journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Exploring an optimal vector autoregressive model for multi-channel pulmonary sound data

Ipek Sen, Murat Saraclar, Yasemin P. Kahya*

Department of Electrical and Electronics Engineering, Bogazici University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 18 August 2011
Received in revised form
6 February 2013
Accepted 18 May 2013

Keywords:
Multi-channel pulmonary sounds
Multi-variate signal analysis
VAR modeling
Goodness of fit
Flow normalization

ABSTRACT

The purpose of this study is to find a useful mathematical model for multi-channel pulmonary sound data. Vector auto-regressive (VAR) model schema is adopted and the best set of arguments, namely, the order and sample size of the model and the sampling rate of the data, is aimed to be determined. Both conventional prediction error criteria and a set of three new criteria which are derived specifically for pulmonary sound signals are used to evaluate the success of the model. In terms of these criteria, the second order 250-point model is selected to be the most descriptive VAR model for 14-channel pulmonary sound data. The preferred sampling rate is the original data acquisition rate, which is 9600 samples per second. The effect of normalization of the data with respect to the air flow is also examined. Six normalization schemes are implemented on the data prior to the model fit, and the resulting model parameters are examined in terms of the proposed criterion measures. It is concluded that normalization with respect to flow is not necessary prior to the VAR modeling of pulmonary sound data.

1. Introduction

Auscultation provides invaluable information about the condition of the lungs, hence is a common practice for the diagnosis of pulmonary diseases. A pathological condition in the airways or alveoli is usually revealed in the altered characteristics of pulmonary sounds heard on the chest. While the physician relies on experience to interpret the sounds heard via a stethoscope, computerized lung sound analysis techniques that have developed over the past decades enable quantitative and more objective analysis.

Multi-channel pulmonary sound acquisition provides means to observe multiple locations on the chest wall simultaneously, which is not possible with stethoscope auscultation. Through the analysis of the multi-channel data, the location of the underlying pathology in the lungs can be estimated and the distribution of adventitious sound components on the chest wall can be extracted. Moreover, visual representations may be constructed to depict these findings, with a view to produce an acoustic mapping of the pulmonary system, which may offer a cheap, practical, and harmless alternative where advanced techniques are not available (e.g., in deprived regions), difficult to use (e.g., for people with physical difficulties), or harmful for the patient (e.g., small children or pregnant women).

In recent years, several studies have been carried on acoustic-based pulmonary system mapping [1–3]. In one study [1], an algorithm to find the location of the original crackle in the lungs using its amplitude and time-delay properties is proposed. In [2] simulated crackles are added onto normal sounds and detected and counted producing an image depicting regional densities of crackles. The work in [3] on the other hand deals with pulmonary sound intensities of the

* *Corresponding author at*: Bogazici University, Electrical and Electronics Engineering Department, 34342 Bebek, Istanbul, Turkey. Tel.: +90 212 3597475; fax: +90 212 2872465.
E-mail addresses: ipek.sen@boun.edu.tr, senipek@gmail.com (I. Sen).
0169-2607/$ – see front matter
http://dx.doi.org/10.1016/j.cmpb.2013.05.007

Nomenclature

A_i	$(K \times K)$ coefficient matrix of VAR(p) model, $i = 1, 2, \ldots, p$
$AIC(m)$	Akaike's Information Criterion for model order m
$\mathbf{B}$	matrix in form $[\mathbf{v}, \mathbf{A}_1, \ldots, \mathbf{A}_p]$ in VAR estimation
$\hat{\mathbf{B}}$	estimated $\mathbf{B}$ after omitting $\hat{\mathbf{v}}$, i.e., in form $[\hat{\mathbf{A}}_1, \ldots, \hat{\mathbf{A}}_p]$
$\hat{\mathbf{B}}^s_{\{i,j\},w}$	$\hat{\mathbf{B}}$ for the wth segment of the jth flow-phase of the ith full flow-cycle of the sth subject
$(EFCN)^s_j$	effective number of full flow-cycles for the jth flow-phase of the sth subject, such that $(EFCN)^s_j \le (FCN)_s$
$(FCN)_s$	number of full flow-cycles for the sth subject
$FPE(m)$	final prediction error for model order m
$HQ(m)$	Hannan-Quinn Criterion for model order m
J	set of flow-phase indices, $j = 1, 2, \ldots, 6$
K	dimensionality of the VAR(p) model (i.e., number of data channels)
$\mathbf{M}^s_{i,j}$	average of $\hat{\mathbf{B}}^s_{\{i,j\},w}$ over indices w, i.e., over segments fitting into $P^s_{i,j}$
$\mathbf{M}^s_{\cdot j}$	average of $\mathbf{M}^s_{i,j}$ over indices i
$MSE(m)$	mean squared error associated with the model fit when the order is m
N	sample size of the K-variate data segment for the model to be fitted on
$P^s_{i,j}$	the jth flow-phase of the ith full flow-cycle of the sth subject
$R^2(m)$	coefficient of determination for model order m
S	set of subject indices, $s = 1, 2, \ldots, 30$
S_B	between-group scatter score
S_W	within-group scatter score
$SC(m)$	Schwartz Criterion for model order m
$\mathbf{U}$	matrix of noise vectors $\mathbf{u}_n$ in VAR estimation
$\mathbf{u}_n$	$(K \times 1)$ random vector at time sample n, white noise of VAR(p) model
$\mathbf{v}$	$(K \times 1)$ vector of intercept terms in VAR(p) model
$(WN)^s_{i,j}$	maximum number of N-point 50% overlapping segments fitting into $P^s_{i,j}$
$\mathbf{Y}$	matrix of data vectors $\mathbf{y}_n$ in VAR estimation
$\mathbf{y}_n$	$(K \times 1)$ random vector at time sample n, output of VAR(p) model
$\mathbf{Z}$	matrix of $\mathbf{Z}_n$ in VAR estimation.
$\mathbf{Z}_n$	matrix of previous $\mathbf{y}_n$ in VAR estimation
μ^s_1, μ^g_2, μ_3	first measure for subject s, second measure for group g, and third measure for the overall performance
$\lvert.\rvert$	size, if $\cdot$ is a set; determinant, if $\cdot$ is a matrix.

healthy group, calculating the signal envelopes at individual measurement locations and obtaining an intensity map via interpolation to display ventilation. Development of an acoustic mapping technique that uses the multi-variate information and reveals the characteristics of spatio-temporal relationships for both healthy and pathological groups regardless of the existence of a specific type of adventitious sound component is still open for exploration.

As an initial step toward the final aim of developing such a technique, the motivation behind this work is to find a useful mathematical model for multi-channel pulmonary sound data. Physical lung models have been proposed previously in literature [4–7], where generally the lungs were represented by equivalent electrical circuits. Besides, studies exist that explore how sound is transmitted within the lungs [8–14]. Instead of adopting a physical approach, a statistical mathematical model is pursued in this work because the model parameters are intended to be later employed in a diagnostic system within the aforementioned frame of acoustic mapping.

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 without defying its general behavior is still a good starting point before developing a deeper perspective, since the complex mechanism is unknown. The sound signal acquired at each microphone location on the chest wall is actually 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 acquired. Auto-regressive (AR) models, which imply causal linear filters, are known to be successful to model pulmonary sounds [2,15–19]. Since the channels are also interrelated when multi-channel pulmonary sound data are measured simultaneously, the multi-variate version of the linear univariate AR setup, vector auto-regressive (VAR) model schema, has been adopted in this work with an aim to find the optimum parameters to represent these signals.

The two arguments (i.e., free variables) that need to be determined to estimate the VAR model parameters are the model order and the sample size. 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 in this study, introducing data sampling rate as the third argument. 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.

The first purpose of this study is to find the optimal set of these three arguments that yields the most descriptive VAR model of pulmonary sound data that will eventually provide diagnostic value to pulmonary sound analysis. If the aim of the VAR modeling were to predict the future values of the signal (e.g., as in economics or meteorology), then the model that minimizes the prediction error would have been selected as the best. There are several criterion functions defined in this sense in literature, the use of which constitute the traditional procedure to determine the goodness of fit of the model. In this study, however, the aim of the VAR modeling is to eventually use the model parameters in a diagnostic system. Therefore, the model that maximizes the ability of those parameters to represent the pulmonary sound data characteristics is regarded to be the most suitable model. Accordingly, three new criteria have been proposed and the associated