is the elder, and Fındık for the relief and joy she offered. I am grateful to my family for their endless patience, trust and love. Finally, I would like to thank deeply and literally all the people and events that I have encountered on the path called life, since they contributed to my degree of doctor of philosophy, with a particular emphasis on the *philosophy* part.

This thesis was supported in part by the Scientific and Technical Research Council of Turkey (TÜBİTAK) Integrated Doctorate Program (BDP), and in part by Boğaziçi University Research Fund under Project No. 09A203D.

ABSTRACT

MULTIVARIATE MODELING AND DIAGNOSTIC CLASSIFICATION OF PULMONARY SOUNDS

Computerized pulmonary sounds analysis has become prevalent in the recent decades as it provides means for quantitative and objective evaluation, contrary to the limited and subjective nature of stethoscope auscultation. Multi-variate methods disclose the inherent spatial information in the multi-channel measurements and enable to explore the system characteristics altered due to pathological conditions that have developed in the lungs. In this study, a multi-variate mathematical model, namely, vector auto-regressive (VAR) model, has been considered, and the optimal VAR model to represent the pulmonary sounds data is pursued through new goodness of fit criteria proposed specifically for the data and the application. The estimated model parameters are employed in classification using the k-nearest neighbor (k-NN), support vector machine (SVM), and Gaussian mixture model (GMM) classifiers, with an eventual aim to develop a diagnostic classifier for clinical setups. Various classifier schemes are experimented with different data sets in the quest for the most useful classifier design. The healthy and the pathological groups are discriminated successfully. In the classification of conditions including the healthy group and the obstructive and restrictive types of pathologies, a hierarchical framework is suggested. Generally, the healthy and the restrictive groups are discriminated more successfully than the obstructive group. GMM is generally the most competent classifier among all, however, SVM is also successful for certain feature arrangements. The improvement of the diagnostic classifier so as to make it appropriate for clinical setups is still open for exploration, especially with additional features to enhance the distinctive characteristics further as to prevent the confusion of the obstructive diseases.

ÖZET

SOLUNUM SESLERİNİN ÇOK DEĞİŞKENLİ MODELLENMESİ VE TANIYA YÖNELİK SINIFLANDIRILMASI

Bilgisayarlı solunum sesleri analizi, stetoskop ile dinlemenin aksine nicel ve nesnel değerlendirme imkanı sunması ile, son bir kaç onyılda öne çıkmıştır. Çok değişkenli yöntemler, çok kanallı ölçümlerin barındırdığı uzamsal bilgiyi açığa çıkararak, akciğerlerde gelişmiş patolojik koşullardan dolayı değişen sistem özelliklerini araştırmaya olanak sağlar. Bu çalışmada, çok değişkenli matematiksel bir model olan vektör öz-bağlanımlı (VÖB) model ele alınmış, solunum ses verisini ifade eden eniyi VÖB model, veri ve uygulama için özel olarak önerilen yeni başarım kriterleri kullanılarak araştırılmıştır. Kestirilen model parametreleri, klinik ortamlar için tanıya yönelik bir sınıflandırıcı geliştirme nihai amacı ile, k-en yakın komşu (k-EYK) algoritması, destek vektör makineleri (DVM) ve Gauss karışım modelleri (GKM) kullanılarak sınıflandırılmıştır. En işe yarar sınıflandırıcıyı tasarlamak amacı ile farklı veri kümeleri üzerinde çeşitli sınıflandırıcı tasarımları denenmiştir. Sağlıklı ve hasta gruplar başarılı bir şekilde ayrışmaktadır. Sağlıklı grubu ve tıkayıcı ve sınırlayıcı hastalık tiplerini içeren durumların sınıflandırılmasında hiyerarşik bir yapı önerilmiştir. Genel olarak, sağlıklı ve sınırlayıcı gruplar tıkayıcı gruptan daha başarılı şekilde ayrışmaktadır. GKM genel olarak tüm sınıflandırıcılar içinde en yetkini olmakla birlikte, belirli öznitelik düzenlemeleri için DVM de başarılıdır. Tanısal sınıflandırıcının klinik ortama uygun hale gelecek şekilde geliştirilmesi problemi halen, özellikle tıkayıcı hastalıkların karıştırılmasını önleyecek şekilde ayırdedici özellikleri daha da vurgulayan ek öznitelikler ile, araştırılmaya açıktır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\hat{\mathbf{A}}_i$	Estimated $(K \times K)$ parameter matrix of VAR(p) model for $i = 1, 2, ..., p$
$\hat{\mathbf{B}}$	Matrix composed of $\hat{\mathbf{A}}_i$ in form $[\hat{\mathbf{A}}_1, ..., \hat{\mathbf{A}}_p]$
$\hat{\mathbf{B}}^s_{\{i,j\},w}$	$\hat{\mathbf{B}}$ for the w^{th} segment of the j^{th} flow-phase of the i^{th} full flow-cycle of the s^{th} subject
$\mathcal{B}^s_j$	Set of row vectors $\hat{\mathbf{b}}^s_{\{i,j\},w}$ concatenated over w and i for a fixed j per s
$\hat{\mathbf{b}}^s_{\{i,j\},w}$	$\hat{\mathbf{B}}^s_{\{i,j\},w}$ converted to its vector form
$\mathbf{C}^s_g$	The representative parameter matrix of form $[\mathbf{M}^s_{.\{j\in\mathcal{J}\}}]$, where $s \in \mathcal{S}_g$, the set of subjects in group g
C_r	Class r in classification
$E\{X\}$	Expected value of random variable X
$(EFCN)^s_j$	The effective full flow cycle number for the j^{th} flow-phase of the s^{th} subject
$f_{1\|2}$	Frequency of $n_{1\|2}$ over all C_2 instances
F	SVM penalty factor
$(FCN)_s$	Total number of full flow cycles for subject s
K	Dimensionality of time series in modeling and number of folds of the cross validation scheme in classification
$\mathbf{M}^s_{i,j}$	Average of $\hat{\mathbf{B}}^s_{\{i,j\},w}$ over indices w, i.e., over segments fitting into $P^s_{i,j}$
$\mathbf{M}^s_{.j}$	Average of $\mathbf{M}^s_{i,j}$ over indices i
m	The VAR model order in modeling and GMM component count in classification
N	Sample size of the VAR model
$n_{1\|2}$	Number of instances misclassified as in C_1 while they actually come from C_2
$P(C_r)$	Prior probability of class C_r
$P(\mathbf{x}\|C_r)$	Likelihood probability of vector $\mathbf{x}$ under class C_r
$P(C_r\|\mathbf{y})$	Posterior probability of class C_r given vector $\mathbf{y}$

$P^s_{i,j}$	The j^{th} flow-phase of the i^{th} full flow-cycle of the s^{th} subject
R^2	Coefficient of determination
S_B	Between-class scatter score
S_W	Within-class scatter score
$(WN)^s_{i,j}$	The maximum number of VAR model segments fitting into $P^s_{i,j}$
$\lvert \cdot \rvert$	Determinant if $\cdot$ is a matrix, and size if $\cdot$ is a set
$\lvert\lvert \mathbf{a} - \mathbf{b} \rvert\rvert$	Distance between vectors $\mathbf{a}$ and $\mathbf{b}$
γ	SVM hyper-parameter for the radial basis function.
μ_1, μ_2, μ_3	The three proposed goodness of fit criteria for the VAR model
η	The first success score defined for the classifiers (total correct classification rate)
ζ	The second success score defined for the classifiers (harmonic mean of the frequencies of correctly classified instances)

LIST OF ACRONYMS/ABBREVIATIONS

AIC	Akaike's Information Criterion
AR	Auto-regressive
ANN	Artificial Neural Network
ATS	American Thoracic Society
COPD	Chronic Obstructive Pulmonary Disease
CT	Computed Tomography
TN	True Negatives
DS	Data Set
DWT	Discrete Wavelet Transform
ERS	European Respiratory Society
FC	Full Cycle
FIR	Finite Impulse Response
FN	False Negatives
FP	False Positives
FPE	Final Prediction Error
GMM	Gaussian Mixture Model
HQ	Hannan-Quinn (Criterion)
IIR	Infinite Impulse Response
ILD	Interstitial Lung Disease
LDA	Linear Discriminant Analysis
LS	Least Squares
k-NN	k-Nearest Neighbor
MAP	Maximum A-posteriori
ML	Maximum Likelihood
MSE	Mean Squared Error
P1	Perspective 1
P2	Perspective 2
P3-SOP	Perspective 3 - Sum Of Posteriors

P3-SOLP	Perspective 3 - Sum Of Log-posteriors
P3-SVM	Perspective 3 - SVM
P3-SVML	Perspective 3 - SVM Logarithmic
RBF	Radial Basis Function
SC	Schwartz Criterion
SE	Sensitivity
SP	Specificity
SVM	Support Vector Machine
TN	True Negatives
TP	True Positives
VAR	Vector Auto-regressive
WGN	White Gaussian Noise

1. INTRODUCTION

A pathological condition developed in the lungs is usually revealed in the altered characteristics of pulmonary sounds heard at the chest wall. Therefore, auscultation provides invaluable information about the condition of the lungs, and stethoscope auscultation is still the common and primary practice in clinics for the diagnosis of pulmonary diseases. On the other hand, stethoscope auscultation is subjective by nature, because it depends on the hearing ability and the experience of the physician, and objective comparison is not possible due to the lack of the opportunity to quantify the acoustic information. Besides, a stethoscope attenuates the frequencies above approximately 120 Hz, thereby filtering out the frequency band that is most relevant [1].

Computerized pulmonary sounds analysis techniques that have developed over the past decades enable quantitative and objective analysis. These techniques can be summarized as acquiring the sounds on the chest wall and analyzing them using mathematical methods. They provide means to express the information residing in pulmonary sounds that can not be perceived by ear, to assess the characteristics through signal processing and pattern recognition methods, and to produce numerical and visual outputs to compare and follow pulmonary conditions.

The studies conducted so far to understand the thoracic system using the acoustic data can be summarized in a few categories in terms of their objectives:

(i) Modeling: Fitting a physical model for human lungs [2–5], or exploring the characteristics of sound transmission within the lungs [6–13],

(ii) Detection - Estimation: Detection of abnormal sound components (adventitious components such as crackle and wheeze) which are associated with certain diseases [14–27] and their separation from the background pulmonary sounds [28–30], or estimation of flow cycle from pulmonary sounds data [31, 32],

(iii) Classification: Classification of types of adventitious sound components [33–35], or, classification into pulmonary conditions (e.g., healthy versus pathological)

[36–41],

(iv) Mapping: Building visual representations of the chest based on the acoustic information provided by the acquired pulmonary sounds data [42–44].

There is also an effort to develop a set of standards for the terminology special to respiratory studies, and a set of guidelines about acquisition and processing of the data. European Respiratory Society (ERS) published a review as a task force report covering the above issues in 2000 [45]. Some basic definitions can also be found in the nomenclatures published by American Thoracic Society (ATS) earlier in 1977 [46].

In the analysis of pulmonary sounds, the most common approach is to handle each channel of the measured signals separately, i.e., to extract some mathematical features on the individual channel signal and to employ the features according to the particular purpose. Although this approach is not wrong, it is insufficient since the inter-channel relationships are thereby neglected. The simultaneous channel measurements are not independent from each other and the mutual relationships are expected to differ as the pulmonary condition alters. Multi-variate analyses of the multi-channel measurements provide means to include in the features the spatial relationships between the channels as well, which takes the altered transmission characteristics into account. For these reasons, a multi-variate mathematical model is considered in this thesis study.

Auto-regressive (AR) models are known to be useful in single-channel analyses of pulmonary sound signals [32, 36, 37, 39–41, 44]. Therefore, its multi-variate version, namely, vector AR (VAR) model, is adopted here. The AR model has been widely used in the analysis of pulmonary sounds with the particular purpose to estimate flow-cycle from the sounds acquired at the chest wall [32], to classify into healthy and pathological conditions [36,37,39–41], or to model the crackles to build their distribution maps [44], however the VAR model has only been recently employed in the analysis of pulmonary sounds [41]. The model parameters are employed in classification in [41] as in this study, however, the proper model is determined via one of the conventional goodness of fit criteria defined in the literature in the minimum prediction error sense. In this study, on the other hand, new goodness of fit criteria are proposed specifically for

pulmonary sounds as to enhance the inherent discriminative characteristics, and the classification is performed using the optimal model determined accordingly.

The first part of this study, Chapter 3, is devoted to find the optimal VAR model of pulmonary sounds for diagnostic purposes. Since it aims to find a useful model to describe the condition of the lungs, it belongs under the scope of item (i) in the above list. However, it differs from the other studies under item (i) in the following aspect. The studies in [2–5] propose a physical lung model where generally the lungs are represented by equivalent electrical circuits, and those in [6–12] explore how sound is transmitted within the lungs. On the other hand, this study focuses on a mathematical model where the model parameters are intended to be employed in a diagnostic system. Although the mathematical model subject to this study is not new, the novelty is in the fact that the criteria to select the best model are proposed specifically to the pulmonary sounds and to the application (diagnostic classification).

Most of the pulmonary diseases belong to one of two main categories, namely, obstructive and restrictive. Obstructive diseases (e.g., asthma, chronic obstructive pulmonary disease (COPD), bronchiectasis, emphysema) are associated with short of breath due to airway obstruction, and are in some cases caused by bacterial infection (e.g., bronchiectasis). Restrictive diseases (e.g., interstitial lung disease (ILD), pneumonia), on the other hand, are accompanied with a stiffness of the tissue around the alveoli, and are classified according to the cause (e.g., infection, certain drugs, long term exposure to materials such as asbestos).

Studies on classification of pulmonary sounds (item (iii) in the above list) are grouped under two main categories. The first group mainly focuses on adventitious sound components (e.g., crackle, wheeze), where the aim is either to detect a particular type of adventitious component [14–27], or to discriminate between various types of adventitious sounds [33–35]. On the other hand, the focus of the second group is on *pulmonary conditions* (e.g., healthy versus pathological, obstructive versus restrictive) [36–41]. In the studies in [36, 37, 40], healthy versus pathological classification schemes are considered, where the pathological classes consist of various diseases from

obstructive and restrictive groups. In [38] and [41], on the other hand, single types of pathologies are considered in the healthy versus pathological scheme, namely, ILD and asthma, respectively. The only study in the literature that addresses the obstructive versus restrictive distinction in classification (in a three-class classification scheme with the healthy, obstructive and restrictive classes) is presented in [39]. Moreover, the only study that approaches a multi-class case is again the one in [39] as the other studies deal with the binary case.

As addressed above, AR model parameters are widely used as features in the classification of pulmonary sounds [36,37,39–41], whereas, only one study has adopted VAR models so far [41]. Other features that are used as inputs to classifiers are various power spectral parameters [18, 19, 24, 34, 35], wavelet [33, 38] and Fourier [17] coefficients, percentile frequencies [41], and eigenvalues of covariance matrices [41]. Artificial neural networks (ANNs) are adopted for classification in the majority of the studies [17, 18, 33, 34, 38, 39, 41], and k-nearest neighbor (k-NN) is also used frequently [19, 27, 36, 40].

The second and the main part of this study, Chapter 4, is devoted to the use of the VAR model parameters for diagnostic classification. Various experimental setups with k-NN, support vector machine (SVM), and Gaussian mixture model (GMM) classifiers are designed and used with different data sets. The pathological data include asthma, COPD, bronchiectasis and ILD as the conditions to be diagnosed. In most of the experiments conducted in Chapter 4, the general case of healthy versus pathological classification is considered, however, this chapter in general differs from [36] and [40] both in terms of the features (VAR parameters as opposed to univariate AR) and the classifier algorithms (SVM and GMM besides k-NN). The studies in [38] and [41] perform binary classifications into healthy and pathological groups, however, they do not assume clinical setups as this study does in general since they consider single types of diseases for the pathological groups against the healthy. The only common point of [41] with this study is that it employs the VAR model, yet a prominent difference exists in terms of the selection of the optimal model, as addressed above.

As the next step in developing a diagnostic classifier, the discrimination between the obstructive and restrictive classes is considered. In a three-class scheme including the healthy group besides the two pathological types, a hierarchical framework (i.e., healthy versus pathological classification followed by obstructive versus restrictive) is proposed. Although the most comparable study is the one in [39] with its three-class schema, the classification is performed for the three classes at once in [39] as opposed to the hierarchical framework offered by this study. Moreover, the study in [39] addresses data from a single channel recording thus is limited to a single location on the posterior chest.

A special case, namely, asthma versus COPD discrimination, is also included in Chapter 4. Asthma is a complex respiratory disorder characterized by chronic inflammation of the airways, reversible airflow obstruction, bronchial hyper-responsiveness, and recurrent episodes of wheezing, breathlessness, chest tightness, and coughing [47, 48]. COPD is defined as a disease state characterized by airflow limitation that is not fully reversible [49], where the airflow limitation is usually progressive and is associated with an abnormal inflammatory response of the lungs to harmful particles or gases, primarily caused by cigarette smoking [49]. Since the symptoms of asthma and COPD overlap, misdiagnosis of the two diseases is not rare in clinical practice [50–53]. The only objective assessment of the condition can be done through a pulmonary function test called spirometry. Certain limits of spirometric parameters to differentiate between asthma and COPD are defined in the literature, however other limits are also used or have been suggested [50, 54]. Particularly in the absence of wheezing, the spectral differences between asthma and COPD [55] can hardly be detected via stethoscope auscultation. The study in [55] proposes that changes in asthma and COPD might be reflected in pulmonary sounds and can be objectively quantified via computerized sound analysis techniques. Although the study in [55] reports statistical differences between asthma and COPD in terms of percentile frequencies and the total spectral power calculated on pulmonary sound signals, it does not propose a classification algorithm to help diagnosis in clinical setup. Moreover, there does not exist any other study in the pulmonary literature that proposes a computerized method for acoustic based discrimination of asthma and COPD. Therefore, the classification into asthma

and COPD is considered to be the focus of one of the experiments in order to explore the discrimination of these two diseases based on the acoustic data.

Diagnostic classification implies the correct assessment of the existing pathology, hence requiring a multi-class framework including the largest possible variety of diseases besides the healthy group. The final part of Chapter 4 is on building the diagnostic classifier according to the results of the previous experiments, using the data set at hand with the largest variety of conditions, namely, healthy, asthma, COPD, bronchiectasis and ILD. There do not exist any studies in the literature that consider diagnostic classification where the correct assessment of the pathology is pursued among various types. Therefore, the final part of the chapter should be put aside from all those studies discussed here both in terms of the material and the methodology. Moreover, as the proposed diagnostic classifier is a combination of binary classifiers (healthy versus pathological and obstructive versus restrictive), the individual results of each such division step reported at the final part contribute to the novelty of this study as well in terms of the data sizes.

Finally, none of the studies in literature that deal with pulmonary sounds classification with the aim to assess pulmonary conditions employs GMM, moreover, none of them employs SVM at all. The objective of the studies that use GMM [24, 26, 35] is to recognize and classify wheezes and crackles, whereas the aim in using GMM in this study is to recognize and classify different disease conditions.

To summarize, the first purpose of this thesis study is to find the optimal VAR model of pulmonary sounds data that will eventually provide diagnostic value to the analysis. The second purpose is to compare classifier algorithms and schemes with different data sets with the eventual aim to build a diagnostic system for pulmonary sounds. The novelty of this study lies in the following aspects: (i) The proposal of new criteria to evaluate the goodness of fit of the VAR model specifically for pulmonary sounds, (ii) The employment of the VAR model that has been optimized via the proposed criteria in diagnostic classification, (iii) The assessment of the SVM and GMM classifiers which have not been adopted yet to recognize pulmonary conditions, and (iv)

The presentation of the diagnostic framework with the largest variety of conditions.

The remaining chapters of the thesis are organized as follows: In Chapter 2, the multi-channel pulmonary sounds data are introduced and the basic specifications of the data acquisition system are mentioned. The description of the VAR model, the definitions of the conventional and the proposed goodness of fit criteria, and the experimental procedure to find the optimal model are explained in Chapter 3. Chapter 4 gives brief information about the theories of the k-NN, SVM and GMM classifier algorithms, and elaborates on various designs of those classifier algorithms with different data sets, proceeding from a basic level towards the most developed, in a diagnostic framework. Finally, the conclusions, and the suggested perspectives for the future studies, are presented in Chapter 5.

2. DATA

The database referred to in this study has been acquired using the 14-channel pulmonary sound data acquisition and processing system that was designed and implemented in the Boğaziçi University Lung Acoustics Laboratory [56,57]. Pathological data have been acquired mainly in İstanbul Yedikule Teaching Hospital for Chest Diseases and Thoracic Surgery, from subjects who were diagnosed with various diseases by the adjunct specialists. Healthy subjects were non-smokers who had no history of any serious lung disease. The data acquisition procedure of this study has been approved by the Second Ethical Committee on Clinical Research of Istanbul (is in compliance with the Declaration of Helsinki).

The system is composed of 14 air-coupled electret microphones (SONY ECM-44 BPT) attached on the posterior chest wall (see Figure 2.1), an analog amplifier-filter unit (with a gain of 100 and a pass-band of 80 to 4000 Hz), a Fleisch type pneumatachograph (Validyne CD379) to measure the flow-cycle simultaneously for synchronization, a data acquisition card (NI DAQCard-6024E, 12-bit) for digitization, and a laptop computer to control the process (via an interface program implemented in LabVIEW) and to store the data. The data are sampled at a rate of 9600 samples per second, and an acquisition session lasts 15 seconds. A sample waveform is given in Figure 2.2. During the recording sessions, the subjects were sitting upright and wearing a nose clip while breathing through the mouthpiece of the flowmeter. The database has been formed over time, therefore the data sets used in the successive sections are augmented accordingly at each step. Moreover, they are arranged as to fit the aim of the particular experimental setup. The list of the data sets summarizing their properties is given below.

- DS0: 15 healthy and 15 pathological subjects (bronchiectasis common, 4 of them had also asthma, 2 of them COPD, 1 of them pneumonia).
- DS1: 20 healthy and 20 pathological subjects (bronchiectasis is the common diagnosis, 6 of them also have asthma, 3 have COPD and 1 has pneumonia

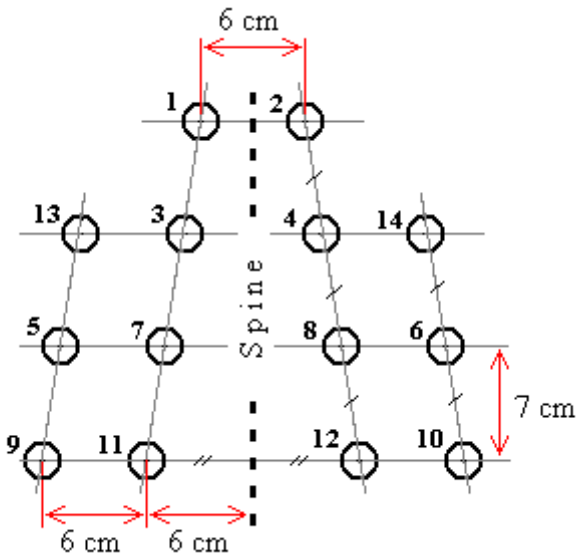


Figure 2.1. Microphone locations on the posterior chest wall.

besides bronchiectasis). This data set is the augmented version of DS0 by newly acquired data.

- DS2: 20 healthy (same healthy subjects as in DS1), 10 bronchiectasis (selected among the pathological subjects of DS1 such that the diagnosis is limited to bronchiectasis only), 10 ILD (no other accompanying diagnosis).
- DS3: 30 asthma (10 of DS2 + 20 selected among the newly acquired data), 20 COPD (10 of DS2 + 10 selected among the newly acquired data).
- DS4: 25 healthy, 60 asthma, 35 COPD, 15 bronchiectasis, 15 ILD. This data set is the augmented version of DS2+DS3 by newly acquired data.

As a pre-processing step for Chapters 3 and 4, the flow signal is divided into

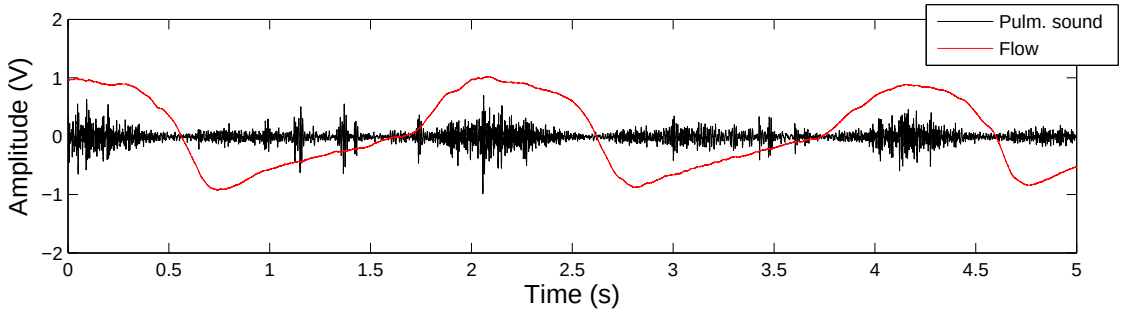


Figure 2.2. Sample waveform of one channel of pulmonary sound signals together with the flow rate signal.

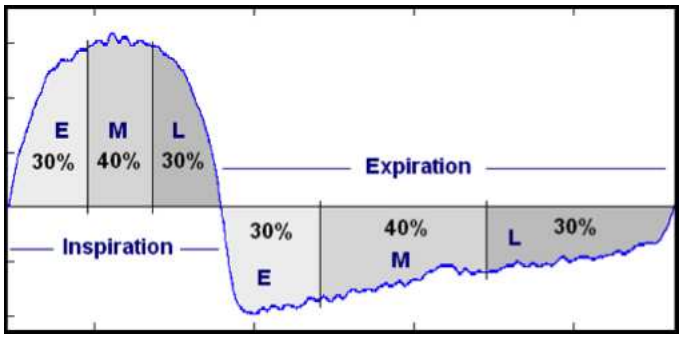


Figure 2.3. The scheme for the division of the flow-cycle into six respiratory sub-phases. E: Early, M: Mid, L: Late.

respiratory sub-phases, namely, early/mid/late inspiration/expiration. Early and late phases of inspiration (or expiration) represent 30 percent of total air volume inhaled (or exhaled) during one inspiration (or expiration) period, while the mid phase represents the remaining 40 percent (see Figure 2.3). Thereby, one complete inspiration-expiration cycle is composed of six respiratory sub-phases (flow-phases hereafter). The start and end indices of inspiration and expiration are determined automatically, such that the positive portions of the flow rate signal denote inspiration while the negative portions denote expiration. In order to avoid errors due to noise, a dead band of one per cent of maximum amplitude deviation is allocated around the zero level, hence positive and negative signs imply being above and below this band, instead of the zero level. The start and end indices of the flow-phases (early/mid/late) are in turn calculated using the area under each inspiration and expiration. Once these indices are determined on the flow signal, the 14-channel pulmonary sound data are labeled accordingly for each subject.

3. MULTIVARIATE MODELING

The sound production within the lungs with respiration does not obey a simple mechanism and is highly complicated due to physical reasons such as turbulence in the airways and different sound speeds as per tissue type. However, assuming a simple model to approach a complex system, especially when both the system and its inputs are unknown such as in this case, provides means to understand its general behavior and to adjust for the next evolutionary step with more complexity.[1]

Auto-regressive (AR) models are known to be useful in single-channel pulmonary sound signal analyses [32, 36, 37, 39–41, 44]. The mathematical assumption behind the ordinary univariate AR model is that the measured signal is a realization of a stochastic process that is auto-regressive in nature, i.e., each time sample of the signal can be written as a combination of the previous time samples up to an error term. The error term is assumed to be statistically white and Gaussian-distributed, and is to be minimized for model estimation. Physically, an AR filter is a finite impulse response (FIR) filter whose input is a previous snapshot of the signal, and whose output is the signal itself, plus white Gaussian noise (WGN) with the smallest possible variance. Equivalently, an AR filter is also an infinite impulse response (IIR) filter, whose input is small-variance WGN, and whose output is the measured signal.

Basically, the sound signal acquired at each microphone location on the chest wall is a filter output, where the filter is the path the sound follows in the thorax from the point that it is produced to the point that it is recorded. Since the sounds are sourced from *multiple points* within the lungs, the actual filter associated with each measurement point is a *combination of path filters*. If those filters are assumed to be linear, as in the AR model, a combination of them is reduced to a *single linear filter* that accepts as input a *combination of sounds* produced at multiple locations. According to the central limit theorem, additive combination of statistically independent distributions converges to normal (Gaussian) distribution as the number of components increases,

[1] "Essentially, all models are wrong, but some are useful" George E. P. Box

therefore the input signal of such a linear path filter would approximate Gaussian noise. Moreover, it would also tend to be white due to the superposition of independent natural phenomena. These two inferences verify that the WGN assumption with the AR model is a suitable assumption for the mechanism at hand, where an exception may occur in the existence of adventitious pulmonary sound components.

The ordinary univariate AR model estimates the linear filter associated with each microphone location on the chest wall individually. In multi-channel analysis of pulmonary sounds, it is still a possible approach to model each channel individually and then combine the model parameters for further processing. However, this approach ignores the inter-channel relationships, thereby losing a certain amount of the spatial information. When a set of time-correlated univariate time series is also interrelated in multivariate sense such as in this case, the improved approach is to replace the AR process assumption with its *multivariate* or *vector* version, namely, vector AR (VAR) process. Then in the model, each time sample of any of the channels is written as a linear combination of the previous time samples of both itself and the other channels. Throughout this study, VAR schema is adopted to model multi-channel pulmonary sounds data. By doing so, the model is improved to include the spatial relationships between the channel measurements on the chest wall, even though the actual internal filters still can not be estimated (due to the fact that the actual internal signals are unknown).

VAR processes are adopted in economy and meteorology to model trends and to predict future behaviors, and also have recently been employed in the analysis of pulmonary sounds [41] as multi-channel measurements have become prevalent. Two free arguments should be determined cautiously for the VAR model to be useful, namely, the model order and the sample size. Traditionally, these arguments are selected using several *goodness of fit* criteria defined in the literature (see Section 3.2.1) based on prediction error. If the aim of the VAR modeling is to predict the future values of the signal (as in economics or meteorology), then the conventional criteria are useful. However, if the aim is not prediction but the employment of the model parameters in further analysis (as in this case), then the criteria should be redefined accordingly.

The reason that the classification results are poorer with VAR parameters than with the other features in [41] might be that the sample size is chosen heuristically and the model order is chosen according to one of the conventional goodness of fit criteria based on the prediction error, despite the fact that the aim is classification. This chapter focuses on finding the best VAR model in this sense, and proposes new criteria specific to multi-channel pulmonary sound measurements with an aim to use the model parameters in a diagnostic system, namely, in classification.

For the aim of this study, the model that maximizes the ability of the VAR model parameters to represent the pulmonary sound data characteristics is regarded to be the most useful model. Accordingly, three new goodness of fit criteria are proposed, where the associated measures have been defined on calculated model parameters so as to reveal the inherent structure that pulmonary sounds have. The structure can be summarized as similar (different) characteristics for the same (different) phases of flow-cycle within a subject and across different subjects within a subject class, and different characteristics across two different classes (here, healthy and pathological). Therefore, the measures are intra-subject, inter-subject (intra-class), and inter-class measures, respectively, and the maximum scores imply that the model reveals the similarities and differences most distinctively.

Since a particular choice of sample size corresponds to a particular time duration depending on the sampling rate directly, the effect of sampling rate on the model parameters is also examined. This introduces the data sampling rate as the third argument, together with the model order and the sample size. If the results show that working with a lower sampling rate does not deteriorate the success of the model, downsampling is to be adopted to save the computational cost. Moreover, the effect of changing the sampling rate is still a point of interest for exploratory purposes, notwithstanding the fact that a single model fit requires significantly less computational effort once the model order and the sample size are fixed.

Pulmonary sounds are produced by turbulence due to air flow in the lungs, and flow rate levels are variable from subject to subject (being an habitual behavior and

depending on the conditions that may restrict the subject) and from data acquisition session to session (even for the same subject). Although the focus of modeling is on transfer characteristics of the lungs as a filter system affected by changing conditions (e.g., existence of a pathology), it still needs to be explored whether different flow levels affect the model order. The motivation is to be able to compare two model parameter sets reliably regardless of the fact that the associated flow levels are different. Accordingly, six flow normalization schemes are defined as divisions by functions of the flow rate signal, and applied on the acquired multi-channel data prior to the VAR modeling. Results with and without normalization are evaluated and compared using the proposed measures mentioned above.

To summarize, the main purpose of this chapter is to find the optimal set of the three model arguments, namely, the model order, sample size, and the sampling rate, which yields the most descriptive VAR model of pulmonary sounds data that will eventually provide diagnostic value to the analysis. The second purpose is to examine the effect of different flow levels on the model.

The novelty is in proposing new criteria to evaluate the goodness of fit of the VAR model. Although the proposed criteria are specific to a particular type of data and application, they can be adapted and employed for other types of data or applications as well, wherever the aim is to extract distinctive characteristics of the underlying system.

3.1. Description of VAR Model

A K-dimensional VAR process of order p (VAR(p) process) can be expressed as [58]

$$\mathbf{y}_n = \mathbf{v} + \mathbf{A}_1\mathbf{y}_{n-1} + ... + \mathbf{A}_p\mathbf{y}_{n-p} + \mathbf{u}_n, \quad n = 0, \pm1, \pm2, ... \tag{3.1}$$

where $\mathbf{y}_n \triangleq [y_{1n}, ..., y_{Kn}]^T$ is a $(K \times 1)$ random vector, $\mathbf{v} \triangleq [v_1, ..., v_K]^T$ is a fixed $(K \times 1)$ vector of intercept terms allowing for the possibility of a nonzero mean $E\{\mathbf{y}_n\}$, $\mathbf{A}_i$ are

fixed $(K \times K)$ coefficient matrices as

$$\mathbf{A}_i \triangleq \begin{bmatrix} \alpha_{11,i} & \cdots & \alpha_{1K,i} \\ \vdots & \ddots & \vdots \\ \alpha_{K1,i} & \cdots & \alpha_{KK,i} \end{bmatrix}, \quad i = 1, ..., p \tag{3.2}$$

and finally, $\mathbf{u}_n \triangleq [u_{1n}, ..., u_{Kn}]^T$ is a K-dimensional white noise, i.e., $E\{\mathbf{u}_n\} = 0$ and

$$E\{\mathbf{u}_{n_1}\mathbf{u}_{n_2}^T\} = \begin{cases} \Sigma_u, & \text{if } n_1 = n_2 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

where Σ_u can be either diagonal or not, i.e., the noise is not necessarily uncorrelated in space as it is in time. The process $\mathbf{y}_n$ is assumed to be stable and stationary, the latter assumption allowing $\mathbf{v}$ and $\mathbf{A}_i$ to be accepted as fixed (time invariant).

We have a K-variate time series available at hand (as a result of measuring K variables at successive time instances), which may be interpreted as a realization of a K-dimensional stochastic process. Then, VAR(p) modeling means estimating the model parameters as if the time series is generated by a stationary, stable VAR(p) process.

One way of estimating the model parameters is the multivariate (generalized) version of ordinary least squares (LS) solution. Let us assume that a time series $\mathbf{y}_1, ..., \mathbf{y}_N$ of the y variables is available, that is, we have a sample size of N for each of the K variables for the same sample interval. In addition, assume that p presample values for each variable, $\mathbf{y}_{-p+1}, ..., \mathbf{y}_0$, are available. Let us define [58]

$$\begin{aligned} \mathbf{Y} &\triangleq [\mathbf{y}_1, ..., \mathbf{y}_N] && (K \times N), \\ \mathbf{B} &\triangleq [\mathbf{v}, \mathbf{A}_1, ..., \mathbf{A}_p] && (K \times (Kp+1)), \\ \mathbf{Z}_n &\triangleq [1, \mathbf{y}_n^T, ..., \mathbf{y}_{n-p+1}^T]^T && ((Kp+1) \times 1), \\ \mathbf{Z} &\triangleq [\mathbf{Z}_0, ..., \mathbf{Z}_{N-1}] && ((Kp+1) \times N), \\ \mathbf{U} &\triangleq [\mathbf{u}_1, ..., \mathbf{u}_N] && (K \times N). \end{aligned} \tag{3.4}$$

Then, VAR(p) model in Equation 3.1 can be written compactly as

$$\mathbf{Y} = \mathbf{BZ} + \mathbf{U} \tag{3.5}$$

and the multivariate LS estimator of $\mathbf{B}$ is [58]

$$\widehat{\mathbf{B}} = \mathbf{YZ}^T(\mathbf{ZZ}^T)^{-1}. \tag{3.6}$$

The VAR setup assumes statistical stationarity, whereas the pulmonary sound data is inherently non-stationary (i.e., statistical characteristics of pulmonary sound data change across the flow-cycle). For the model to be valid, it should be fitted on short time intervals where the data can be assumed stationary. Accordingly, the sample size N of the model in Equations 3.4 should correspond to fragments of a second for pulmonary sound signals.

3.2. Goodness of Fit

As introduced in Section 3, two viewpoints to measure the goodness of model fit are considered. The traditional approach refers to *minimizing the prediction error* and is adopted either in minimum mean squared error (minimum MSE) sense or as proportion of variance explained. The proposed approach, on the other hand, is referred here as *maximizing the pattern coherence*, and corresponds to maximizing the ability of the model parameters to represent the pulmonary sound characteristics.

3.2.1. Minimizing the Prediction Error

Let the true model order be denoted by p, its estimate by $\hat{p}$ and the order of the fitted model by m.

Each one of the following four criterion measures (final prediction error (FPE), Akaike's Information Criterion (AIC), Hannan-Quinn criterion (HQ) and Schwarz cri-

terion (SC)) offers a solution for the problem of estimating the model order, once K and N are set [58]:

$$FPE(m) = \left[\frac{N + Km + 1}{N - Km - 1}\right]^K |\widehat{\Sigma}_u(m)|, \tag{3.7}$$

$$AIC(m) = ln|\widehat{\Sigma}_u(m)| + \frac{2mK^2}{N}, \tag{3.8}$$

$$HQ(m) = ln|\widehat{\Sigma}_u(m)| + \frac{2lnlnN}{N}mK^2, \tag{3.9}$$

$$SC(m) = ln|\widehat{\Sigma}_u(m)| + \frac{lnN}{N}mK^2, \tag{3.10}$$

where $\widehat{\Sigma}_u(m)$ can be estimated as

$$\widehat{\Sigma}_u(m) = \frac{1}{N}(\mathbf{Y} - \widehat{\mathbf{B}}\mathbf{Z})(\mathbf{Y} - \widehat{\mathbf{B}}\mathbf{Z})^T, \tag{3.11}$$

where $\widehat{\mathbf{B}}$ is the LS estimate of $\mathbf{B}$ (see Equation 3.6).

The estimated error covariance matrix can be written explicitly as

$$\widehat{\Sigma}_u(m) = \begin{bmatrix} \hat{\sigma}^2_{u_1}(m) & \hat{\sigma}_{u_{12}}(m) & \cdots & \hat{\sigma}_{u_{1K}}(m) \\ \hat{\sigma}_{u_{21}}(m) & \hat{\sigma}^2_{u_2}(m) & & \vdots \\ \vdots & & \ddots & \\ \hat{\sigma}_{u_{K1}}(m) & \cdots & & \hat{\sigma}^2_{u_K}(m) \end{bmatrix}, \tag{3.12}$$

and the data covariance matrix as

$$\Sigma_y = \begin{bmatrix} \sigma_{y_1}^2 & \sigma_{y_{12}} & \cdots & \sigma_{y_{1K}} \\ \sigma_{y_{21}} & \sigma_{y_2}^2 & & \vdots \\ \vdots & & \ddots & \\ \sigma_{y_{K1}} & \cdots & & \sigma_{y_K}^2 \end{bmatrix}. \tag{3.13}$$

After estimating the VAR(m) models for $m = 1, 2, ..., M$ and computing the corresponding measures, $\hat{p}$ is chosen such that

$$\hat{p} = \underset{m}{\operatorname{argmin}}\ f_c(m), \tag{3.14}$$

where $f_c(m)$ stands for the error criterion function that is used in calculation.

The terms with m, N and K in Equations 3.7 - 3.10 imply that there are complexity costs as m and K increase, the effect of which decreases by increasing N. This allows choosing the $\hat{p}$ that balances the trade-off between their individual effects. Without the correction terms, the prediction error reduces to the MSE

$$MSE(m) = |\widehat{\Sigma}_u(m)|, \tag{3.15}$$

which monotonically decreases with increasing model order m. In modeling K - dimensional data, this implies that for a fixed N the best model is the one with the highest model order m. Although this observation is straightforward, with an aim to compare the errors over different combinations of m, N, and the sampling rate without the effect of the correction terms, MSE is also examined besides the four measures in Equations 3.7 - 3.10.

One alternative error measure is the percentage of the data represented by its estimate, that is, the proportion of variance explained. It is also known as the *coefficient of determination* in statistics, and is defined as

$$R^2(m) = 1 - \frac{\sum_{k=1}^{K} \hat{\sigma}^2_{u_k}(m)}{\sum_{k=1}^{K} \sigma^2_{y_k}}, \tag{3.16}$$

where $\hat{\sigma}^2_{u_k}(m)$ and $\sigma^2_{y_k}$ are the diagonal elements of estimated error covariance and data covariance matrices, respectively, as shown in Equations 3.12 and 3.13.

$MSE(m)$ should be as small as possible, equivalently, $R^2(m)$ should be large (close to unity), to accept the model as a good fit. Geometrically, MSE corresponds to the volume that the prediction error vectors enclose in space (consider the determinant of the covariance matrix), while the coefficient of determination uses the sum of individual error variances and ignores the space correlations of the error vectors (consider the trace of the error covariance matrix). Except for this detail, both are equivalent in meaning, that is to say, they both accept that the information is carried in variance.

Note that the LS estimation is to maximize Equation 3.16 by definition, i.e., it already corresponds to choosing the parameter set that makes the total estimation error (sum of squares of error terms) minimum, hence the name *least squares*. In other words, $\widehat{\mathbf{B}}$ is the parameter set that minimizes the estimation (prediction) error for a particular data set once the model order and the sample size are fixed. There is an associated prediction error per model estimation, which is the minimum possible for the particular model fit, and the conventional criteria based on prediction error are about searching for the minimum of those minimum prediction errors in order to choose among the various possible model order and sample size combinations. In this context, the proposed argument is about using alternative criteria to choose among those combinations, instead of using the prediction error once more.

3.2.2. Maximizing the Pattern Coherence

Spatial and temporal relationships that multi-channel pulmonary sounds data have are expected to be in different ways for distinct types of flow-phases, for distinct subjects, or, for distinct conditions (healthy and pathological). Therefore, the computed model parameters are expected to follow some patterns if the model is to be

considered successful. Three new goodness of fit criteria are proposed and the associated measures are defined below accordingly. With this point of view, the information is not accepted to lie in the variance, rather, in the spatio-temporal relationships.

Each measure is defined as the ratio of between-group scatter over within-group scatter, and should be maximized to validate the success of the model, akin to as in Fisher's Linear Discriminant Analysis (LDA).

Let $P_{i,j}^{s}$ denote one of the six flow-phases of one subject (each full inspiration-expiration cycle of each subject is divided into six flow-phases, as described in Section 2), where s is the subject index, $s \in \mathcal{S}$, i is the *full cycle* (one inspiration followed by one expiration will be called a full cycle (FC) hereafter) index, $i = 1, 2, ..., (FCN)_s$, $(FCN)_s$ being the total number of FCs for subject s, and j is the flow-phase index from early inspiration to late expiration, $j \in \mathcal{J}$, $|\mathcal{J}| = 6$ ($|\cdot|$ denotes the size of the set).

Let $(WN)_{i,j}^{s}$ denote the maximum number of N-point segments fitting into $P_{i,j}^{s}$. Estimated parameter matrices $\hat{\mathbf{B}}$ are calculated as in Equation 3.6 and are in the form of $\mathbf{B}$ in Equations 3.4. Suppose the estimated mean vector $\hat{\mathbf{v}}$ is omitted and the new matrix is in the form of $\hat{\mathbf{B}} = [\hat{\mathbf{A}}_1, \hat{\mathbf{A}}_2, ..., \hat{\mathbf{A}}_p]$. More precisely, estimated parameter matrices are denoted as $\hat{\mathbf{B}}_{\{i,j\},w}^{s}$, $s \in \mathcal{S}$, $i = 1, 2, ..., (FCN)_s$, $j \in \mathcal{J}$, $w = 1, 2, ..., (WN)_{i,j}^{s}$. Note that $\hat{\mathbf{B}}_{\{i,j\},w}^{s}$ is a $(K \times Kp)$ matrix.

Let the characteristics of each flow-phase in each FC be summarized by

$$\mathbf{M}_{i,j}^{s} = \frac{1}{(WN)_{i,j}^{s}} \sum_{w=1}^{(WN)_{i,j}^{s}} \hat{\mathbf{B}}_{\{i,j\},w}^{s}. \tag{3.17}$$

Criterion 1: One type of flow-phase (e.g., mid inspiration) should have high resemblance to the same type of flow-phase throughout the same subject (it should be similar to every mid inspiration of the same subject from flow-cycle to flow-cycle);

whereas that type of flow-phase should have low resemblance to any other types (e.g., mid inspiration should not resemble early expiration or other types, in any of the flow-cycles.). This should hold for each subject.

The first measure can be defined mathematically as

$$\mu_1^s = \frac{S_{B,1}^s}{S_{W,1}^s}, \tag{3.18}$$

where

$$S_{B,1}^s = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}(EFCN)_j^s \times tr[(\mathbf{M}_{.j}^s - \mathbf{M}_{..}^s)(\mathbf{M}_{.j}^s - \mathbf{M}_{..}^s)^T] \tag{3.19}$$

is the between-group scatter and

$$S_{W,1}^s = \frac{1}{\sum_{j\in\mathcal{J}}(EFCN)_j^s} \times \sum_{j\in\mathcal{J}} \sum_{i=1}^{(EFCN)_j^s} tr[(\mathbf{M}_{ij}^s - \mathbf{M}_{.j}^s)(\mathbf{M}_{ij}^s - \mathbf{M}_{.j}^s)^T] \tag{3.20}$$

is the within-group scatter for subject s, where $|\mathcal{J}|$ is the size of set $\mathcal{J}$ and $|\mathcal{J}| = 6$, tr stands for the *trace* operator for matrices, $\mathbf{M}_{.j}^s$ indicates averaging $\mathbf{M}_{ij}^s$ over indices i, and $(EFCN)_j^s$ is the effective FC number for the j^{th} flow-phase of the s^{th} subject, such that $(EFCN)_j^s \leqslant (FCN)_s$. Inequality occurs if the sound acquisition for subject s did not start with *early inspiration* and/or did not end with *late expiration*, values of j depending on which flow-phases are lacking.

The maximum score per subject implies that the flow-phases with the same index are similar and those with different indices are dissimilar to the largest extent, within the subject.

Criterion 2: One flow-phase of one subject should be more similar to the same type of flow-phase of any other subject in the same group (Healthy/Pathological) than it is to any other type of flow-phase (of even himself/herself). (E.g., the mid inspiration of one subject should resemble the mid inspiration of another subject in the same group,

but not the early expiration of even himself/herself.) This should hold for each group (Healthy/Pathological).

Mathematically,

$$\mu_2^g = \frac{S_{B,2}^g}{S_{W,2}^g}, \tag{3.21}$$

where

$$S_{B,2}^g = \frac{|\mathcal{S}_g|}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} tr[(\mathbf{M}_{\cdot j}^{\cdot} - \mathbf{M}_{\cdot\cdot}^{\cdot})(\mathbf{M}_{\cdot j}^{\cdot} - \mathbf{M}_{\cdot\cdot}^{\cdot})^T] \tag{3.22}$$

and

$$S_{W,2}^g = \frac{1}{|\mathcal{J}||\mathcal{S}_g|} \times \sum_{j \in \mathcal{J}} \sum_{s \in \mathcal{S}_g} tr[(\mathbf{M}_{\cdot j}^{s} - \mathbf{M}_{\cdot j}^{\cdot})(\mathbf{M}_{\cdot j}^{s} - \mathbf{M}_{\cdot j}^{\cdot})^T], \tag{3.23}$$

where g is the group index, $g = 1, 2$, denoting the healthy and pathological groups, respectively, $\mathcal{S}_g$ denotes the set of subjects in group g, and $|\mathcal{S}_g|$ is the size of the set.

The maximum score per group implies that the flow-phases with the same index are similar and those with different indices are dissimilar to the largest extent, within the group.

Criterion 3: Parameter sets of one group (Healthy/Pathological) should be more similar to each other than they are to those of the other group. (E.g., parameters from the healthy group should be alike to each other but not alike to the parameters from the pathological group.)

This corresponds to

$$\mu_3 = \frac{S_{B,3}}{S_{W,3}}, \tag{3.24}$$

where

$$S_{B,3} = \frac{1}{2}\sum_{g=1}^{2} |\mathcal{S}_g| \times tr[(\mathbf{C}_g^{\cdot} - \mathbf{C}_{\cdot}^{\cdot})(\mathbf{C}_g^{\cdot} - \mathbf{C}_{\cdot}^{\cdot})^T] \tag{3.25}$$

and

$$S_{W,3} = \frac{1}{|\mathcal{S}|}\sum_{g=1}^{2}\sum_{s\in\mathcal{S}_g} tr[(\mathbf{C}_g^{s} - \mathbf{C}_g^{\cdot})(\mathbf{C}_g^{s} - \mathbf{C}_g^{\cdot})^T], \tag{3.26}$$

where $|\mathcal{S}|$ is the size of the set of subjects, $|\mathcal{S}_g|$ is the size of the set of subjects in group g, and $\mathbf{C}_g^s$ is formed by concatenating matrices $\mathbf{M}_{\cdot j}^s$ horizontally such that $\mathbf{C}_g^s = [\mathbf{M}_{\cdot\{j\in\mathcal{J}\}}^s]$, where $s \in \mathcal{S}_g$.

The maximum score per model fit implies that the representative parameter sets $\mathbf{C}$ with the same group index are similar and those with different indices are dissimilar to the largest extent. In other words, the two groups (healthy and pathological) are separable to the largest extent in terms of the representative parameter sets. Since the aim of VAR modeling is diagnostic classification, the most pertinent criterion is this one among all, albeit the first two are also necessary to complete the analysis and strengthen the choice of the third criterion in case of an agreement.

Some pathologies cause changes only in a specific flow-phase, or only in inspiration or expiration, instead of the whole pulmonary signal. This causes that the representative parameter sets of the pathological subjects resemble those of the healthy group to a larger extent than those of the other pathological types, in some portions of the set. Therefore, the actual values of the second and the third measures are sensitive to the type of diseases selected for the pathological group. However, that does not affect the validity of the criterion in the attempt to find the optimal model, because the relevant question is when the maximum value is reached, rather than what the actual values are. The similar portions of the parameter sets across the groups affect the measure in the same manner for each different choice of free variables (m, N and sampling rate), therefore are neutral elements in the quest for the total maximum diversity between

the groups.

3.3. Effect of Flow Levels on the Model

As introduced before, the data are segmented into short time intervals for the VAR model to be valid. Over those intervals, the data can be assumed stationary and the model parameters are constant. This arises the question how, or whether, the changes in the statistical characteristics of the data due to different flow levels, such as changes in the variance, are observed in the model parameters.

With an attempt to investigate the direct effects of different flow levels on the parameters, six types of normalization schemes are defined as divisions by functions of air flow rate (in liters per second). For comparison, the case without normalization is also included in the list as the first item.

(i) No normalization,
(ii) Division by the instantaneous flow rate,
(iii) Division by the absolute instantaneous flow rate,
(iv) Division by the square of the instantaneous flow rate,
(v) Division by the total air volume (in liters) inhaled or exhaled during the flow-phase,
(vi) Division by the total energy of the flow rate,
(vii) Division by the mean of the flow rate.

The functions are chosen as they are physiologically meaningful, and included here for a full list of possible normalization schemes, although some of them are mathematically equivalent to each other. For example, normalization schemes (v) to (vii) are actually divisions by scalars, therefore the estimated model parameters should not be effected by them due to the linearity of Equation 3.6. Moreover, the schemes (ii) and (iii) are expected to yield equal results since the only difference between them, which occurs for expiration periods, is effectively a sign change in the data matrices on both sides of Equation 3.5, leaving the parameter matrix unaffected.

These schemes are applied on the data prior to the VAR model fit, and the proposed criteria are calculated on the model parameters to observe the effects. Any of the normalization schemes ((ii) to (iv), ignoring the ineffective ones) is accepted to be useful if it decreases the within-class scatters and increases the between-class scatters so as to increase the value of the corresponding measure. The healthy group is expected to be relatively more uniform than the pathological group due to the wide variety that the latter depicts in terms of the type and spatial location of the pathology. Then, hypothetically, a decrease in the within-class scatter scores associated with the healthy group is still accepted as a success even if the corresponding pathological scores do not decrease. To observe the individual effects of normalization, the within-class and between-class scatters as defined in Section 3.2.2 are handled separately for the healthy and pathological groups, for the first two measures. The third measure, however, does not allow separate analysis by nature since it yields only one score for all subjects per model fit.

3.4. Experiments and Results

Pulmonary sounds data of 15 healthy and 15 pathological subjects (data set DS0, see Chapter 2) are used for the experiments. Accordingly, $|\mathcal{S}| = 30$, and $|\mathcal{S}_g| = 15$ for $g = 1, 2$, referring to the definitions in Section 3.2.2. The subjects and the data acquisition sessions to be included in the pathological group have been selected such that adventitious pulmonary sound components are absent, or at least in negligible amount, with an attempt to satisfy the model assumptions to the largest possible extent (see the third paragraph in Section 3).

To observe the effect of the data sampling rate, the 14-channel data are downsampled by d, $d = 1, 2, ..., 10$, where $d = 1$ corresponds to the case of no downsampling. Downsampling by d implies an equivalent of sampling the data with a sampling rate of $9600/d$ samples per second, i.e., keeping one out of every d samples after the data is digitally processed through a proper anti-aliasing filter.

Considering the solution in Equation 3.6, there is a maximum allowable p value

per choice of N, or, a minimum allowable N value per choice of p, since the calculation of the parameter matrix requires $\mathbf{Z}\mathbf{Z}^T$ to be non-singular, which in turn requires that $Km+1 < N$ (see Equations 3.4). The model orders $m = 1, 2, ..., 10$ are experimented. Letting $K = 14$ and $m = 10$, the constraint is found to be $N > 141$. Accordingly, the minimum sample size N is set to 200. Moreover, considering the typical flow-phase lengths, the maximum N is set to 1900. The step size is set to 50 samples from 200 to 1900.

For each $P^s_{i,j}$ as defined in Section 3.2.2, the data are downsampled by d, $d = 1, 2, ..., 10$, and a 14-variate N-point VAR(m) model is fitted on the N-point 50% overlapping segments of the data, where $m = 1, 2, ..., 10$ and $N = 200, 250, ..., 1900$. Calculations for $MSE(d, m, N)$ and $R^2(d, m, N)$ together with the other four conventional goodness of fit measures, and the three proposed measures $\{\mu_1, \mu_2, \mu_3\}(d, m, N)$ are carried out[2] as explained in Section 3.2 for all selected d, m and N values. Please note that d, m and N are also variable indices of every matrix and scalar beginning from Section 3.2.2, however dropped for clarity.

The best sets of model arguments according to the conventional goodness of fit criteria are given in Table 3.1. MSE, R^2, FPE and AIC choose the highest model order m and the smallest sample size N available ($m = 10$, $N = 200$), whereas HQ and SC choose a lower model order m and the largest sample size N possible ($m = 6$, $N = 1900$. Refer to Equations 3.7 - 3.10 for the correction terms that causes this difference). All those measures deteriorate for $d > 1$.

The upper two plots in Figure 3.1 depict the three proposed coherence-related measures in percentage values with respect to their maxima, for selected ranges of d, m and N ($d = 1$, $m = 1, 2, ..., 5$ and $N = 200, 250, ..., 850$). For each measure, the maximum occurs at a single point within the range. The sets of model arguments that yield those measure values above 99% of their maxima are given in Table 3.2, the figures in bold depicting the best sets. Among the conventional measures, R^2 is chosen

[2]Calculations are done using MATLAB v7.2 running on a Linux workstation with 16 CPUs each working at 3 GHz and with 32 GB RAM in total

Table 3.1. The best set of model arguments found using the conventional goodness of fit measures. d: downsampling ratio, m: VAR model order, N: sample size.

	d	m	N
MSE	1	10	200
R^2	1	10	200
FPE	1	10	200
AIC	1	10	200
HQ	1	6	1900
SC	1	6	1900

for comparison since its values can be comprehended more intuitively. R^2 values are also given in the lower plot in Figure 3.1 for $m = 2, 3$ and the selected range of N values.

Outside the range of Figure 3.1, all the three proposed measures deteriorate (therefore, not included in the figure). This implies that, the attempt to choose the optimal model according to the best values of the conventional measures would lead to poor values of the proposed measures. On the other hand, the figure verifies that choosing the optimal model according to the best values of the proposed measures leads to sufficiently good models in terms of prediction successes as well. To give numerical examples, if one of the selections given in Table 3.1 is adopted, e.g., $d = 1$, $m = 10$ and $N = 200$, μ_1, μ_2 and μ_3 drop to 10.5%, 26.4% and 41.5% of their maxima, respectively. The other selection, $d = 1$, $m = 6$ and $N = 1900$, yields μ_1, μ_2 and μ_3 at 38.9%, 50.3% and 65.8% of their maxima, respectively. Whereas, the selections given in Table 3.2 yield the minimum R^2 value to be 0.9978 (i.e., the selected models represent the data with a prediction loss of 0.22% at worst), implying that all of the choices in Table 3.2 are preferable both in terms of the conventional and the proposed goodness of fit criteria. According to the proposed criteria, the acceptable choices for the VAR model of pulmonary sounds data are limited and should be determined cautiously.

The first measure μ_1 reaches its maximum for $m = 3$ and $N = 600$, whereas μ_2

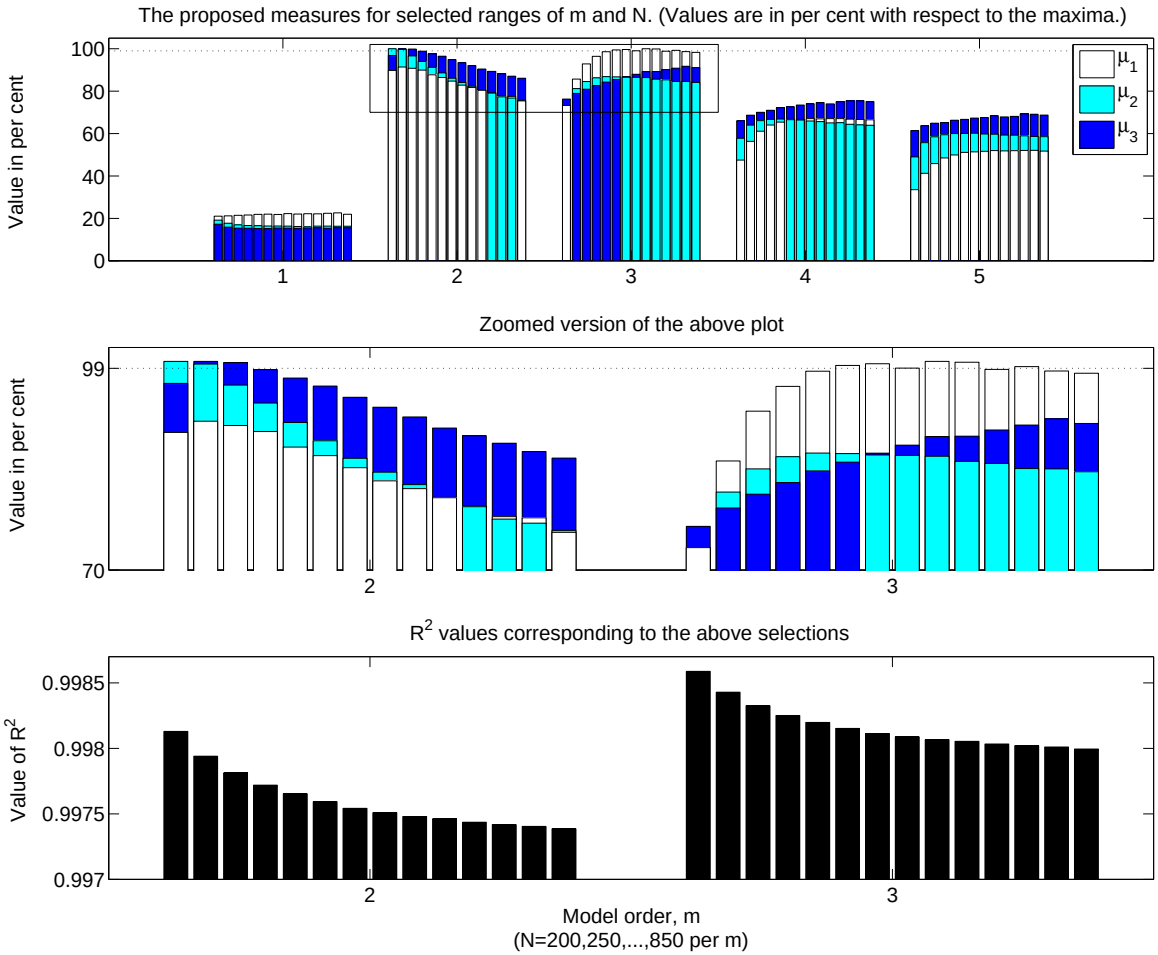


Figure 3.1. The proposed coherence-related measures in percentage with respect to their maxima, for selected ranges of m and N ($d = 1$). The dotted line is the 99 % threshold. $N = 200, 250, ..., 850$ in increasing order for each m lot in the x-axis.

and μ_3 reach their maxima for $m = 2$ (common for both), and $N = 200$ and $N = 250$, respectively. The sample size selection of μ_3, $N = 250$, also yields a μ_2 value above 99% of its maximum, therefore it is also acceptable in terms of μ_2 for practical purposes. Consequently, the best sets of arguments $\{d, m, N\}$ are determined in this study to be $\{1, 3, 600\}$ according to μ_1, and $\{1, 2, 250\}$ according to μ_2 and μ_3. Among the two sets, the latter is a better choice because it yields μ_1 at 91.3% of its maximum (a larger percentage than those of μ_2 and μ_3 when the former is selected, see Table 3.2). Hence, 250-point VAR(2) model is proposed to be an optimal VAR model for pulmonary sounds data. That is the optimal model selection of the third criterion, which is also in accordance with the fact that the third criterion is the most relevant

Table 3.2. Model arguments selected by the proposed goodness of fit criteria (those that yield the three measures above 99% of their maxima). The best sets are in bold. All the measures are in percentage values with respect to their own maxima.

	Argument selections			Measures (%)		
	d	m	N	μ_1	μ_2	μ_3
μ_1	**1**	**3**	450, 500, 550, **600**, 650, 750	100	86.4	89.2
μ_2	**1**	**2**	**200**, 250	89.8	100	96.8
μ_3	**1**	**2**	**250**, 300	91.3	99.6	100

criterion for the final aim of subject-based classification using the whole respiratory cycle.

As a preprocessing step for normalization, flow signals are first smoothed to eliminate the quantization noise. For this purpose, a cubic spline interpolator is adopted after experimental observations, since it proved superior to polynomial fits of order one to ten. A step size of 100 is chosen, and starting from the time index equal to the half of the step size, samples are picked to be used for interpolation. The cubic spline interpolated signal is a smoother version of the former with the same time duration, mean value, and standard deviation. For early and late inspiration (expiration), the flow rates below (above) 25% of the maximum (minimum) flow rate of that particular sub-phase are floored to an empirical threshold flow value (after smoothing). This is a preventive step against divisions by zero or near-zero values which would cause explosions near the start or end portions of the sub-phase, thus violating the stationarity assumption.

The six normalization procedures listed in Section 3.3 are applied on the data, after the segmentation into early/mid/late inspiration and expiration phases, prior to the model fit. Here the focus is on how the coherence-related measures change with flow normalization. To observe each one of the three measures, the maximizing set of arguments is used, i.e., the set of arguments $\{d, m, N\}$ that makes the measure maximum without normalization. Accordingly, after applying each normalization scheme,

three models are fitted on the data with different sets of arguments. When the set of arguments is $\{1, 3, 600\}$, μ_1 is calculated only. Similarly, when the set of arguments is $\{1, 2, 200\}$, μ_2 is calculated, and when it is $\{1, 2, 250\}$, μ_3 is calculated. The percentage changes of all scores with respect to their values before normalization are given in Table 3.3. In the table, the values in bold depict the most meaningful percentage changes, i.e., the maximal decreases in within-scatter scores, and the maximal increases in between-scatter scores and the three measures). An explanation about handling the healthy and pathological scores separately is given in Section 3.3.

The schemes (v) to (vii) yield results that are exactly equal to each other and also to the case without normalization as deduced in Section 3.3, therefore are not reported in Table 3.3. The schemes (ii) and (iii) are reported together since they yield exactly equal results. Observations on Table 3.3 reveal that all of the within-class scatters decrease with normalization, with maximum decrease when scheme (iv) is applied, making it the best in terms of within-class scatters. The between-class scatters on the other hand do not increase with normalization except for the two bold figures in the table. Yet, μ_1 and μ_2 improve with normalization, using normalization schemes (ii) (therefore (iii)) and (iv), respectively. The most consistent improvement occurs in terms of μ_2, and the most proper normalization scheme is (iv) accordingly. The third measure, μ_3, does not improve upon normalization although the associated within-class scatter decreases, therefore it is concluded that flow normalization does not increase the discriminative ability of the model parameters into healthy and pathological cases.

Table 3.3. Effect of normalization on the three measures in percentage changes with respect to the case without normalization. H: Healthy, P: Pathological, O: Overall, S_B: Between-class scatter, S_W: Within-class scatter, $\mu = S_B/S_W$.

		Normalization schemes	
		(ii),(iii)	**(iv)**
$\Delta S_{W,1}$	H	-0.69	**-1.09**
	P	-1.31	**-2.22**
	O	-1.14	**-1.91**
$\Delta S_{W,2}$	H	-0.02	**-0.07**
	P	-0.07	**-0.23**
	O	-0,05	**-0,16**
$\Delta S_{W,3}$	O	-0.08	**-0.22**
$\Delta S_{B,1}$	H	-0,47	-0,94
	P	-1,24	-2,79
	O	-0.98	-2.16
$\Delta S_{B,2}$	H	-0,01	-0,03
	P	0,20	**0,67**
	O	0,04	**0,14**
$\Delta S_{B,3}$	O	-0,17	-0,61
$\Delta \mu_1$	H	**0,06**	-0,41
	P	**0,29**	-0,25
	O	**0,15**	-0,35
$\Delta \mu_2$	H	0,01	**0,04**
	P	0,27	**0,91**
	O	0,07	**0,23**
$\Delta \mu_3$	O	-0,09	-0,40

3.5. Summary and Discussion

According to both the conventional and the proposed criteria, the case without downsampling has been determined to be the best choice in terms of the sampling rate. The reason for obtaining the highest scores for the original sampling rate can be explained as that the data are represented with more detail as the sampling rate increases, hence yielding a more correct modeling.

The conventional goodness of fit criteria select quite different m and N values from those selected by the coherence-related criteria, which yield models that perform poorly in terms of the latter. On the other hand, selecting the model arguments using the proposed coherence-related criteria causes only a slight deterioration in the conventional measures, therefore, is preferable in both respects.

The first measure differs from the second and the third in terms of the best set of arguments $\{d, m, N\}$ it chooses. The reason is that the former is a within-subject measure while the latter two are between-subject measures (see the definitions given in Section 3.2.2). Among the two sets, the simultaneous maximization of all the three measures is obtained for the latter, namely, $\{1, 2, 250\}$. In other words, 250-point VAR(2) model applied without downsampling has been determined to be the optimal VAR model to represent 14-channel pulmonary sound data. This selection coincides with the one optimized by the third criterion, which is the one that enhances the ability of the model to discriminate between the pulmonary conditions (healthy and pathological) by definition.

Flow normalization does not seem to improve the third proposed measure, i.e., the discriminative ability of the model parameters. The first and the second measures depict minor improvements, however, they do so for different normalization schemes. In general, the improvements, if any, are so small that normalization with respect to the flow rate is shown to be a non-crucial step before the VAR modeling of the pulmonary sounds data.

Any comparison of the results of this chapter with the literature can not be made since any study that proposes goodness of fit criteria specific to VAR modeling of pulmonary sounds is not available. However, a discussion on the results of two studies, [36] and [41], is included here to emphasize the importance of choosing the model order cautiously.

In [41], a binary classification problem into healthy and pathological groups (interstitial pulmonary disease only) is addressed. Four methods of feature extraction have been explored, two of which are univariate AR (referred as UAR) and multivariate AR (referred as MAR, same as VAR in this study) modeling. Model parameters have been used as inputs of a neural network classifier. The results with VAR have been shown to be the worst among the four, in terms of the thresholded final decisions. As a difference in [41] from this study, there are 25 microphones, each inspiratory period is divided into 30 temporal windows (therefore each segment is of varying length depending on the duration of the inspiration), and only the inspiration cycles are used. The VAR model order is chosen to be four according to Akaike's criterion. The best classification success is obtained using univariate AR, where the model order is chosen to be four as that of VAR, this time by observing the residual correlations.

In [36], binary classification into healthy and pathological (mixed group of both obstructive and restrictive types) groups is performed using univariate AR parameters with k-NN classifier. The microphone count is two, the model is fitted on data segments of 256 samples-long (fixed), and both the inspiration and expiration periods are used. The best order for the model is proposed to be six according to the resulting classification successes. The explicit correct classification rates for healthy and pathological classes for the other model orders are not presented separately. Moreover, the overall success rates are not written explicitly for the other model orders, however those are given in a graphical figure.

A direct comparison between the two studies is impossible to make since the channel counts, sampling rates, data segmentation procedures, contents of the subject groups (especially the variations in the pathological groups), and the classifier algo-

rithms are all different. However, the following discussion on the results of [36] and [41] with univariate AR would still give an idea about the importance of choosing the optimal model in accordance with the purpose of the analysis: The correct healthy and pathological classification rates in [41] for model order four (optimal based on residual correlations) are 75% and 93%, respectively. In [36], the correct healthy and pathological classification rates for model order six (optimal based on classification successes) are 85.7% and 100%, respectively, with an overall correct classification rate of 93.75%. The overall correct classification rate for model order four, on the other hand, can be read in the graphical figure to be 85%-90%, a lower score than that with the sixth order model. These results show that it is crucial to choose the optimal model according to the discriminative power of the parameters.

The proposed criteria are specific to pulmonary sounds analysis, nevertheless they can be adapted and employed for other types of data or applications as well, wherever the aim is to extract distinctive characteristics of the underlying system. Even though these criteria are to be used for pulmonary signals, where they can be employed without adaptation, one should note that the optimal set of model arguments would possibly turn out to be different, since the one suggested in this study is specific to the channel count and sampling rate of our data, possibly to the transfer characteristics and design of the hardware as well.

4. DIAGNOSTIC CLASSIFICATION

The aim of this chapter is to explore various classification schemes with different data sets with the eventual aim to build a diagnostic classifier. The features are the VAR model parameters estimated for the pulmonary sounds data using the optimal model proposed in the previous chapter. For clinical purposes, the detection of a deviation from the healthy case is valuable, therefore the binary classification problem into healthy and pathological cases is sufficiently interesting. Moreover, diagnosis requires differentiation between distinct types of pathologies, hence making the multi-class classification problem even more pertinent. This chapter follows a progressive route in terms of both methodology and material, i.e., it begins with the basic *healthy versus pathological* classification with the smallest data set and proceeds towards the *diagnostic* (multi-class) classification with the largest data set. Accordingly, at each intermediate step, the data set is augmented by enlarging with the newly acquired data and/or dividing into subsets for a particular experimental setup. Besides, the classifier algorithms are compared and/or improved.

In pattern recognition, k-NN classifier is a simple non-parametric method, which is widely used in biomedical data classification. As it predicts the unknown class label of a new instance depending on its closest neighbors in the labeled feature set, it does not assume any statistical models for the distributions of the classes, or does not impose any structured discriminants to separate them. Therefore it is preferable in cases where the data size is small, or where any assumptions on the distribution of the features in space are not available yet. Here, k-NN is adopted since it is widely used in pulmonary sounds analyses [19, 27, 36, 40].

As a second method, a kernel based discriminative classifier, namely, SVM algorithm is adopted. Although its basic linear version is already useful in many applications, the actual power comes from the usage of kernel functions, which provides the ability to handle linearly non-separable cases. SVM algorithm with a nonlinear kernel is equivalent to transferring the data to a higher dimensional space and performing

the linear separation there. Although SVM is already popular in speech and image recognition, bioinformatics, and various biomedical signal classification applications, it has not been adopted for classification of pulmonary sounds yet.

Finally, the GMM classifier is included as the third method. Unlike k-NN and SVM classifiers, GMM classifier is parametric (generative), assuming statistical models for features to yield probabilistic values for class memberships. The features are assumed to lie in space as a combination of distinct Gaussian-distributed subsets (rather than obeying a single Gaussian distribution) and the class labels are predicted using *maximum likelihood* or *maximum a posteriori* decision rules. GMMs have recently begun to be adopted to classify and label adventitious pulmonary sound components [24, 26, 35] rather than pulmonary conditions (e.g., healthy versus pathological).

In this chapter, Section 4.3.1 is on the binary scheme of healthy versus pathological classification with a relatively small data set. All the classifier algorithms are incorporated with some basic preferences to explore the comparative successes. Using the classifier algorithms selected according to the results of Section 4.3.1, Section 4.3.2 includes the obstructive versus restrictive discrimination in a three-class schema including the healthy group as well, with the data set augmented accordingly. Section 4.3.3 introduces the special case of asthma versus COPD discrimination with the classifier that has been observed to be the most successful in the previous two experiments. Since the symptoms of the two diseases overlap, they are often confused in clinical diagnosis (see Chapter 1), and there does not exist any study in the literature that proposes a method for their differential diagnosis using computerized pulmonary sounds analysis. Therefore, the inclusion of this special case in this chapter embodies particular importance. Finally, Section 4.3.4 presents the diagnostic classification scheme with the largest variety of conditions, namely, with the healthy, asthma, COPD, bronchiectasis and ILD classes, with some additional improvements in the classifier design. This final part of the chapter contributes to the novelty of this study at different levels, first, it proposes a diagnostic framework with multiple disease conditions together with the healthy case contrary to the studies in the literature, second, each binary classification step is novel in terms of the data size and the classifier (see Chapter 1).

To summarize, the purpose of this chapter is to compare and improve classifier algorithms with the eventual aim to build a diagnostic system for pulmonary sounds, and to assess their performances on particular data sets. The novelty lies in the following aspects: (i) The employment of the VAR model that has been optimized for pulmonary sounds for diagnostic purposes, (ii) The assessment of the SVM and GMM classifiers which have not been adopted yet to recognize pulmonary conditions, (iii) The inclusion of the special case of asthma versus COPD classification, and (iv) The observation of the diagnostic framework with the largest variety of conditions.

4.1. Classifier Algorithms

4.1.1. k-NN

Given a set $\mathcal{X}$ of class-labeled data vectors $\mathbf{x}$, k-NN algorithm predicts the class label of a new vector $\mathbf{y}$ by looking at the neighboring instances of $\mathbf{x}$. Let the distance d between two vectors $\mathbf{a}$ and $\mathbf{b}$ be defined as $d = ||\mathbf{a} - \mathbf{b}||$, where the distance metric can be any suitable choice. To predict the label of $\mathbf{y}$, the distances $||\mathbf{x}_t - \mathbf{y}||$ are calculated for each $\mathbf{x}_t \in \mathcal{X}$, $t = 1, 2, ..., |\mathcal{X}|$, then they are sorted in the ascending order, such that $d_1(\mathbf{y}) = min_t||\mathbf{x}_t - \mathbf{y}||$. The labels of vectors $\mathbf{x}_t$ with the closest k distances, $d_1, d_2, ..., d_k$, are considered, and the label of $\mathbf{y}$ is predicted by majority voting over those labels (therefore, k should be an odd number for binary classification).

4.1.2. SVM

For binary classification problems, SVM algorithm proposes a method to find the optimal separating hyperplane. Suppose we have data vectors $\mathbf{x}_t \in \mathcal{X}$, $t = 1, 2, ..., |\mathcal{X}|$, and the class labels are denoted by r_t, $r_t = +1$ if $\mathbf{x}_t \in C_1$ and $r_t = -1$ if $\mathbf{x}_t \in C_2$. Then the optimal separating hyperplane is such that

$$r_t(\mathbf{w}^T\mathbf{x}_t + w_0) \geqslant +1, \qquad \forall t, \tag{4.1}$$

where $\mathbf{w}$ is the normal vector of the hyperplane. If there is no hyperplane to separate the two classes, the one that incurs the least error is searched. Including slack variables $\xi_t \geqslant 0$ to represent the deviations from the margin, the relation becomes

$$r_t(\mathbf{w}^T\mathbf{x}_t + w_0) \geqslant 1 - \xi_t, \qquad \forall t. \tag{4.2}$$

To solve for $\mathbf{w}$, one should minimize

$$L_p = \frac{1}{2}||\mathbf{w}||^2 + F\sum_t \xi_t - \sum_t \alpha_t[r_t(\mathbf{w}^T\mathbf{x}_t + w_0) - 1 + \xi_t] - \sum_t \mu_t\xi_t \tag{4.3}$$

where α_t and μ_t are the Lagrange multipliers for the two constraints (for the one in Equation 4.2 and for $\xi_t \geqslant 0$, respectively), and F is the penalty factor to trade off complexity and data misfit [59] (the letter F is used here to represent the penalty factor, since the conventional letter C is reserved to stand for *class*).

For those cases where the separating boundaries are nonlinear, the idea behind SVM classification is to map the data to a new space through a nonlinear transformation and to find a separating hyperplane in this new space. The nonlinear transformation is done using a suitably chosen basis function. Let the d_1-dimensional $\mathbf{x}$ space be mapped to the d_2-dimensional $\mathbf{z}$ space as $\mathbf{z} = \boldsymbol{\phi}(\mathbf{x})$, where $z_i = \phi_i(\mathbf{x})$, $i = 1, 2, ..., d_2$ (d_2 is usually much larger than d_1). Then the new discriminant is written as $g(\mathbf{z}) = \mathbf{w}^T\mathbf{z} = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$, where now w_0 is inside $\mathbf{w}$ with the assumption that $z_1 = \phi_1(\mathbf{x}) \equiv 1$. The constraint $r_t(\mathbf{w}^T\mathbf{x}_t + w_0) \geqslant 1 - \xi_t$ in Equation 4.3 is replaced by $r_t\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_t) \geqslant 1 - \xi_t$, except for which the problem is the same. In the solution of this problem, inner product of these basis functions $\boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{y})$ appear. The idea in kernel machines is to replace this inner product by a kernel function $K(\mathbf{x}, \mathbf{y})$, which provides means for direct application in the original space instead of mapping the two instances to the new space and perform the inner product there [59].

It is a convex quadratic optimization problem to solve for $\mathbf{w}$ in Equation 4.3. Throughout this study, LIBSVM [60] is used to solve for $\mathbf{w}$, in other words, to train

(and validate) the SVM classifiers. For multi-class problems, LIBSVM follows one-against-one approach, which can be summarized as considering two of the classes at each time to predict a label for the instance, then deciding on the true label by majority voting. As the basis functions, linear, polynomial, and radial basis functions as defined in [60] are considered.

4.1.3. GMM

In the single Gaussian case the likelihood of vector $\mathbf{x} \in \mathcal{X}$ under class r is $p(\mathbf{x}|C_r) \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ where the mean vector $\boldsymbol{\mu}_r$ and the covariance matrix $\boldsymbol{\Sigma}_r$ are maximum likelihood (ML) estimates calculated over $\mathcal{X}_r$, such that $\mathcal{X} = \bigsqcup_r \mathcal{X}_r$, where $\bigsqcup$ denotes the exclusive union of sets (the intersections are empty).

In the GMM case, on the other hand, the likelihood $p(\mathbf{x}|C_r)$ is written as

$$p(\mathbf{x}|C_r) = \sum_{m=1}^{M} P(C_r^m) p(\mathbf{x}|C_r^m) \tag{4.4}$$

where M is the total number of components in the mixture, $P(C_r^m)$ is the probability of the m-th component, and $p(\mathbf{x}|C_r^m) \sim \mathcal{N}(\boldsymbol{\mu}_r^m, \boldsymbol{\Sigma}_r^m)$ is the likelihood of $\mathbf{x}$ under the m-th component ($\boldsymbol{\mu}_r^m$ and $\boldsymbol{\Sigma}_r^m$ are ML estimates calculated over $\mathcal{X}_r^m$, such that $\mathcal{X}_r = \bigsqcup_m \mathcal{X}_r^m$).

Then, given a vector $\mathbf{y}$ lying in the same space, the posterior probability of class r is calculated by Bayes' rule as

$$p(C_r|\mathbf{y}) = \frac{p(\mathbf{y}|C_r)P(C_r)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|C_r)P(C_r)}{\sum_r p(\mathbf{y}|C_r)P(C_r)}, \tag{4.5}$$

where $P(C_r)$ and $p(\mathbf{y})$ are called the class *prior* probability and the *evidence*, respectively, and the likelihood $p(\mathbf{y}|C_r)$ is calculated using the mixture model in Equation 4.4.

The components in the GMM are estimated using unsupervised learning (cluster-

ing) methods, the most popular being the expectation maximization (EM) algorithm, which is the one adopted in this study as well. To search for the best fit, mixtures with various component counts are considered for class distributions.

4.2. Performance Evaluation

The estimated VAR parameter matrices are denoted (in Section 3.2.2) as $\hat{\mathbf{B}}^{s}_{\{i,j\},w}$, where $s \in \mathcal{S}$ is the subject index, $i = 1, 2, ..., (FCN)_s$ is the FC index, $(FCN)_s$ being the total number of FCs for subject s, $j \in \mathcal{J}$ is the flow-phase index from early inspiration to late expiration ($|\mathcal{J}| = 6$), and $w = 1, 2, ..., (WN)^s_{i,j}$ is the index of the N-point segments in one flow-phase, $(WN)^s_{i,j}$ being the maximum number of such segments fitting into the flow-phase. The 14-channel pulmonary sounds data are divided (automatically) into the six flow-phases, then further divided (automatically) into 250-point 50% overlapping segments, and a VAR(2) model is fitted on each segment. Accordingly, $\hat{\mathbf{B}}^{s}_{\{i,j\},w}$ is in the form $[\hat{\mathbf{A}}_1, \hat{\mathbf{A}}_2]$ and is a (14×28) matrix. To build the feature space for classification, the matrices are converted to their vector forms, i.e., each matrix $\hat{\mathbf{B}}^{s}_{\{i,j\},w}$ is converted to a row vector $\hat{\mathbf{b}}^{s}_{\{i,j\},w} \in \mathbb{R}^d$, where $d = 14 \times 28 = 392$. Suppose a set $\mathcal{B}^s_j$ is defined as the set of these row vectors concatenated over w and i for a fixed j per s. Mathematically, $\mathcal{B}^s_j \triangleq \{\hat{\mathbf{b}}^{s}_{\{i,j\},w} : 1 \leqslant w \leqslant (WN)^s_{i,j}, 1 \leqslant i \leqslant (FCN)_s\}$.

Classifiers are trained on combinations of these $\mathcal{B}^s_j$ and validated on the complementing data set, which is also a combination of $\mathcal{B}^s_j$, where the rule of combination is determined by the particular classification scheme. At the lowest level, there are only feature vectors, and the success rate associated with each validation set corresponds to the percentage of correctly classified *vectors* (equivalently, correctly classified *segments* of the data). At the next level, the predicted labels (or, probabilistic scores) of all the vectors in $\mathcal{B}^s_j$ are combined to predict a label for $\mathcal{B}^s_j$, which bears the notion of correctly classified *flow-phases*. Finally, at the upmost level, the predicted labels (or, probabilistic scores) of the six flow-phases are combined to predict a single label for the *subject*. The three perspectives to evaluate the classification performances are explained below in more detail.

Note that the classifiers are trained for the six flow-phases separately in order to observe their relative pertinences. Since the data size is not large enough to partition the data as to have separate and sufficiently large training, validation and test sets, K-fold cross validation scheme is adopted throughout the experiments.

• *Perspective 1 (P1) - On the level of segments:* Each data segment, i.e., each one of the 250-point segments of the sound data that the VAR model is fitted on, is assigned a class label and successes are calculated at this level. Algorithmically, one class label is predicted for every vector $\hat{\mathbf{b}}$ (dropping the indices of $\hat{\mathbf{b}}^{s}_{\{i,j\},w}$ for simplicity) in the validation set, the percentages of correctly and incorrectly classified vectors are calculated per j per fold, and the averages over the folds yield the entries of the confusion matrix for j.

For k-NN and SVM, the classifier yields a class label per vector directly. For GMM on the other hand, the outcome is a probability value (likelihood). Then, the decision rule for a single $\hat{\mathbf{b}}$ is the maximum a-posteriori (MAP) decision rule

$$\log p(C_1|\hat{\mathbf{b}}) >^{C_1} \log p(C_2|\hat{\mathbf{b}}), \tag{4.6}$$

or, equivalently (refer to Equation 4.5)

$$\log p(\hat{\mathbf{b}}|C_1) + \log p(C_1) >^{C_1} \log p(\hat{\mathbf{b}}|C_2) + \log p(C_2), \tag{4.7}$$

where the likelihoods are calculated as in Equation 4.4 and the prior probabilities are calculated as the frequencies of classes in the training set.

• *Perspective 2 (P2) - On the level of flow-phases:* Segment decisions are combined to produce one common decision per flow-phase per subject and successes are calculated at this level. Algorithmically, one class label is predicted for $\mathcal{B}^s_j$ for every j and s, and the percentages of correctly and incorrectly classified $\mathcal{B}^s_j$'s per j yield the entries of the confusion matrix for j.

For k-NN and SVM, the common decision for a flow-phase is taken by majority voting over the labels of segments in the flow-phase. However, for GMM, this is done by comparing the class posterior probabilities given the flow-phase. That is, the decision rule is

$$\log p(C_1|\mathcal{B}_j^s) >^{C_1} \log p(C_2|\mathcal{B}_j^s), \tag{4.8}$$

in other words

$$\log p(\mathcal{B}_j^s|C_1) + \log p(C_1) >^{C_1} \log p(\mathcal{B}_j^s|C_2) + \log P(C_2). \tag{4.9}$$

Assuming statistical independency between observations (segments), we have

$$\log p(\mathcal{B}_j^s|C_r) = \log \prod_{\hat{\mathbf{b}}\in\mathcal{B}_j^s} p(\hat{\mathbf{b}}|C_r) = \sum_{\hat{\mathbf{b}}\in\mathcal{B}_j^s} \log p(\hat{\mathbf{b}}|C_r), \qquad r = 1, 2. \tag{4.10}$$

Then the MAP decision rule for this perspective becomes

$$\sum_{\hat{\mathbf{b}}\in\mathcal{B}_j^s} \log p(\hat{\mathbf{b}}|C_1) + \log p(C_1) >^{C_1} \sum_{\hat{\mathbf{b}}\in\mathcal{B}_j^s} \log p(\hat{\mathbf{b}}|C_2) + \log p(C_2). \tag{4.11}$$

Successive segments of pulmonary sound signals are not necessarily independent from each other, however, assuming statistical independence in Equation 4.10 is still appropriate because absolute entities are not aimed here. Rather, sums of log-likelihoods are used for comparison, as measures of superimposed distances from the class means.

• *Perspective 3 (P3) - On the level of subjects:* The six flow-phase decisions are combined to produce one common decision per subject, and successes are calculated at this level. That yields one confusion matrix for the overall classification.

For k-NN and SVM, the common decision for a subject is taken by majority voting over the flow-phases of the subject. For GMM, on the other hand, this is done by comparing the combined class posterior probabilities of flow-phases. As methods of

combination, four possibilities are considered:

(i) The combined decision for subject s is basically made by comparing the sums of class posterior probabilities over the flow-phases. The decision rule is then

$$\sum_{j\in\mathcal{J}} p(C_1|\mathcal{B}_j^s) >^{C_1} \sum_{j\in\mathcal{J}} p(C_2|\mathcal{B}_j^s). \tag{4.12}$$

For discriminative purposes, this is called P3-SOP (*sum of posteriors*) hereafter. In the extreme case that all posterior probabilities are either zero or one (the case sufficiently far from the class boundaries), this is equivalent to majority voting. Except for this extreme case, this combination method, as well as the succeeding ones, promises higher classification scores since it reduces the indecisive cases.

(ii) The combined decision for subject s is made by comparing the sums of *log-posterior* probabilities. The decision rule is then

$$\sum_{j\in\mathcal{J}} \log p(C_1|\mathcal{B}_j^s) >^{C_1} \sum_{j\in\mathcal{J}} \log p(C_2|\mathcal{B}_j^s). \tag{4.13}$$

This method is expected to increase the success scores in cases where the extreme case mentioned above (all probabilities are either zero or one) equates P3-SOP to majority voting, which is apt to create indecisive cases. Besides, the *addition of log-posteriors* is equivalent to the *multiplication of posteriors*, which corresponds to the probability that the six events occur at the same time if the events are statistically independent, the events being that all the six class indices are r given all the six flow-phases. (Note that the above decision rule is not equivalent to comparing sums of *joint* likelihoods of flow-phases, if the priors are not equal.) This combination method is called P3-SOLP (*sum of log-posteriors*) hereafter.

(iii) P3-SOP assumes equal weights for the class posterior probabilities of the six flow-phases. However, the flow-phases may be unequal in pertinence to predict the true class label of the subject, and assigning larger weights to those which are consistently more successful may increase the overall classification success. As a systematic way of determining the optimal weights, the SVM algorithm with

linear basis function is adopted.

The weighted version of the rule in Equation 4.12 is

$$\sum_{j\in\mathcal{J}} u_j p(C_1|\mathcal{B}_j^s) >^{C_1} \sum_{j\in\mathcal{J}} v_j p(C_2|\mathcal{B}_j^s). \tag{4.14}$$

Substituting $1 - p(C_1|\mathcal{B}_j^s)$ for $p(C_2|\mathcal{B}_j^s)$, it is reduced to

$$\sum_{j\in\mathcal{J}} w_j p(C_1|\mathcal{B}_j^s) >^{C_1} \sum_{j\in\mathcal{J}} v_j \tag{4.15}$$

where $w_j = u_j + v_j$. Note that it is equivalent in form to

$$\mathbf{w}^T\mathbf{x} >^{C_1} \tau, \tag{4.16}$$

where

$$\begin{aligned} \mathbf{w} &= [w_1, w_2, ..., w_6]^T, \\ \mathbf{x} &= [p(C_1|\mathcal{B}_1^s), p(C_1|\mathcal{B}_2^s), ..., p(C_1|\mathcal{B}_6^s)]^T, \\ \tau &= \sum_{j\in\mathcal{J}} v_j. \end{aligned} \tag{4.17}$$

The vector $\mathbf{w}$ of weights in Equation 4.16 is the normal vector $\mathbf{w}$ of the separating hyperplane in Equation 4.2, and τ in Equation 4.16 is found by LIBSVM as a result of the solution of Equation 4.3. This combination method is called P3-SVM hereafter. In the extreme case that all the weights turn out to be one, it is equivalent to P3-SOP.

(iv) As P3-SOLP is the logarithmic version of P3-SOP, the logarithmic version of P3-SVM, namely, P3-SVML is also worth mentioning. That is,

$$\sum_{j\in\mathcal{J}} u_j \log p(C_1|\mathcal{B}_j^s) >^{C_1} \sum_{j\in\mathcal{J}} v_j \log p(C_2|\mathcal{B}_j^s), \tag{4.18}$$

where Equation 4.16 still holds with

$$\begin{aligned}\mathbf{w} &= [u_1, u_2, ..., u_6, -v_1, -v_2, ..., -v_6]^T, \\ \mathbf{x} &= [\log p(C_1|\mathcal{B}_1^s), ..., \log p(C_1|\mathcal{B}_6^s), \log p(C_2|\mathcal{B}_1^s), ..., \log p(C_2|\mathcal{B}_6^s)]^T, \\ \tau &= 0.\end{aligned} \tag{4.19}$$

A reduction as done from Equation 4.14 to Equation 4.15 is not valid in this case because of the nonlinearity, therefore the dimensionality of the feature space is kept twice higher.

The format of the confusion matrix for the binary classification scheme with C_1 and C_2 is given in Table 4.1. In the table, n stands for the *number* of instances, and f for the *frequency* of instances. For example, $n_{2|1}$ denotes the number of instances misclassified as in C_2 while they are actually members of C_1, and $f_{2|1}$ is the frequency of those instances such that

$$f_{2|1} = 100 \times \frac{n_{2|1}}{n_{1|1} + n_{2|1} + n_{X|1}} \quad (\%), \tag{4.20}$$

where X is for indecisive cases when applicable.

Two success metrics are defined to summarize the whole confusion matrix in one single value with the aims to assess the performance of a classifier and to compare several classifiers. These are η and ζ, respectively, the former being the simple *total correct classification rate*, and the latter being the harmonic mean of the two class correct classification rates. Mathematically,

$$\eta \triangleq 100 \times \frac{n_{1|1} + n_{2|2}}{n_{1|1} + n_{2|1} + n_{X|1} + n_{1|2} + n_{2|2} + n_{X|2}} \quad (\%), \tag{4.21}$$

and

$$\zeta \triangleq \sqrt{f_{1|1} f_{2|2}} \quad (\%). \tag{4.22}$$

Table 4.1. The format for confusion matrices. $C/\hat{C}$: True/Predicted class label, C_1/C_2: Class 1/2, X:Indecisive.

	$\hat{C} = C_1$	$\hat{C} = C_2$	$\hat{C} = \text{X}$
$C = C_1$	$n_{1\|1}$ $(f_{1\|1})$	$n_{2\|1}$ $(f_{2\|1})$	$n_{X\|1}$ $(f_{X\|1})$
$C = C_2$	$n_{1\|2}$ $(f_{1\|2})$	$n_{2\|2}$ $(f_{2\|2})$	$n_{X\|2}$ $(f_{X\|2})$

The rationale behind defining two different metrics for the same purpose (to summarize the confusion matrix in a single value) is the following. The first metric η is defined especially for the first two of the three evaluation perspectives (P1 and P2) explained above. To examine the total correct classification rates of a particular classifier in box-plots with respect to different flow-phases and different GMM component counts, a linear metric is chosen. The second metric ζ on the other hand is defined especially for the third perspective (P3) and to compare different classifier schemes. Harmonic mean is chosen because the correct classification rates of both classes are desired to be high simultaneously. When used for P1 and P2 to complement the analysis with P3, ζ is calculated on the average confusion matrix over the six flow-phases.

Two more success measures are worth mentioning, namely, *recall* and *precision* rates, which are used in information retrieval. These are evaluated separately per class, and are intuitively appealing, therefore are incorporated here to observe the behavior of a particular classifier in more detail than with the scores η and ζ. For C_1, the recall rate is

$$Recall(C_1) = 100 \times \frac{n_{1|1}}{n_{1|1} + n_{2|1} + n_{X|1}} \quad (\%), \tag{4.23}$$

and is equal to the correct classification rate of C_1 $(f_{1|1})$. On the other hand, the precision rate for C_1 is

$$Precision(C_1) = 100 \times \frac{n_{1|1}}{n_{1|1} + n_{1|2}} \quad (\%), \tag{4.24}$$

and interpreted as the proportion of true members of class C_1 in all those instances

Table 4.2. The confusion matrix for SE and SP calculation. $C/\hat{C}$: True/Predicted class labels, C_1/C_2: Class 1/2, T/F: True/False, N/P: Negative/Positive.

	$\hat{C} = C_1$(N)	$\hat{C} = C_2$ (P)
$C = C_1$ (N)	TN	FP
$C = C_2$ (P)	FN	TP

classified as C_1. While the recall rate implies success in recognition, the precision rate implies reliability of claim. In medical applications such as the detection of an abnormal case, i.e., normal versus abnormal classification, a high prediction rate for the normal class should especially be pursued, since there should be as little doubt as possible when a subject passes the diagnostic test.

In the medical field, the popular success measures are *sensitivity* (SE) and *specificity* (SP). Either one corresponds to the correct classification rate of one of the two classes, the attributions depending on which class is defined as the positive one. Referring to Table 4.2, SE and SP are defined as

$$SE = 100 \times \frac{TP}{TP + FN} \quad (\%) \tag{4.25}$$

$$SP = 100 \times \frac{TN}{TN + FP} \quad (\%) \tag{4.26}$$

where TP/FP and TN/FN denote the numbers of true/false positives and negatives, respectively. In healthy versus pathological classification, since the existence of a pathology is a deviation from the normal situation, which is to be detected, the pathological class is usually the positive one. The co-evaluation of Tables 4.1 and 4.2 reveals the following equalities

$$SE = Recall(C_1) = f_{1|1},$$
$$SP = Recall(C_2) = f_{2|2},$$

the equalities of SE and SP holding if indecisive cases are ignored or included in the denominators in the definitions, and

$$Precision(C_1) = 100 \times \frac{TN}{TN + FN},$$
$$Precision(C_2) = 100 \times \frac{TP}{FP + TP},$$

for which the counterpart in medical nomenclature is called the *positive predictive value*.

Although the confusion matrices and the definitions are given here for binary classification, they can also be extended to the multi-class case. While the extension is straightforward for η, ζ, and the recall and precision rates, it is not for SE and SP since they form a complementary pair per classification. To be adapted, SE versus SP needs to be defined for each class versus all others, in which case they contradict the aim of multi-class classification. Therefore, SE and SP are mentioned here for the reader who is more comfortable with the medical nomenclature, however, they are only calculated for the healthy versus pathological classification, where the measures adhere to their true nature.

4.3. Experiments and Results

4.3.1. Binary Classification: Healthy/Pathological

For the first experiment, pulmonary sounds data of 20 healthy and 20 pathological subjects (data set DS1, see Chapter 2) are used. The first aim is to validate the ability of VAR parameters to discriminate between healthy and pathological groups. The second aim is to compare different algorithms and schemes of classifiers and assess the most successful ones that can be employed for further experiments.

At each validation fold, features of one subject are omitted from the data set and the classifier is trained on the remaining data, then the omitted subset is used to validate the classifier performance at that fold ($K = 40$ in the K-fold cross validation

scheme). All the classification schemes are performed for the six flow-phases separately.

In k-NN, values of $k = 1, 3, 5, ..., 99$ are experimented. Euclidean and Mahalanobis distances are adopted for the distance metric. In calculation of the Mahalanobis distance, two viewpoints are considered: (i) Two separate covariance matrices for the two classes, (ii) One covariance matrix for the overall data set. Moreover, both the diagonal and full versions of them are considered for exploratory purposes, increasing the total number of schemes to four. For SVM, linear, polynomial (of orders two to five), and radial basis function (RBF) are tested. GMMs are fitted with component counts from one (equivalent to the conventional *single Gaussian fit*) to ten.

Figure 4.1 summarizes the performances of five selected classifiers. Since the aim is to compare different classifiers, the success score ζ defined in Equation 4.22 is used in the bar plot. The five selected classifiers are:

(i) k-NN with Euclidean distance measure (kNN(E)),
(ii) k-NN with Mahalanobis distance measure using *full class covariances* (kNN(M)),
(iii) SVM with linear basis function (SVM(Lin)),
(iv) SVM with radial basis function (SVM(RBF)), and
(v) GMM.

In calculating the Mahalanobis distances, considering two separate class covariances and using them in their full form (rather than diagonal) maximizes the scores, therefore is coded briefly by kNN(M) as the representative of the four Mahalanobis schemes. SVM(Lin) and SVM(RBF) are the two SVM schemes selected to be the best among all after many experiments, therefore the two are included in the list while the others are omitted.

Note that the figure depicts the best versions of the five classifiers, i.e., the best k for k-NN, the best F (and γ) for SVM, and the best component count for GMM. The best k-NN performances are yielded with $k = 1$. The optimal penalty factor is $F = 1000$ (optimization range is $F = 10^{-5}, 10^{-4}, ..., 10^{5}$) for SVM(Lin). For SVM(RBF), the

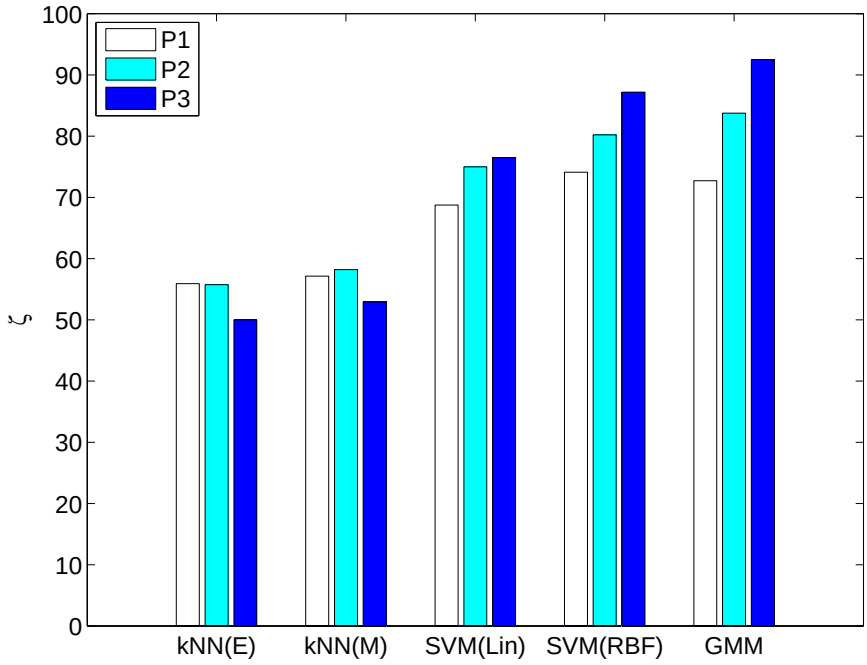


Figure 4.1. Bar-plot of ζ scores from P1, P2 and P3, for the five most successful schemes in healthy versus pathological classification.

optimal $F = 100, 1000$ (exactly equal confusion matrices for the two) and the optimal $\gamma = 0.1$ (optimization range is $\gamma = 10^{-5}, 10^{-4}, ..., 10^{2}$). The values for GMM in the figure are for component count five.

Mahalanobis distance measure in k-NN yields equivalent results to Euclidean. As for the SVM basis functions, the RBF has higher scores than the linear. Observation of the figure reveals that k-NN is in general poor in separating the two classes, whereas, SVM is considerably successful. The highest success is observed for GMM, especially in P2 and P3, the latter being the most pertinent perspective for the evaluation of performances since it is subject-level. Therefore, the best classifier algorithm among all is concluded to be the GMM. Table 4.3 depicts the numerical values explicitly (the confusion matrices from P1 to P3, together with the recall and precision rates of the two classes, SE and SP values, and the two success scores η and ζ, all of which are calculated on the confusion matrix from P3). kNN(E) and kNN(M) are comparable to each other in terms of η and ζ, albeit in opposite tendencies in terms of recall, precision, SE and SP. Since η and ζ are rather low, a further comparison including the two is skipped. SVM(RBF) has all the scores higher than SVM(Lin). While that

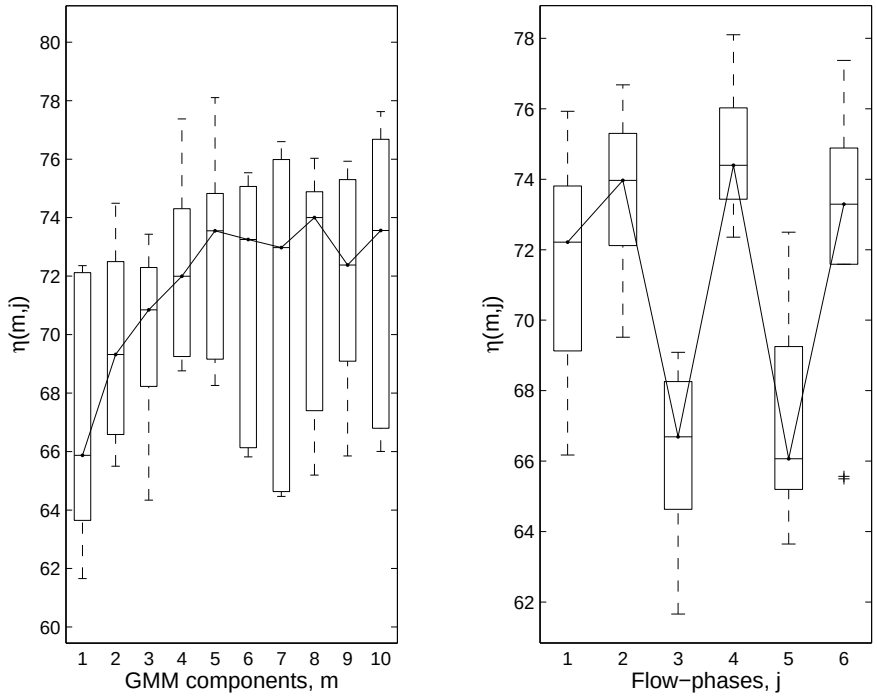
Table 4.3. Healthy versus pathological classification results. P1, P2 and P3: Perspectives 1, 2 and 3. C / $\hat{C}$: True / Predicted class label, H: Healthy, P: Pathological.

		k-NN (E)		k-NN (M)		SVM (Lin)		SVM (RBF)		GMM	
		$\hat{C}=H$	$\hat{C}=P$	$\hat{C}=H$	$\hat{C}=P$	$\hat{C}=H$	$\hat{C}=P$	$\hat{C}=H$	$\hat{C}=P$	$\hat{C}=H$	$\hat{C}=P$
P1	$C=H$	80.3	19.7	47.5	52.5	76.8	23.2	81.7	18.3	78.9	21.1
	$C=P$	61.1	38.9	31.3	68.7	38.5	61.5	32.8	67.2	33	67
P2	$C=H$	90.8	9.2	46.7	53.3	83.3	16.7	86.7	13.3	85	15
	$C=P$	65.8	34.1	27.5	72.5	32.5	67.5	25.8	74.2	17.5	82.5
P3	$C=H$	100	0	35	65	90	10	95	5	90	10
	$C=P$	75	25	20	80	35	65	20	80	5	95
		Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
	Healthy	100	57	35	63.6	90	72	95	82.6	90	94.7
	Pathological	25	100	80	55.2	65	86.7	80	94.1	95	90.5
	SE/SP	25 / 100		80 / 35		65 / 90		80 / 95		95 / 90	
	η / ζ	62.5 / 50		57.5 / 52.9		77.5 / 76.5		87.5 / 87.2		92.5 / 92.5	

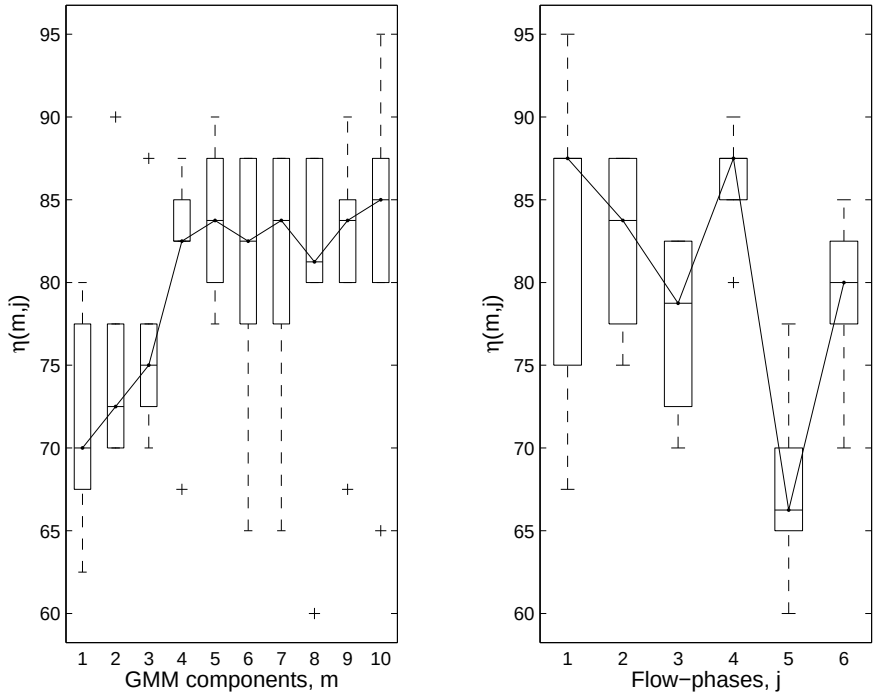
renders SVM(RBF) more appealing, two drawbacks render it dispensable for further experiments: (i) A high computational cost even for learning a single classifier, (ii) Two hyper-parameters to optimize (γ as well as the penalty factor F). Therefore, the linear basis function is also considered for the next experiment together with the RBF, to observe the performances with a different data set.

SVM algorithm is used here in its original form, i.e., it predicts class labels for instances according to which side of the hyperplane they lie (either in the original space or in the kernel space). Alternatively, the individual distances of support vectors to the hyperplane can be considered, even, can be turned into probabilities, which would give more information than two discrete class labels, which in turn would possibly increase the success of the method. Nevertheless, these are still derivations of the original algorithm, with more other assumptions and estimations. As the Gaussian approach is already probabilistic and its flexibility is increased with the mixture model approach, a search for other variants of SVM than those included here is kept out of scope of this thesis study.

As GMM proved to be the most successful classifier, additional observations for healthy versus pathological classification can be done using the GMM results. Calculating η values for the six flow-phases, we have η as a function of j, and also of m (m being the index for the GMM component count hereafter). Figure 4.2 depict the box-plots of $\eta(m, j)$ for P1 (upper) and P2 (lower), versus m (left) and j (right). The clear observation made in Figure 4.2a is that the first, second, fourth and sixth flow phases (early and mid inspiration, and early and late expiration) are notably more successful compared to the third and fifth (late-inspiration and mid-expiration), and the clear observation made in Figure 4.2b is that the GMM component counts greater and equal to four are notably more successful than the smaller ones. In the lower subfigure, although the first, second, fourth and sixth flow-phases do not depict the same breakthrough as in the upper, a closer observation reveals that the median scores of the four flow-phases are greater and equal to a threshold (80%) while those of the remaining two are below.



(a) η from P1 (the segment-level evaluation perspective).



(b) η from P2 (the flow-phase-level evaluation perspective).

Figure 4.2. Box-plots of η scores, versus GMM component count (on the left), and versus flow-phase index (on the right).

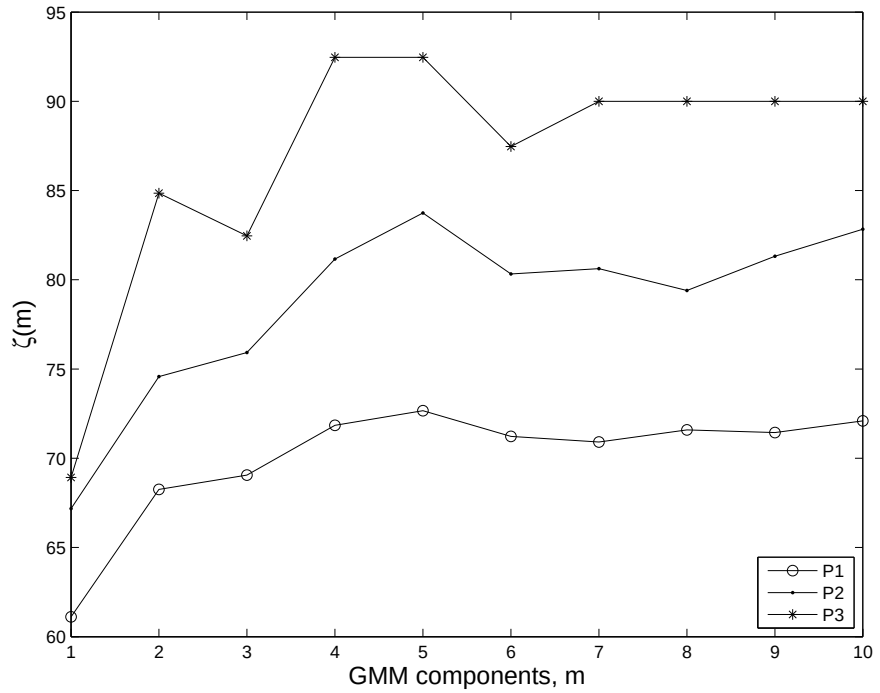


Figure 4.3. The ζ scores from P1, P2 and P3 versus GMM component count.

Figure 4.3 plots $\zeta(m)$ versus m from the three perspectives, P1 to P3. From P3, the most successful models are those with four and five mixture components, the latter being the best in terms of P1 and P2 as well. For component counts above five the successes decrease or stay more or less constant, and for those below four (or five for P1 and P2) the successes also decrease as the component count decreases. Moreover, the scores increase as we go from P1 to P3, i.e., combining the individual decisions to reach a single decision on behalf of the group increases the diagnostic ability of the classifier. This is also the case with SVM, leaving k-NN as the only exception among the algorithms that are subject to this experiment (see Figure 4.1).

This experiment verifies that VAR parameters are successful to model pulmonary sounds data, and reveals that the GMM classifier is successful to separate the healthy and pathological classes represented by VAR parameters. However, one should keep in mind that the optimal component count is specific to the distribution of the classes in the feature space, therefore may turn out to be different for other diagnostic classification setups with different pathological groups.

4.3.2. Three-class Classification: Healthy/Obstructive/Restrictive

Most of the pulmonary diseases accompanied with auscultation findings are grouped under two main categories, namely, obstructive and restrictive (see Chapter 1). Before proceeding to the problem of differentiating between diseases within the same category, differentiating between these two distinct categories is considered. For the experiments, bronchiectasis (of obstructive type) and interstitial pulmonary disease (of restrictive type) are selected as the representative diseases of their categories. Accordingly, 40 subjects are used, 20 of them being healthy, 10 of them being diagnosed with bronchiectasis and the remaining 10 with interstitial lung disease (ILD) (DS2, see Chapter 2). The aims of this section are to observe the classification performance in general with the two sub-categories of pathology, and to compare different designs of classifiers to obtain the best scores. Accordingly, eight different schemes in total, as variants of SVM and GMM, are defined and experimented.

First, the healthy versus pathological classification is considered for a comparison of the results with those of the previous section, and the scores are updated for the new data set. For the comparison, the three most successful classifiers of the previous section, namely, SVM(Lin), SVM(RBF) and GMM, are employed in the 40-fold cross validation scheme as before. The resulting confusion matrices and all the related success scores are given in Table 4.4. With the new data set, the scores of SVM(Lin) are observed to be improved, approaching those of SVM(RBF). Moreover, the confusion matrices of both SVM schemes are observed to be more balanced (compare SE versus SP). GMM yields the highest scores among all, as before. Since the performances of the two SVM schemes are comparable to each other, only one of them is to be chosen besides GMM for further experiments.

Although the power of SVM is based on the usage of kernel functions, one of the most popular being the RBF, basic linear SVM is also useful in certain applications, especially with high dimensional feature spaces. Actually, the superiority of the RBF over the linear is expected to get more subtle as the dimensionality increases. Here, the linear basis function is almost equally successful as RBF, and SVM(RBF) introduces

Table 4.4. Healthy versus pathological classification results for the three-class classification scheme. P1, P2 and P3: Perspectives 1, 2 and 3. $C/\hat{C}$: True/Predicted class label, H: Healthy, P: Pathological.

		SVM (Lin)		SVM (RBF)		GMM	
		$\hat{C} = H$	$\hat{C} = P$	$\hat{C} = H$	$\hat{C} = P$	$\hat{C} = H$	$\hat{C} = P$
P1	$C = H$	78.5	21.5	79.5	20.5	82.3	17.7
	$C = P$	31.5	68.5	32.8	67.2	29.1	70.9
P2	$C = H$	88.3	11.7	87.5	12.5	87.5	12.5
	$C = P$	21.7	78.3	26.7	73.3	20	80
P3	$C = H$	90	10	90	10	90	10
	$C = P$	20	80	15	85	10	90
		Recall	Precision	Recall	Precision	Recall	Precision
	Healthy	90	81.8	90	85.7	90	90
	Pathological	80	88.9	85	89.5	90	90
	SE/SP	80 / 90		85 / 90		90 / 90	
	η / ζ	85 / 84.9		87.5 / 87.5		90 / 90	

a minor improvement over SVM(Lin) at the expense of a much higher computational cost (see the drawbacks mentioned in the previous section). Therefore, SVM(RBF) is abandoned in the subsequent analyses, and the code "SVM" simply refers to SVM(Lin) hereafter. In the following paragraphs, the design of the various classifier schemes is explained in more detail.

As the number of classes is three, the following two approaches are possible: (i) Treating all the classes at once and perform a multi-class classification, (ii) Classifying the data into healthy versus pathological, then the pathological group into obstructive and restrictive (into bronchiectasis and ILD), which corresponds to performing two successive binary classifications. The former will be referred to as *all at once* and coded with an "O", and the latter will be referred to as *hierarchical* and coded with a "T" of *tree* (see Figure 4.4). In the *all at once* case, SVM follows one-against-one approach (see Section 4.1.2).

Combining the six flow-phases in the feature space prior to the classification is also considered. Although in all of the experiments, the classifiers are trained separately for the six flow-phases to inspect their relative pertinences, the combination is done here for exploratory purposes. Accordingly, the two approaches are: (i) Training the classifier for each respiratory flow-phase separately, with a later combination at the decision level, (ii) Combining the flow-phases at the feature level and training the classifier on the new feature space (whenever applicable). The former is in the original feature space (of dimensionality 392), therefore is coded with number "1". The latter, on the other hand, is coded with number "6" since the feature space has six times higher dimensionality (392x6). For GMM, the higher dimensionality version is not applicable since the number of feature vectors are insufficient to estimate the covariance matrices. SVM on the other hand is expected to work even better as dimensionality increases.

For the pre-combination of the six flow-phases at the feature-level, another pair of assumptions is considered: (i) Concatenating the feature vectors directly, i.e., by direct concatenation of vectors having the same order indices in the subsequent flow-phases, or, (ii) Concatenating the averages of the feature vectors in the flow-phases. Averaging

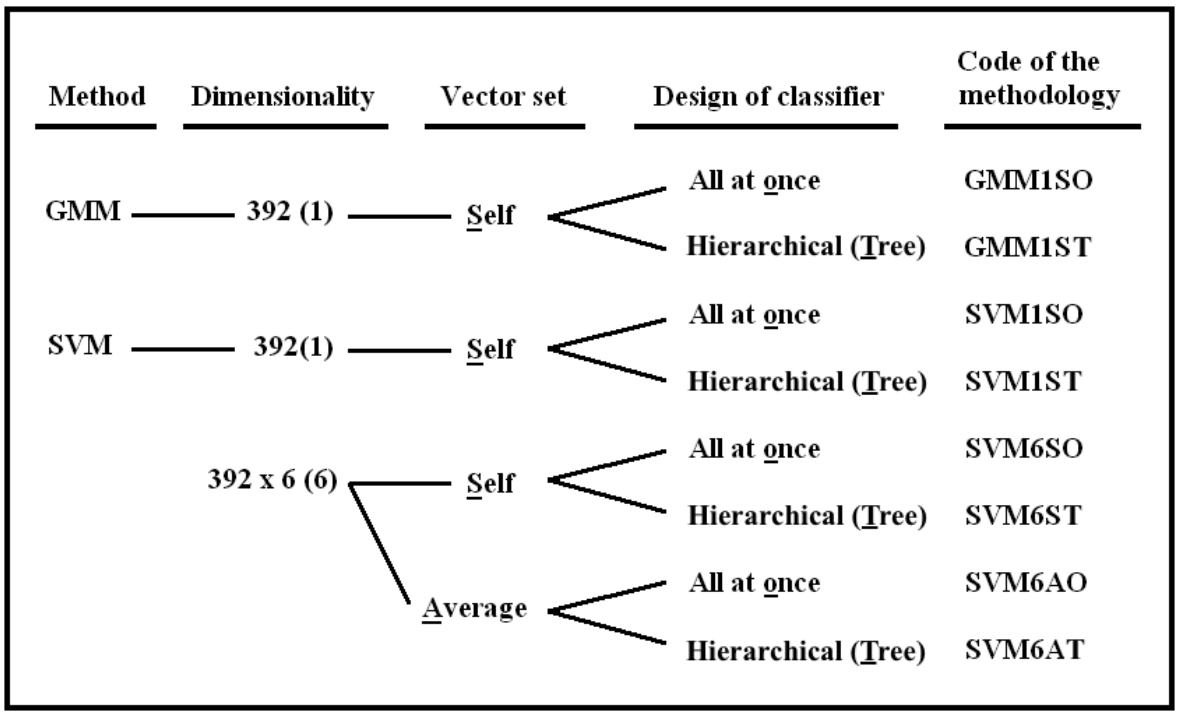


Figure 4.4. The diagram depicting the design of the classifiers and the corresponding letter codes.

is also done for exploratory purposes, as a precaution to prevent the concatenation of asynchronous instances erroneously, and also to see whether the average vectors are sufficient to summarize the classes. Direct use of vectors will be referred to as *self* and coded with an "S", and averaging them will be referred to as *average* and coded with an "A". All pairs of assumptions make $2 \times 2 \times 2 = 8$ schemes in total, which are summarized in the diagram in Figure 4.4.

The bar plot of ζ is given in Figure 4.5 for the eight classifier schemes, in the order from top to bottom in the diagram of Figure 4.4. Note that P2 becomes irrelevant for those cases where pre-combination of flow-phases is performed, therefore the corresponding bars are absent in the figure. The results of GMM schemes are for component count five for GMM1SO and seven for GMM1ST. The optimal SVM penalty factors are searched over the range $F = 10^{-7}, 10^{-6}, ..., 10^{7}$. In the figure, $F = 10^{5}$ for the first two SVM schemes (SVM1SO and SVM1ST), and they are in various values in the range $F = 0.001$ to $F = 0.1$ for the remaining four (SVM6-SO/ST/AO/AT).

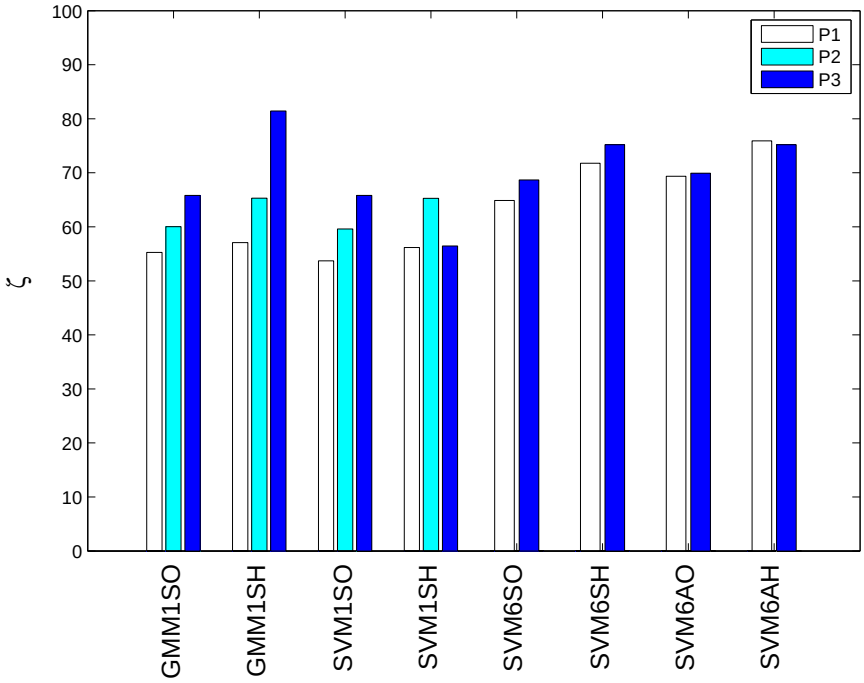


Figure 4.5. Bar-plot of ζ scores from P1, P2 and P3, for the eight classifier schemes in the three-class classification.

Considering that the average features are used, the last two classifiers perform better than expected, and they are comparable to their counterparts where the dimensionality is again 392x6 yet feature vectors are used directly as themselves (no averages are taken). The reason might be that the average feature vectors are sufficient to summarize (or imitate) the distribution of the classes in the feature space. In other words, they lie in the feature space in such a manner that the optimal separating hyperplane is not distorted to the extent of violating the boundary. SVM schemes with dimensionality code 6 are rather successful, having the highest P1 and P3 scores of all except for P3 of GMM1ST. Recalling that the maximum optimal penalty factor is $F = 0.1$ for these four schemes whereas it is $F = 10^5$ for the other two SVMs (those with the dimensionality code 1), one can conclude that when the flow-phases are pre-combined at the feature-level, the discrimination of the two classes by SVM is easier.

An important observation is that the hierarchical versions are consistently better than the counterparts, where the only exception is holding for P3 of SVM1ST (P1 and P2 of SVM1ST obey this general tendency). A closer observation on the predicted class labels of subjects (not reported here explicitly) reveals that the indecisive cases

Table 4.5. Three-class (HBI) classification confusion matrix for GMM1ST (7 components). $C/\hat{C}$: True/Predicted class label, H: Healthy, B: Bronchiectasis, I: ILD

		$\hat{C}=H$	$\hat{C}=B$	$\hat{C}=I$
P1	$C=H$	82.3	11.9	5.8
	$C=B$	45.2	31.4	23.4
	$C=I$	12.8	15.3	71.9
P2	$C=H$	87.5	8.3	4.2
	$C=B$	36.7	36.7	26.6
	$C=I$	3.3	10.0	86.7
P3	$C=H$	90	10	0
	$C=B$	20	60	20
	$C=I$	0	0	100
		Recall	Precision	
	Healthy	90	90	
	Bronchiectasis	60	75	
	ILD	100	83.3	
	η / ζ	85 / 81.4		

in the combined decisions of flow-phases are incidentally very frequent for this scheme.

Although GMM1ST has comparable P1 and P2 scores with its three neighbors (the moderately successful schemes GMM1SO, SVM1-SO/ST), it makes an apparent breakthrough at P3, attaining the highest score among all. Although the last four SVM schemes (those with the dimensionality code 6) are considerably more successful than GMM1ST in P1, the discriminative ability of GMM1ST increases at each step of decision combination, eventually surpassing the others in P3. The power of GMM comes from that it is a probabilistic model, and going from P1 to P3 corresponds to combining the probabilistic scores rather than the discrete decisions.

As discussed above, GMM1ST is selected to be the most successful among all

Table 4.6. Three-class (HAI) classification confusion matrix for GMM1ST (7 components). $C/\hat{C}$: True/Predicted class label, H: Healthy, A: Asthma, I: ILD

		$\hat{C}=H$	$\hat{C}=A$	$\hat{C}=I$
P1	$C=H$	83.9	11.6	4.4
	$C=A$	31	46.6	22.4
	$C=I$	10.7	3.1	86.2
P2	$C=H$	89.2	6.7	4.2
	$C=A$	18.3	53.3	28.3
	$C=I$	1.7	0	98.3
P3	$C=H$	95	0	5
	$C=A$	20	60	20
	$C=I$	0	0	100
		Recall	Precision	
	Healthy	95	90.5	
	Asthma	60	100	
	ILD	100	76.9	
	η / ζ	87.5 / 82.9		

the eight classifiers. The confusion matrices and the success scores calculated for P3 corresponding to this selection (GMM1ST with component count seven) are given in Table 4.5. Although the majority of the bronchiectasis instances are labeled correctly, the recall rate is still low, moreover, bronchiectasis is observed to be confused equally with the other two classes. Healthy and ILD recall and precision rates are rather high, which validates that the classifier is competent in general, except in the discrimination of bronchiectasis.

GMM1ST with seven components is also trained for an alternative data set, where bronchiectasis is replaced with another obstructive disease, namely, asthma (10 subjects as with bronchiectasis). The confusion matrices together with the success scores are given in Table 4.6. The precision rate of asthma turns out to be 100% due to the fact that none of the other two classes are confused with asthma. However, the

general picture is not different from the case with bronchiectasis, and asthma is equally confused with the other two classes, with the same recall rate as bronchiectasis.

In this section, the hierarchical form of the GMM classifier is determined to be the best among the eight schemes. General observation is that the two obstructive diseases are confused with the healthy and the restrictive classes, which causes a low recall rate for the obstructive disease, and lower precision rates for the other classes than they could have had without this confusion.

4.3.3. Special Case: Asthma/COPD

Pulmonary sounds data of 50 subjects are used in this section, where 30 of them are diagnosed with asthma and 20 of them with COPD (DS3, see Chapter 2). As addressed above, asthma and COPD are often confused in clinical diagnosis due to the overlap in their symptoms. The primary aim of this section is to explore the diagnostic ability of the GMM classifier in this special case. Besides, this section is introduced as an intermediate step, from the one on the quest of a useful classifier and a useful algorithm design, to the next on the diagnostic classification with the widest data set at hand. Therefore, while the diagnostic ability of the GMM classifier is being explored, additional perspectives as variants of P3 are considered as well, taking the unequal pertinences of flow-phases into consideration. Hence, an improvement of the P3 success scores is the secondary aim of the section.

To obtain subject-level classification results, P3-SOP has been the conventional methodology throughout the experiments conducted so far. In this section, its logarithmic version, P3-SOLP, and two SVM-based methods, namely, P3-SVM and P3-SVML, are considered for the improvement of P3 scores (see Section 4.2 for the definitions). For the GMM classifier, 10-fold cross validation scheme is adopted such that the validation set is arranged to have three of asthma and two of COPD subjects at each fold. Moreover, the same 10-fold cross validation scheme is adopted for P3-SVM and P3-SVML as well, for them to be in accordance with the GMM step. GMM is used with component counts from one to ten, and the classification is performed for each of

the six flow-phases separately, as before.

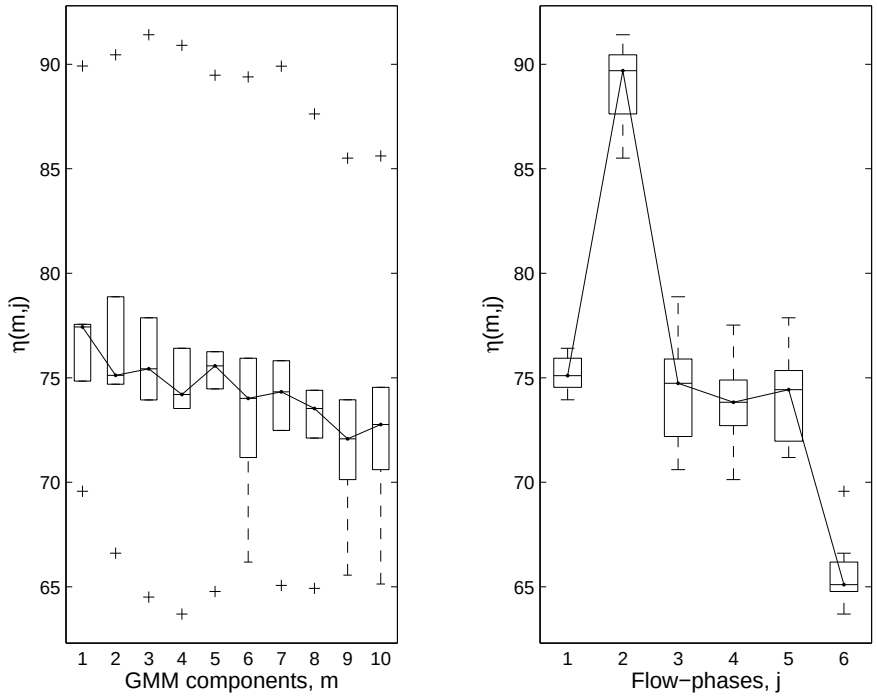
To evaluate the classification results from the first and second perspectives (P1 and P2), success scores $\eta(m, j)$ are calculated for all m and j. The box-plots of $\eta(m, j)$ for P1 and P2 are given in Figures 4.6a and 4.6b respectively, versus m (left) and j (right). As observed in both of the sub-figures, the success scores do not depict significant differences across GMM component counts, however they do across flow-phases, such that the second flow-phase (mid-inspiration) is significantly the most successful of all and the sixth flow-phase (late-expiration) is significantly the least successful.

For P3-SVM and P3-SVML, the values for the penalty factor are optimized over the range $F = 10^{-10}, 10^{-9}, ..., 10^{10}$. Therefore, the corresponding success score ζ is a function of m and F for the two variants of P3. The smallest F yielding the maximum success score is selected to be the optimal penalty factor per m, that is

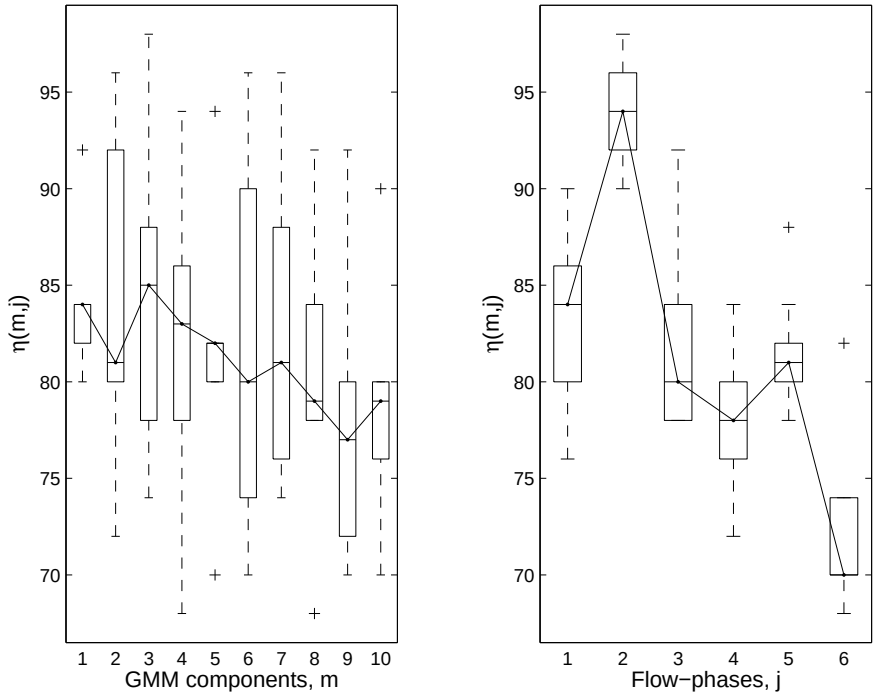
$$F_m = \min(\underset{F}{\operatorname{argmax}}\{\zeta(m, F)\}).$$

The classifier is trained on the whole data set once more using F_m to learn the final weights corresponding to the particular GMM component count selection, i.e., the vector $\mathbf{w}_m$. The final decisions are made using this vector, and the classifier performances are reported accordingly. Note that this vector turns out to be approximately the average of the vectors calculated per fold, achieving equal or higher successes in the final decisions than the average vector does (observation not reported explicitly).

The calculated ζ scores versus m are plotted in Figure 4.7 for P1, P2, and the variants of P3. Since the most prominent flow-phase is observed to be the second one (mid-inspiration) in Figure 4.6, the scores based only on the second flow-phase is also shown for comparison (P3-SP2). In Figure 4.7, a downward trend is observed both for P1 and P2 (also for P3-SOP and P3-SOLP) as the GMM component count increases, which is in accordance with Figure 4.6. Moreover, the performances are observed to improve as proceeding from P1 to P2, and from P2 to P3, as observed in the two



(a) η from P1 (the segment-level evaluation perspective).



(b) η from P2 (the flow-phase-level evaluation perspective).

Figure 4.6. Box-plots of η scores, versus GMM component count (on the left), and versus flow-phase index (on the right)

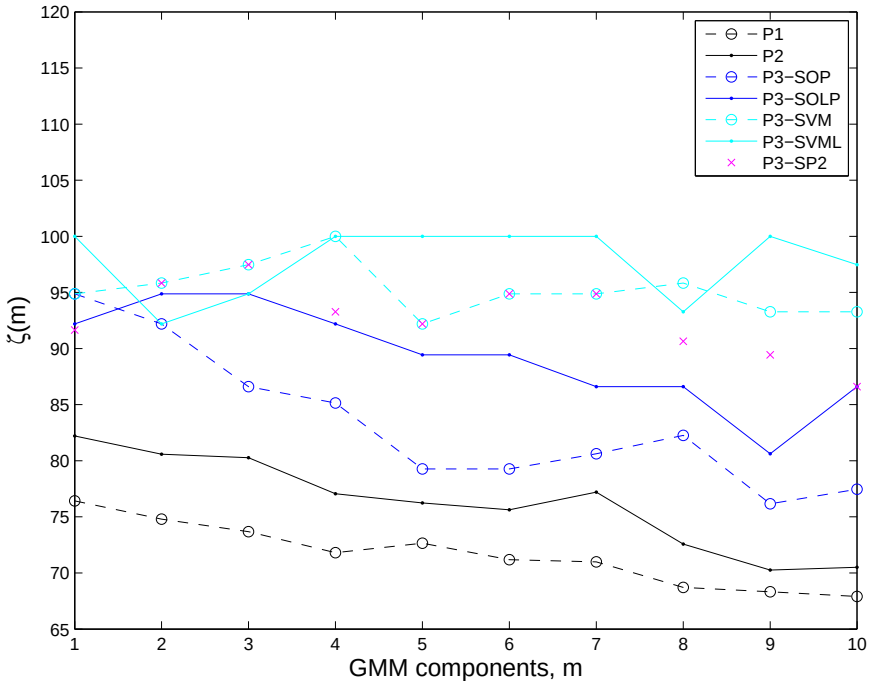


Figure 4.7. The ζ scores versus GMM component count from all evaluation perspectives.

previous sections as well. The explicit confusion matrices for GMM component counts one to five are also given in Table 4.7 for reference.

P3-SOLP outperforms P3-SOP for all m except $m = 1$, and P3-SVM in turn outperforms P3-SOLP for all m, reaching at 100% for $m = 4$. P3-SVML on the other hand is comparable to P3-SVM, albeit inconsistent in terms of superiority and inferiority. The former falls below the latter for $m = 2, 3, 8$, without an intuitive explanation. On the other hand, it reaches the 100% success level at $m = 1, 4, 5, 6, 7, 9$, hence for the majority of GMM component count selections.

Making the subject-level decisions by depending only on the second flow-phase (P3-SP2) turns out to be a better approach than P3-SOP and P3-SOLP, except for $m = 1$. However, the weighted sums of posterior (P3-SVM) and log-posterior probabilities (P3-SVML) attain even higher (or equal, if not higher, if P3-SVM is considered) scores than P3-SP2 do. This result emphasizes the importance of taking the unequal pertinences of flow-phases into account. P3-SVM curve is either above, or coincides with, the P3-SP2 points. For those component count selections where the former is

Table 4.7. Confusion matrices for GMM component counts one to five, for all evaluation perspectives. A hat is used to denote the predicted class label. A: Asthma, C: COPD.

		P1		P2		P3-SOP		P3-SOLP		P3-SVM		P3-SVML	
		$\hat{A}$	$\hat{C}$	$\hat{A}$	$\hat{C}$	$\hat{A}$	$\hat{C}$	$\hat{A}$	$\hat{C}$	$\hat{A}$	$\hat{C}$	$\hat{A}$	$\hat{C}$
1	A	82.9	17.1	91.1	8.9	100	0	100	0	100	0	100	0
	C	29.6	70.4	25.8	74.2	10	90	15	85	10	90	0	100
2	A	83.4	16.6	92.8	7.2	100	0	100	0	96.7	3.3	100	0
	C	32.9	67.1	30	70	15	85	10	90	5	95	15	85
3	A	84.9	15.1	96.7	3.3	100	0	100	0	100	0	100	0
	C	36	64	33.3	66.7	25	75	10	90	5	95	10	90
4	A	85.5	14.5	95	5	96.7	3.3	100	0	100	0	100	0
	C	39.7	60.3	37.5	62.5	25	75	15	85	0	100	0	100
5	A	85.3	14.7	95.6	4.4	96.7	3.3	100	0	100	0	100	0
	C	38.1	61.9	39.2	60.8	35	65	20	80	15	85	0	100

Table 4.8. The weight vectors $\mathbf{w}_m$ learned by P3-SVM, and the corresponding ζ scores, for GMM component counts, m.

m	w_{m1}	w_{m2}	w_{m3}	w_{m4}	w_{m5}	w_{m6}	ζ
1	0.33	0.45	0.33	0.26	0.33	0.30	94.9
2	0.00	1.32	0.68	0.00	0.00	0.00	95.8
3	0.32	1.58	0.21	0.11	0.31	-0.11	97.5
4	1.99	1.99	0.00	1.99	0.00	-1.99	100
5	0.40	0.90	0.40	0.30	0.40	0.00	92.2
6	0.53	1.47	0.32	-0.16	0.16	0.00	94.9
7	0.58	0.98	0.37	0.12	0.32	0.07	94.9
8	1.99	1.88	0.35	-0.12	0.00	-0.12	95.8
9	1.99	2.23	0.00	0.00	0.00	0.00	93.3
10	1.99	1.99	0.00	0.00	0.00	0.00	93.3

above ($m = 1, 4, 8, 9, 10$), an examination of Table 4.8 reveals the other flow-phases that were taken into account. The case with $m = 1$ is in accordance with Figure 4.6, with the weights in the table following the same pattern as that of the median success scores in the figure. The common observation with $m = 8, 9, 10$ is that the first and second flow-phases are the only pertinent ones. For $m = 4$, where P3-SVM score reaches the 100% level making its single maximum, the weights are equal in magnitude for the first, second, fourth and sixth flow-phases, with a minus sign for the sixth, and are zero for the remaining (the third and fifth) flow-phases. Except for the case with $m = 1$, these observations seem contradictory with the box-plots of Figure 4.6, however reveal that the first flow-phase has more pertinence, while the third and fifth have less, to reach the subject-level decisions than they imply in P1 and P2. Moreover, the sixth flow-phase is almost inversely effective.

The performances reached by P3-SOP and P3-SOLP may be interpreted to be more consistent and reliable than P3-SVM and P3-SVML since they are basically the summations of the posterior and log-posterior probabilities calculated once the GMM is fitted, and they promise good subject-level discrimination to the extent that the GMM

is successful to model the components of hopefully distinct mixtures. The performances reached by P3-SVM and P3-SVML, on the other hand, depend both on the outcomes of the first classifier (GMM) and on the design of the second classifier (SVM), therefore the weights corresponding to high diagnostic ability for this experiment are not the final weights that should be used with any GMM classification of asthma versus COPD. Rather, the experiments of this section are conducted to see whether it is *possible* to outperform the simple combination (sum with equal weights) by a proper combination (weighted sum with proper weights), keeping in mind that the flow-phases *may* not be equivalent in pertinence. The considerably high success scores obtained with these two SVM-based decision combination methodologies verify that it is. Nevertheless, the final weights to be employed in clinical setup should be learned using a larger data set and various schemes for cross validation and testing.

Even for a mediocre success level in Figure 4.7, e.g., the common level of P3-SOP at $m = 1$ and P3-SOLP at $m = 2$, the total correct classification rate is 96%, with the correct classification rates of 100% and 90% for asthma and COPD respectively, which implies a quite successful classification. The results of this experiment are very promising in terms of demonstrating that the differential diagnosis of the two obstructive diseases in clinical practice can benefit from the incorporation of computerized techniques in pulmonary sounds analysis.

4.3.4. General Case: Diagnostic Classification

The data set of this section is the largest data set used in the experiments, consisting of 150 subjects, where 25 of them are healthy, 60 of them are diagnosed with asthma, 35 of them with COPD, 15 of them with bronchiectasis, and the remaining 15 with ILD. Depending on the results of the experiments conducted so far, GMM classifier with hierarchical setup in K-fold cross validation scheme, and P3-SOP, P3-SOLP and P3-SVM as decision combination methods, are adopted. The aim of this section is to experiment with a multi-class data set, with the eventual aim to devise a methodology for diagnostic classification of pulmonary sounds. The quest for the best among several possible division schemes for the hierarchical framework is also in the

Table 4.9. The best common and distinct GMM component counts for the two classes and the corresponding ζ scores for H/P classification. Best ζ is given in bold. CC: Component count, H: Healthy, P: Pathological.

Perspective	ζ (Best common CC)	ζ (Best distinct CCs)
P1	70.7 (4)	73.1 (H:6, P:2)
P2	75.3 (4)	80.2 (H:5, P:2)
P3-SOP	79.6 (1)	89.6 (H:6, P:2)
P3-SOLP	81.0 (2)	83.9 (H:9, P:2)
P3-SVM	86.5 (3)	**92.8 (H:5, P:2)**

scope of this section. Besides, a new perspective is considered for the GMM classifier, namely, assuming a different GMM component count per class and finding the optimal combination for the classifier.

The intuitive division scheme is to start the hierarchical structure with healthy versus pathological (H/P hereafter) classification, and to divide the pathological group into the obstructive and restrictive groups (O/R). Note that the obstructive group consists of asthma, bronchiectasis and COPD, whereas the restrictive group consists only of ILD. In Tables 4.9 and 4.10, ζ scores are presented for the binary classifiers for H/P and O/R classification. In the tables, the left column is for the best scores that are yielded based on the assumption that the GMM component counts are the same

Table 4.10. The best common and distinct GMM component counts for the two classes and the corresponding ζ scores for O/R classification. Best ζ is given in bold. CC: Component count, O: Obstructive, R: Restrictive.

Perspective	ζ (Best common CC)	ζ (Best distinct CCs)
P1	56.4 (2)	58.7 (O:1, R:4)
P2	63.1 (2)	66.1 (O:1, R:4)
P3-SOP	71.3 (2))	**81.8 (O:1, R:3)**
P3-SOLP	71.7 (2)	80.8 (O:1, R:2)
P3-SVM	76.0 (1)	76.0 (O:1, R:1)

Table 4.11. The best common and distinct GMM component counts for the two classes and the corresponding ζ scores for U/L classification. Best ζ is given in bold. CC: Component count, U: Upper airways, L: Lower airways.

Perspective	ζ (Best common CC)	ζ (Best distinct CCs)
P1	53.9 (2)	55.9 (U:1, L:3)
P2	58.2 (2)	60.4 (U:2, L:3)
P3-SOP	62.9 (1)	72.4 (U:2, L:3)
P3-SOLP	64.9 (1)	**72.9 (U:1, L:2)**
P3-SVM	66.5 (1)	66.5 (U:1, L:1)

for the both classes. The right column, on the other hand, is for the new perspective adopted in this section, i.e., different component counts for the mixtures of the two classes. In parentheses, the optimal GMM component counts are given. On the right column, the zeta scores are higher than those on the left, but not with large differences. Nevertheless, a closer observation of the confusion matrices of all reveals that, the associated confusion matrices (observed but not reported explicitly) are more balanced, i.e., the correct classification rates of the two classes are closer to each other, when the new approach is adopted.

As an alternative to the O/R division of the pathological group, another division is considered as to put asthma and COPD in one group, and bronchiectasis and ILD into the other. The rationale behind this division is that the former pair is associated with the upper (U) airways of the lungs, whereas the latter pair is associated with the lower (L) airways. This division is included for exploratory purposes. If the success scores turned out to be considerably higher, a new insight would be gained about how to approach diagnostic (multi-class) classification of pulmonary sounds in clinical practice.

The scores associated with the U/L division are given In Table 4.11. The confusion matrices for O/R and U/L division schemes are given in Tables 4.12 and 4.13, respectively. Note that each binary matrix is the matrix that corresponds to the highest

Table 4.12. The binary confusion matrices for H/P and O/R classifications and the resulting H/O/R confusion matrix. H: Healthy, O: Obstructive, R: Restrictive.

	$\hat{H}$	$\hat{P}$
H	92	8
P	6.4	93.6

	$\hat{O}$	$\hat{R}$
O	83.6	16.4
R	20	80

	$\hat{H}$	$\hat{O}$	$\hat{R}$
H	92	8	0
O	7.3	77.3	15.4
R	0	20	80
ζ=82.9 (η=80)			

(bold) score in the relevant ζ score table. The 3×3 confusion matrices on the right are the resulting three-class performances based on the successive decisions of the binary classifiers. Note that the performance of the U/L division scheme is poorer than that of the O/R, due to the low recall rate of the L class (bronchiectasis together with ILD). Therefore, the O/R division scheme is adopted for the rest of the experiment.

To further divide the obstructive group, various schemes are considered:

(i) Asthma versus COPD together with bronchiectasis (A/CB),
(ii) Asthma together with COPD versus bronchiectasis (AC/B),
(iii) Asthma together with bronchiectasis versus COPD (AB/C),
(iv) Asthma, COPD and bronchiectasis as separate classes, to be classified *all at once* (A/C/B).

Calculating the ζ scores as with the H/P, O/R and U/L cases, and considering

Table 4.13. The binary confusion matrices for H/P and U/L classifications and the resulting H/U/L confusion matrix. H: Healthy, L: Lower airways, U: Upper airways.

	$\hat{H}$	$\hat{P}$
H	92	8
P	6.4	93.6

	$\hat{U}$	$\hat{L}$
U	87.4	12.6
L	40	60

	$\hat{H}$	$\hat{U}$	$\hat{L}$
H	92	4	4
U	5.3	72.6	22.1
L	10	23.3	66.7
ζ=76.4 (η=74.6)			

Table 4.14. The binary confusion matrices for the first three schemes of obstructive division

	$\hat{A}$	$\hat{CB}$
A	86.7	13.3
CB	44	56
ζ=69.7		

	$\hat{AC}$	$\hat{B}$
AC	74.7	25.3
B	53.3	46.7
ζ=59.8		

	$\hat{AB}$	$\hat{C}$
AB	72	28
C	31.4	68.6
ζ=70.3		

the maximum score to select the best component count combination, the confusion matrices for the binary schemes listed above (the first three) are presented in Table 4.14. Although none of the three is considerably superior to the other two, the third one (AB/C) has the highest ζ score. The further step for asthma versus bronchiectasis (A/B) classification is also performed, and the resulting binary and 5×5 confusion matrices together with the corresponding recall and precision rates are given in Table 4.15. A comparison with the results of A/C/B scheme (see Table 4.16) reveals that performing the classification with binary schemes at each step results in the highest ζ score, although with a slight difference.

In the H/O/R scheme (Table 4.12), the healthy correct classification rate is rather high, and the obstructive and restrictive correct classification rates are low but not unacceptably, yielding a total correct classification rate of 80%. A comparison of Table 4.12 with the results of Section 4.3.2 reveals that the new confusion matrix is more balanced, despite having comparable ζ and lower η scores. The healthy class is the least confused, whereas the obstructive and restrictive classes are confused with each other. Table 4.16 reveals the details for the confusion pattern of the obstructive diseases (refer to the 3×3 matrix on the left). Accordingly, asthma and COPD are confused with each other even though they were successfully separated in the binary classification scheme of Section 4.3.3. Moreover, bronchiectasis is confused mainly with asthma, which explains why the AB/C division turns out to be the most successful among the three binary division schemes.

The observation of Tables 4.15 and 4.16 reveals that the highest recall and precision rates belong to the healthy class, and the lowest precision rates to the ILD (the

Table 4.15. The binary confusion matrices for AB/C and A/B classification, and the resulting 5×5 confusion matrix, together with the corresponding precision and recall rates.

	$\hat{AB}$	$\hat{C}$
AB	72	28
C	31.4	68.6

	$\hat{A}$	$\hat{B}$
A	86.7	13.3
B	46.7	53.3

	$\hat{H}$	$\hat{A}$	$\hat{C}$	$\hat{B}$	$\hat{I}$
H	**92**	4	0	4	0
A	8.3	**48.3**	21.7	6.7	15
C	0	20	**54.3**	5.7	20
B	20	13.3	26.7	**33.3**	6.7
I	0	13.3	0	6.7	**80**
ζ=57.8					

	Recall	Precision
H	92	74.2
A	48.3	70.7
C	54.3	52.8
B	33.3	38.5
I	80	41.4

Table 4.16. The 3×3 confusion matrix for the third scheme of obstructive division, and the resulting 5×5 confusion matrix, together with the corresponding precision and recall rates.

	$\hat{A}$	$\hat{C}$	$\hat{B}$
A	71.7	21.7	6.6
C	42.8	54.3	2.9
B	40	20	40

	$\hat{H}$	$\hat{A}$	$\hat{C}$	$\hat{B}$	$\hat{I}$
H	**92**	4	4	0	0
A	8.3	**55**	18.3	3.3	15
C	0	34.3	**42.8**	2.8	20
B	20	26.7	13.3	**33.3**	6.7
I	0	13.3	0	6.7	**80**
ζ=56.5					

	Recall	Precision
H	92	74.2
A	55	63.5
C	42.8	51.7
B	33.3	55.6
I	80	41.4

lowest in Table 4.15 and the second lowest in Table 4.16), although ILD has the second highest recall rate among the five classes. The implication is that a high percentage of ILD subjects are correctly diagnosed, however, the reliability of the diagnosis is poor. The result that the highest precision belongs to the healthy class is appropriate, since the healthy diagnosis is the one that requires the highest reliability. Referring to Tables 4.15 and 4.16, the probability that the subject is actually healthy is 74.2%, if s/he is diagnosed as healthy by the classifier. The 5×5 confusion matrices of Tables 4.15 and 4.16 are comparable, with the following prominent differences. The correct classification rate of asthma is lower than that of COPD in the former, and vice versa in the latter. Besides, bronchiectasis is mainly confused with COPD in the former, whereas with asthma in the latter.

The poor scores in the 5×5 confusion matrices are due to the confusion of the obstructive diseases with each other. A hierarchical framework is suggested for the diagnostic classification of pulmonary sounds, since it achieves higher success scores than the *all at once* point of view does. O/R division is suggested as it outperforms the alternative U/L division. The poorest scores are yielded by the further division of the obstructive class, which bears the necessity of incorporating new features as to capture the specific distinctive characteristics.

4.4. Summary and Discussion

Section 4.3.1 concludes that the GMM classifier is the best among all those experimented with, however, the performance of the SVM classifier can not be neglected as well. Rather high success scores obtained with the SVM classifier even with the average feature vectors (see SVM6-AO/AT in Section 4.3.2) are very promising, and the other variants of SVM with probabilistic outputs are open for exploration. Combining the flow-phases at the feature level is suggested for SVM classification of VAR parameters, if the observation of the relative pertinences of flow-phases is not an aim. Adaptation of the hierarchical framework is encouraged for both GMM and SVM, though the performance improvement is more remarkable with the GMM.

GMM on VAR parameters is shown to separate asthma and COPD successfully. Although summing the posterior probabilities of flow-phases (P3-SOP and P3-SOLP) is an already successful decision combination method to make the subject-level decisions, a weighted summation where the weights are learned by an SVM classifier (P3-SVM and P3-SVML) further improves the performances. In asthma and COPD classification, the prominent flow-phase seems to be the second one (mid-inspiration) in the segment and flow-phase performances, while the first and the fourth (early-inspiration and expiration) appear to be important as well in the combined decisions. The sixth flow-phase (late-expiration) seems to be the least prevalent. The first three sections also reveal that the performances always improve as proceeding from the segment level towards the subject level, which verifies the importance of combining individual decisions to decide for the group.

The diagnostic classifier designed in Section 4.3.4 can not reach acceptable levels of performance for a clinical setup. Nevertheless, it builds a framework that is to be adopted for further explorations. First, it verifies the intuitive division of obstructive versus restrictive for pulmonary sounds classification studies in general. Second, it verifies the expectation that considering different component counts for classes yields an improved GMM classifier performance.

The optimal component counts combination for the GMM classifier in Section 4.3.4 is five and two for the healthy and pathological classes. In the obstructive versus restrictive classification, they are one and three, respectively. Actually, the pathological class consists of four distinct pathologies, hence is expected to be formed of several clusters, while the healthy class is expected to fit into one cluster. Similarly, the obstructive class consists of three distinct pathologies, whereas the restrictive class consists of a single type of pathology. Therefore, the optimal component counts (smaller for those classes with more variety) may seem intriguing. However, one should keep in mind that the component counts that the best classification performances are obtained with may be different than those that yield the best cluster model for the individual class. The generalization ability of the model may be the factor that increases the separability in classification, whereas, the maximum likelihoods for the individual data

groups would be the measure of the goodness of fit in clustering.

Although considering the results of Section 4.3.4 for physiological reasoning is not proper with such low success scores, the following real life experience is worth mentioning. During the data acquisition sessions held in the hospital environment, cases where asthma and bronchiectasis coexist have been encountered more often than those where COPD and bronchiectasis do. The subjects with a report of a mixed diagnosis have been omitted while building the data sets for the experiments of this study. While seven subjects with COPD-bronchiectasis have been omitted from the data set, 21 subjects with asthma-bronchiectasis have been omitted, which implies three times higher field-observation of asthma coexistence. Therefore, the result that the best binary division scheme of the obstructive group turns out to be AB/C is in accordance with the subjective expectation, taking into consideration that those instances included in the study might not be singular pathologies either. If the current complaints of the subject are prominently due to bronchiectasis (e.g., chest pain) rather than asthma (e.g., short of breath and wheezing), the diagnosis of asthma, or its reporting to the data acquisition attendant, might be neglected. The same point also applies to the case with COPD. In Table 4.15, bronchiectasis is confused with COPD more often than vice versa, which leads to a suspect that even if the reported diagnosis was 'bronchiectasis', COPD might also have been accompanying, although not reported. Since the obstructive diseases frequently coexist in other combinations as well, the further division of the obstructive class for diagnostic pulmonary sounds classification studies should be handled cautiously. Including singular diagnoses in the study and paying extra attention to the absence of any accompanying pathologies is suggested besides the quest for better mathematical models to extract features and for better algorithms or schemes of classifiers. Defining separate classes for combined pathologies might also be considered, perhaps at the further steps of improvement of the diagnostic classifier.

In Tables 4.17 and 4.18, the numerical results of the four sections are summarized together with those of the relevant studies in the literature. The total correct classification rates and the methodologies in the tables are given only for the feature-classifier combinations that yield the best scores among all those experimented in the reference

work. Note that for [36] and [37] (Table 4.17) the given sample sizes for the healthy and pathological classes are for respiratory cycles that are employed, not for subjects. Hence, the data size is very small in this study, in [38] as well, which are the two points to be taken into consideration while comparing the results. In [40], k-NN with Euclidean distance measure is found to be successful in healthy versus pathological classification. The results of Section 4.3.1 for the k-NN classifier, on the other hand, are not even comparable to this score. Note that the k-NN classifier is known to work best where the redundancy is the least. Therefore, the poor performance may be attributed to the high feature space dimensionality and the large data size. In [41], the correct classification rates of the healthy and ILD classes are 0% and 40%, respectively for the VAR model, while they are 75% and 93%, respectively for the the univariate AR (rates that are used to calculate the numeric result in Table 4.17). The optimal VAR model order is determined by Akaike's criterion, the microphone array is composed of 25 microphones, the data segmentation is done as to keep the number of segments constant (rather than the lengths), and only the inspiration phases are included in the classification (see Section 3.5 for more details). Moreover, the final decisions are made on the subject level via a thresholding. The low success rates may be attributed to the listed preferences, rather than the features or the classifier adopted.

In [39] (Table 4.18), the cepstral coefficients are found to be slightly more successful than the univariate AR parameters. By *two-stage* classification, the authors refer to a scheme of feature level classification followed by decision combination. Accordingly, the first stage represents the direct results of the feature vector classifier, whereas the second stage represents the expert decisions made by a second classifier. Hence, the two-stage scheme is not equivalent to the hierarchical scheme offered in this study, but to P3-SOP, P3-SVM, and the like. As in this study, the performances improve upon decision combination. The best results are obtained for the non-parametric Parzen window classifier in the second stage. The explicit class correct classification rates are not given in [39], therefore a comparison of the results is only possible via the total correct classification rate. Both tables depict that the results of the experiments conducted here are comparable to those of the relevant studies in the literature, though a direct comparison of the feature extraction and classification methods would only be

Table 4.17. Summary of the results for the binary case of healthy versus pathological classification. H: Healthy, P: Pathological, TCC: Total correct classification rate (η), ANN: Artificial neural network classifier, DWT: Discrete wavelet transform.

Reference	Material	TCC(%)	Methodology
[36], [37]	12 H* / 13 P*	93.75	AR, k-NN (Itakura)
[40]	20 H / 20 P	88	AR, k-NN (Itakura, Euclidean)
[41]	8 H / 19 ILD	87.7**	AR, ANN
[38]	10 H / 10 A	98	DWT, ANN
Section 4.3.1	20 H / 20 P	92.5	VAR, GMM
Section 4.3.2	20 H / 20 P	90	VAR, GMM
Section 4.3.4	25 H / 125 P	93.3	VAR, GMM

* Given figures are counts of respiratory cycles, not of subjects

** Not given explicitly in [41], but calculated here from class correct classification rates and subject counts

Table 4.18. Summary of the results for the classification into healthy, obstructive and restrictive groups. H: Healthy, O: Obstructive, R: Restrictive, B: Bronchictasis, A: Asthma.

Reference	Material	TCC(%)	Methodology
[39]	20 H / 18 COPD / 19 R	82.1	Cepstral, ANN, Parzen
Section 4.3.2	20 H / 10 B / 10 ILD	85	VAR, GMM P3-SOP followed by GMM P3-SOP
Section 4.3.2	20 H / 10 A / 10 ILD	87.5	VAR, GMM P3-SOP followed by GMM P3-SOP
Section 4.3.4	25 H / 110 O / 15 ILD	80	VAR, GMM P3-SVM followed by GMM P3-SOP

possible with the use of the same data sets.

In selection of the data sets DS0, DS1, DS2 and DS3, the subjects with the least episodes of the adventitious sound components are considered for the VAR model to adhere to its theory. To enlarge the data set to form DS4, the data having adventitious sound components or some acquisition noise are also included compulsorily. As addressed in the introduction of Chapter 3, the existence of adventitious sound components is prone to disrupt the Gaussianity assumption of the VAR model, moreover, the existence of crackles in particular is prone to disrupt the stationarity assumption even within the short segments. This is the most probable reason for the poor discrimination between asthma and COPD in the final experimental setup, whereas the discrimination was rather successful in the special section with DS3. The performances of the classifiers may be improved by the inclusion of new features to complement the VAR parameters, so as to enhance the distinctive information offered by the adventitious sound components. Nevertheless, the performance of the H/O/R classifier of Section 4.3.4 may be considered rather successful regarding the "noise" that the data set accommodates (i.e., adventitious sound components, undeterred noise, deficient diagnostic labels reported in the hospital environment).

5. CONCLUSIONS AND FUTURE PERSPECTIVES

In this thesis study, the problem of finding an optimal model for multi-channel pulmonary sounds and applying the model parameters to classify these sounds for diagnostic purposes is addressed. To this end, new goodness of fit criteria to find the optimal VAR model have been proposed specifically for pulmonary sounds data for diagnostic purposes, and the estimated parameters of the optimal model selected via these criteria have been employed in classification with a final aim to build a diagnostic system for clinical setups. Various classifier algorithms and schemes have been designed for different data sets to evaluate and compare the classification successes and to improve them accordingly for the next step in design.

The VAR model is widely used to predict the future behavior of time series in meteorology and economics, where the best model order and sample size are determined via the conventional goodness of fit criteria. As a multi-variate model to represent interrelated measurements, the theory applies to the multi-channel pulmonary sound signals as well, where this time the focus is not on predicting the future behavior but on learning the distinctive pulmonary system characteristics. Therefore, the measure of the goodness of fit is not the variance of the prediction error but the discriminative ability between distinct types of conditions, as shown in this study. Accordingly, the optimal VAR model to represent pulmonary sounds data for diagnostic purposes has been shown to be the 250-point VAR(2) model. The proposed measures can be employed with the same intention for different channel counts and sampling rates of pulmonary sounds measurements, and even for other types of signals where the aim is to differentiate between varieties of the underlying system. Through adaptation, the measures can also be extended to other mathematical models where the selection of the optimal model requires an exhaustive search over the ranges of free model arguments. The adaptation is straightforward, since the underlying structure is inspired from the well known theory of LDA.

This study verifies that the parameters of the optimal VAR model for pulmonary

sounds data are useful to discriminate between two distinct pulmonary conditions, namely, healthy and pathological, with a total correct classification rate above 90%. SVM and GMM classifier algorithms are shown to be competitively successful, the latter outperforming the former, however k-NN as it is used in this study is poor. To increase its performance, dimensionality reduction can be performed on the feature vectors (the VAR parameter matrices in vector forms), and the averages taken in individual flow-phases can be employed to reduce the redundancy. SVM with the radial basis function is avoidable due to the long computational time, since it provides minor improvements over the linear SVM in return. In a future study, the linear SVM is suggested to be used on the average feature vectors, moreover, on the feature space formed by combining the six-flow phases if they are not to be examined separately. The limitation of the SVM classifier as it is used in this study is that it produces discrete class labels as outputs. That leads to insignificant improvements in the success scores as proceeding from the segment-level evaluation towards the subject-level, as compared to the GMM classifier, hence leading to deficiency in the overall classifier success. Using the SVM classifier with probabilistic outputs for class memberships is encouraged for future studies.

In a diagnostic setup with multiple pathological conditions, a hierarchical framework has been proposed in this study, as to classify the data first into the healthy and pathological groups, then the pathological group further into sub-groups. For the further classification of the pathological group, obstructive versus restrictive discrimination has been suggested. In the case of the three-class scheme with the healthy, obstructive and restrictive classes, the obstructive class is observed to be confused with the others more often. Moreover, the distinct diseases in the obstructive class are also confused with each other in the diagnostic scheme of the final part. Nevertheless, the total correct classification rate for the hierarchical classification of the healthy, obstructive and restrictive classes reaches 80%. The acoustic-based differential diagnosis of asthma and COPD is observed to be successful with a total correct classification rate above 95%, which promises a powerful tool for clinical practice. The reason that the binary classification of asthma and COPD was not as successful in the hierarchical diagnostic scheme of the final part as in the special section on asthma and COPD is most

probably due to the fact that the new data included consisted of adventitious sound components. The difference in the successes emphasizes the necessity to incorporate new features to complement the VAR model parameters, which are designed or selected specifically to represent the distinctive characteristics of adventitious pulmonary sound components.

In the further development of the diagnostic classifier, GMM classifier algorithm is strongly suggested to be incorporated in the study, although different classifier algorithms are to be explored as well. As discussed above, it is also suggested to experiment with the linear SVM to explore new methods to increase its discriminative ability. The further discrimination between the obstructive diseases should be handled cautiously due to the overlaps (both in terms of symptoms and the classification results). The specific acoustic characteristics of distinct obstructive diseases are suggested to be taken into account in further studies. Accordingly, new features to enhance those characteristics, in particular, to represent adventitious sound components, may be adopted. In a future study, the discrimination including various disease stages and comparisons with conventional clinical test results would increase the reliability of the diagnostic system. The preparation of the data set needs caution as to involve singular diagnoses if the diagnostic classifier is aimed to assess singular pathologies. To cope with the co-existence of diseases, new classes may be defined as combinations of multiple pulmonary conditions. An ideal diagnostic system would eventually accept as inputs the information such as the age, weight and height of the patient, the current complaints, the clinical and pathological history, spirometric values and blood test results.

REFERENCES

1. Pasterkamp H., S. S. Kraman and G. R. Wodicka, "Advances Beyond the Stethoscope", *American Journal of Respiratory and Critical Care Medicine*, Vol. 156, pp. 974-987, 1997.

2. Van Den Berg J. W., "An Electrical Analogue of the Trachea, Lungs and Tissues", *Acta Physiologica et Pharmacologica Neerlandica*, Vol. 9, pp. 361-385, 1960.

3. Ishizaka K., M. Matsudaira and T. Kaneko, "Input Acoustic-Impedance Measurement of the Subglottal System", *Journal of the Acoustical Society of America*, Vol. 60, No. 1, pp. 190-197, 1976.

4. Jackson A.C., C.A. Giudanella and H.L. Dorkin, "Density Dependence of Respiratory System Impedances Between 5 and 320 Hz in Humans", *Journal of Applied Physiology*, Vol. 67, No. 6, pp. 2323-2330, 1989.

5. Wodicka G. R., K. N. Stevens, H. L. Golub, E. G. Cravalho and D. C. Shannon, "A Model of Acoustic Transmission in the Respiratory System", *IEEE Transactions on Biomedical Engineering*, Vol. 36, No. 9, pp. 925-934, 1989.

6. Goncharoff V., J. E. Jacobs and D. W. Cugell, "Wideband Acoustic Transmission of Human Lungs", *Medical and Biological Engineering and Computing*, Vol. 27, No. 5, pp. 513-519, 1989.

7. Kraman S. S. and A. B. Bohadana, "Transmission to the Chest of Sound Introduced at the Mouth", *Journal of Applied Physiology*, Vol. 66, No. 1, pp. 278-281, 1989.

8. Wodicka G. R. and D. C. Shannon, "Transfer Function of Sound Transmission in Subglottal Human Respiratory System at Low Frequencies", *Journal of Applied Physiology*, Vol. 69, No. 6, pp. 2126-213, 1990.

9. Wodicka G. R., K. N. Stevens, H. L. Golub and D. C. Shannon, "Spectral Characteristics of Sound Transmission in the Human Respiratory System", *IEEE Transactions on Biomedical Engineering*, Vol. 37, No. 12, pp. 1130-1135, 1990.

10. Wodicka G. R., P. D. DeFrain and S. S. Kraman, "Bilateral Asymmetry of Respiratory Acoustic Transmission", *Medical and Biological Engineering and Computing*, Vol. 32, No. 5, pp. 489-494, 1994.

11. Pasterkamp H., S. Patel and G. R. Wodicka, "Asymmetry of Respiratory Sounds and Thoracic Transmission", *Medical and Biological Engineering and Computing*, Vol. 35, No. 2, pp. 103-106, 1997.

12. Royston T. J., X. Zhang, H. A. Mansy and R. H. Sandler, "Modeling Sound Transmission Through the Pulmonary System and Chest with Application to Diagnosis of a Collapsed Lung", *Journal of the Acoustical Society of America*, Vol. 111, No. 4, pp. 1931-1946, 2002.

13. Bergstresser T., D. Ofengeim, A. Vyshedskiy, J.Shane and R. Murphy, "Sound Transmission in the Lung as a Function of Lung Volume", *Journal of Applied Physiology*, Vol. 93, pp. 667-674, 2002.

14. Murphy R. L., E. A. Del Bono and F. Davidson, "Validation of an Automatic Crackle (Rale) Counter", *American Review of Respiratory Disease*, Vol. 140, pp. 1017-1020, 1989.

15. Kaisla T., A. R. Sovijärvi, P. Piirilä, H. M. Rajala, S. Haltsonen and T. Rosqvist, "Validated Method for Automatic Detection of lung Sound Crackles", *Medical and Biological Engineering and Computing*, Vol. 29, pp. 517-521, 1991.

16. Sankur B., E. Ç. Güler, Y. P. Kahya, "Multiresolution Biological Transient Extraction Applied to Respiratory Crackles", *Computers in Biology and Medicine*, Vol. 26, No. 1, pp. 25-39, 1996.

17. Forkheim K. E., D. Scuse, H. Pasterkamp, "A Comparison of Neural Network Models for Wheeze Detection", *Proceedings of IEEE WES-CANEX*, pp. 214-219, Winnipeg, Canada, 15-16 May 1995.

18. Rietveld S., M. Oud and E. H. Dooije, "Classification of Asthmatic Breath Sounds: Preliminary Results of the Classifying Capacity of Human Examiners versus Artificial Neural Networks", *Computers and Biomedical Research*, Vol. 32, No. 5, pp. 440-448, 1999.

19. Oud M., E. H. Dooijes and J. S. van der Zee, "Asthmatic Airways Obstruction Assessment Based on Detailed Analysis of Respiratory Sound Spectra", *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 11, pp. 1450-1455, 2000.

20. Hadjileontiadis L. J., "Discrimination Analysis of Discontinuous Breath Sounds Using Higher-order Crossings", *Medical and Biological Engineering and Computing*, Vol. 41, pp. 445-455, 2003.

21. Taplidou S.A., L. J. Hadjileontiadis, T. Penzel, V. Gross and S. M. Panas, "WED: An Efficient Wheezing-episode Detector Based on Breath Sounds Spectrogram Analysis", *Proceedings of the 25th Annual International Conference of the IEEE-EMBS*, Vol.3, pp. 2531-2534, 17-21 Sept. 2003.

22. Homs-Corbera A., J. A. Fiz, J. Morera and R. Jane, "Time-frequency Detection and Analysis of Wheezes During Forced Exhalation", *IEEE Transactions on Biomedical Engineering*, Vol.51, No.1, pp.182-186, 2004.

23. Taplidou S.A., L. J. Hadjileontiadis, I. K. Kitsas, K. I. Panoulas, T. Penzel, V. Gross and S. M. Panas, "On Applying Continuous Wavelet Transform in Wheeze Analysis"', *Proceedings of the 26th Annual International Conference of the IEEE-EMBS*, Vol.2, pp. 3832-3835, 1-5 September 2004.

24. Bahoura M. and C. Pelletier, "Respiratory Sounds Classification Using Gaussian Mixture Models", *Proceedings of Canadian Conference on Electrical and Computer*

Engineering, Vol.3, pp.1309-1312, 2-5 May 2004.

25. Taplidou S. A. and L. J. Hadjileontiadis, "Wheeze Detection Based on Time-frequency Analysis of Breath Sounds", *Computers in Biology and Medicine*, Vol. 37, No. 8, pp. 1073-1083, 2007.

26. Xiaoguang L. and M. Bahoura, "An Integrated Automated System for Crackles Extraction and Classification", *Biomedical Signal Processing and Control*, Vol. 3, No. 3, pp. 244-254, 2008.

27. Feng J., S. Krishnan and F. Sattar, "Adventitious Sounds Identification and Extraction Using Temporal-Spectral Dominance-Based Features", *IEEE Transactions on Biomedical Engineering*, Vol. 58, No. 11, pp. 3078-3087, 2011.

28. Ono M., K. Arakawa, M. Mori, T. Sugimoto and H. Harashima, "Separation of Fine Crackles from Vesicular Sounds by a Nonlinear Digital Filter", *IEEE Transactions on Biomedical Engineering*, Vol. 36, No. 2, 1989.

29. Hadjileontiadis L. J. and S. M. Panas, "Separation of Discontinuous Adventitious Sounds from Vesicular Sounds Using a Wavelet-based Filter", *IEEE Transactions on Biomedical Engineering*, Vol. 44, No. 12, pp. 1269-1281, 1997.

30. Yeginer M. and Y. P. Kahya, "Elimination of Vesicular Sounds from Pulmonary Crackle Waveforms", *Computer Methods and Programs in Biomedicine*, Vol. 89, No. 1, pp. 1-13, 2008.

31. Moussavi Z. K., M. T. Leopando, H. Pasterkamp and G. Rempel, "Computerized Acoustical Respiratory Phase Detection without Airflow Measurement", *Medical and Biological Engineering and Computing*, Vol. 38, pp. 198-203, 2000.

32. Çiftci K. and Y. P. Kahya, "Respiratory Airflow Estimation by Time Varying Autoregressive Modeling", *Proceedings of the 30th Annual International Conference of the IEEE-EMBS*, pp. 347-350, Vancouver, Canada, 20-25 August 2008.

33. Kandaswamy A., C. S. Kumar, R. P. Ramanathan, S. Jayaramana and N. Malmurugana, "Neural Classification of Lung Sounds Using Wavelet Coefficients", *Computers in Biology and Medicine*, Vol. 34, No. 6, pp. 523-537, 2004.

34. Dokur Z., "Respiratory Sound Classification by Using an Incremental Supervised Neural Network", *Pattern Analysis and Applications*, Vol. 12, No. 4, pp. 309-319, 2009.

35. Mayorga P., C. Druzgalski, R. L. Morelos, O. H. Gonzalez and J. Vidales, "Acoustics Based Assessment of Respiratory Diseases Using GMM Classification", *Proceedings of the 32nd IEEE-EMBS*, pp. 6312-6316, 31 August - 4 September 2010.

36. Engin T., E. Ç. Güler , B. Sankur and Y. P. Kahya, "Comparison of AR-based Classifiers for Respiratory Sounds", *The 6th European Signal Processing Conference (EUSIPCO)*, pp. 1745-1748, Brussels, Belgium, 24-27 August 1992.

37. Sankur B., Y. P. Kahya, E. Ç. Güler, and T. Engin, "Comparison of AR-based Algorithms for Respiratory Sounds Classification", *Computers in Biology and Medicine*, Vol. 24, No. 1, pp. 67-76, 1994.

38. Sezgin M. C., Z. Dokur, T. Olmez and M. Korurek, "Classification of Respiratory Sounds by Using an Artificial Neural Network", *Proceedings of the 23rd Annual International Conference of the IEEE-EMBS*, pp. 697- 699, 2001.

39. Güler E. Ç., B. Sankur, Y. P. Kahya and S. Raudys, "Two-stage Classification of Respiratory Sound Patterns", *Computers in Biology and Medicine*, Vol. 35, No. 1, pp. 67-83, 2005.

40. Alsmadi S. and Y. P. Kahya, "Design of a DSP-based Instrument for Real-time Classification of Pulmonary Sounds", *Computers in Biology and Medicine*, Vol. 38, No. 1, pp. 53-61, 2008.

41. Charleston-Villalobos S., G. Martinez-Hernandez, R. Gonzalez-Camarena, G. Chi-

Lem, J.G. Carrillo and T. Aljama-Corrales, "Assessment of Multichannel Lung Sounds Parameterization for Two-class Classification in Interstitial Lung Disease Patients", *Computers in Biology and Medicine*, Vol. 41, No. 7, pp. 473-482, 2011.

42. Kompis M., H. Pasterkamp and G. R. Wodicka, "Acoustic Imaging of the Human Chest", *Chest*, Vol. 120, No. 4, pp. 1309-1321, 2001.

43. Charleston-Villalobos S., S. Cortés-Rubiano, R. González-Camerena, G. Chi-Lem and T. Aljama-Corrales, "Respiratory Acoustic Thoracic Imaging (RATHI): Assessing deterministic interpolation techniques", *Medical and Biological Engineering and Computing*, Vol. 42, No. 5, pp. 618-626, 2004.

44. Charleston-Villalobos S., G. Dorantes-Mendez, R. Gonzalez-Camarena, G. Chi-Lem, J. Carrillo and T. Aljama-Corrales, "Acoustic Thoracic Image of Crackle Sounds Using Linear and Nonlinear Processing Techniques", *Medical and Biological Engineering and Computing*, Vol. 49, No. 1, pp. 15-24, 2011.

45. CORSA Guidelines, *European Respiratory Review*, Vol. 10, No. 77, 2000.

46. American Thoracic Society, "Updated Nomenclature for Membership Relation", ATS News, Vol. 3, pp. 5–6, 1977.

47. Ryu J. H. and P. D. Scanlon, "Obstructive Lung Diseases: COPD, Asthma, and Many Imitators", *Mayo Clinic Proceedings*, Vol. 76, No. 11, pp. 1144-1153, 2001.

48. National Asthma Education and Prevention Program Expert Panel Report 2: Guidelines for the Diagnosis and Management of Asthma, No. 97-4051, National Institutes of Health Publication, 1997.

49. Celli B. R., W. MacNee, A. Agusti, A. Anzueto, B. Berg, A.S. Buist, P.M.A. Calverley, N. Chavannes, T. Dillard, B. Fahy, et al., "Standards for the Diagnosis and Treatment of Patients with COPD: A Summary of the ATS/ERS Position Paper", *European Respiratory Journal*, Vol. 23, No. 6, pp. 932-946, 2004.

50. Van Schayck C. P., "Diagnosis of Asthma and Chronic Obstructive Pulmonary Disease in General Practice", *The British Journal of General Practice*, Vol. 46, No. 404, pp. 193-197, 1996.

51. Van Wheel C., "Underdiagnosis of Asthma and COPD: Is the General Practitioner to Blame?", *Monaldi Archives for Chest Disease*, Vol. 57, No. 1, pp. 65-68, 2002.

52. Bellia V., S. Battaglia, F. Catalano, N. Scichilone, R. A. Incalzi, C. Imperiale and F. Rengo, "Aging and Disability Affect Misdiagnosis of COPD in Elderly Asthmatics: The SARA Study", *Chest*, Vol. 123, No. 4, pp. 1066-72, 2003.

53. Tinkelman D. G., D. B. Price, R. J. Nordyke, R. J. Halbert, S. Isonaka, D. Nonikov and C. P. van Schayck, "Symptom-based Questionnaire for Differentiating COPD and Asthma", *Respiration*, Vol. 73, No. 3, pp. 296-305, 2006.

54. Roberts S. D., M. O. Farber, K. S. Knox, G. S. Phillips, N. Y. Bhatt, J. G. Mastronarde and K. L. Wood, "FEV_1/FVC Ratio of 70% Misclassifies Patients with Obstruction at the Extremes of Age", *Chest*, Vol. 130, No. 1, pp. 200-6, 2006.

55. Malmberg L. P., L. Pesu and A. R. A. Sovijarvi, "Significant Differences in Flow Standardised Breath Sound Spectra in Patients with Chronic Obstructive Pulmonary Disease, Stable Asthma, and Healthy Lungs", *Thorax*, Vol. 50, pp. 1285-1291, 1995.

56. Sen I. and Y. P. Kahya, "A Multi-Channel Device for Respiratory Sound Data Acquisition and Transient Detection", *Proceedings of the 27th Annual International Conference of IEEE-EMBS*, pp. 6658-6661, 2005.

57. Şen İ., *A Multi-channel Device for Respiratory Sound Data Acquisition and Transient Detection*, M.S. Thesis, Boğaziçi University, 2005.

58. Lütkepohl H., *New Introduction to Multiple Time Series Analysis*, Springer, Germany, 2005.

59. Alpaydın E., *Introduction to Machine Learning*, The MIT Press, Cambridge, 2004.

60. Chang C.-C. and C.-J. Lin, "LIBSVM : A Library for Support Vector Machines", *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, pp. 27:1-27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

61. Hyvarinen A., J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley, New York, 2001.