

# **SIMULATION OF AN EXPERIMENTAL DATABASE OF INFRARED SPECTRA OF COMPLEX GASEOUS MIXTURES FOR DETECTING SPECIFIC SUBSTANCES. THE CASE OF DRUG PRECURSORS.**

*Marco Calderisi<sup>a,b</sup>, Alessandro Ulrici<sup>a,b,\*</sup>, Sauli Sinisalo<sup>d</sup>, Juho Uotila<sup>d</sup>, Renato Seeber<sup>b,c</sup>*

<sup>a</sup> Dipartimento di Scienze della Vita, Università di Modena e Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122 Reggio Emilia; <sup>b</sup> Consorzio INSTM, Via G. Giusti 9, 50121 Firenze, Italy; <sup>c</sup> Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via G. Campi 183, 41125 Modena; <sup>d</sup> Gasera Ltd., Tykistökatu 4, FIN-20520 Turku, Finland

\*Corresponding Author: [alessandro.ulrici@unimore.it](mailto:alessandro.ulrici@unimore.it)

## **ABSTRACT**

This work is motivated by the need to develop suitable databases in absence of real experimental data, for instance when spectra measured with a newly developed instrumentation on real samples are not available yet. This notwithstanding, in fact, the realization of the physical project should be addressed by a starting database, also invaluable in order to test its effectiveness. In this article we face the issue of simulating gas mixtures spectra for the development of a new sensor for External Cavity-Quantum Cascade Laser Photoacoustic Spectroscopy (EC-QCLPAS) starting from literature FT-IR spectra of pure components: a dataset is realized suitable to realistically represent the ensemble of spectra of the gas mixtures of interest. The informative data deriving from the literature spectra were combined with the stochastic component extracted from a sample spectrum recorded with a prototype instrument, allowing us to build a matrix containing thousands of simulated spectra of gaseous mixtures, accounting for the presence of different components at different concentrations. Signal processing and experimental design techniques were used along the whole path leading to the dataset of simulated spectra. In particular, the goal of the construction of the database lies in the development of a final system to detect drug precursors in the vapour phase. The comparison of some EC-QCLPAS spectra with the corresponding simulated signals confirms the validity of the proposed approach.

**Keywords:** spectra simulation; experimental design; Fast Wavelet Transform; FT-IR; Laser Photoacoustic Spectroscopy

## 1. INTRODUCTION

The increasing need for portable sensors suitable to detect dangerous or illegal chemical substances and requiring fast in situ analysis, forces to develop sophisticated systems, exploiting at best the huge potentialities recently made available by precision mechanics, optics, and electronics. The realization of advanced detection systems, however, is often based on sensors whose development is still in progress and whose response could be, therefore, at least partly unknown. For example, the implementation of a new sensor for the detection of a particular target molecule would require that its response to the target and to possible interfering molecules is already known, but it often happens that experimental data are not available yet, the actual development of the physical device being still in progress. A preliminary database on which to work could be also necessary when the chemical substances of interest are hard to handle, due to their toxicity, to hazards, or to legal issues.

The challenging task of building a suitable starting database can be faced by following different strategies, depending on the specific situation. In particular, in this article we face the issue of simulating gas mixtures spectra for the development of a new sensor for External Cavity-Quantum Cascade Laser Photoacoustic Spectroscopy (EC-QCLPAS) working in the Mid Infrared (MIR) spectral range. From a literature survey, a first possible way to obtain the spectral responses consists in calculating them by physico-chemical simulations based on the chemical structures. *Ab initio* and semi-empirical quantum mechanical methods are widely used to this purpose, constituting invaluable tools in modern vibrational spectroscopy. However, *ab initio* calculations are very time consuming and require powerful computers. For this reason, similar methods cannot be adopted routinely, even for compounds of moderate complexity, e.g., of 100 – 300 Da: a much more practical approach lies in the use of semi-empirical methods [1-3]. A further alternative procedure has been followed by Babkov et al. [4], which simulated vibrational spectra by the method of fragments [5]. Such a method is applicable also to the simulation of the vibrational spectra of large

molecules [6,7]. The choice of computing the spectra using quantum mechanical methods presents in all cases important drawbacks, mainly due to the approximations affecting calculations and to the noise issue: a proper simulation should be suitable to represent not only the chemical information (pure spectrum), but also the contribution of the instrumental noise to the final signal shape. Moreover, in case there is the need to simulate the spectrum of mixtures of different substances, the excessive use of approximations could lead to spectral responses that are far from the real ones, not to mention the inherent computational load.

A literature survey has shown that a few attempts are reported in which the simulation of single gas spectra or spectra of specific gas mixtures takes also into account the contribution of instrumental noise and of background signal variations [8,9]. Huang et al. [10] developed a simulation model based on the infrared transmission theory and on the knowledge of background and interference spectra. Sulub et al. [11] used previously measured molar absorptivities and solvent displacement factors, computing synthetic spectra using experimentally recorded background signals. Bak proposed an interesting procedure based on Principal Component Regression to simulate spectra of a single gas (CO) at varying concentrations, temperatures, and pathlengths [12, 13]. Other approaches have been followed by Corsi et al. [14], that developed a simulation approach to reveal VOC in air, and by Gao et al. [15], who proposed a method to simulate spectra of polluting gases in a complex environment.

In this work we deal with the issue of simulating spectra in the MIR range, for the development of a new EC-QCLPAS sensor [8, 9, 16], devoted to the identification of vapours of drug precursors (target molecules) in air, considering a high variety of possible environmental conditions [17, 18]. The MIR spectral range has been chosen since most chemicals, such as the target gases selected for our purposes, exhibit strong characteristic rotovibrational absorption bands in this wavenumber interval [19]. Due to the need to account for the possible presence of a large number of chemical species mixed in varying amounts, we started from spectra of pure substances taken from literature

databases, that were considered as “building blocks” of mixtures, to simulate the complex spectral profiles of interest which, to a first approximation, will be obtained by the sensor under development. In addition to the target molecules, i.e., the drug precursors, other components of the gaseous mixtures were the most common interfering species and the main components of air. Reasonable concentration ranges were considered for each species.

Noteworthy, the bulk of the developed procedure can be adapted to deal with a number of different situations. Moreover, it should be also noticed that, although the gas sensor considered here operates in the MIR spectral range, the proposed simulation methodology can be easily transferred to any other kind of spectroscopic data.

The approach proposed here requires some basic issues to be addressed, mainly involving signal processing and experimental design methodologies [20]. The gas spectra were first denoised by means of a Fast Wavelet Transform (FWT) [21-23] based algorithm. Then, a procedure was developed to multiply the spectra of the pure components by the corresponding concentration values, in order to operate only on the actual absorption bands. The noise structure of the EC-QCLPAS was analyzed using a spectrum recorded on methanol with a prototype instrument, and added to the “clean” spectra of the simulated mixtures. This approach allowed us to build a matrix containing the simulated EC-QCLPAS spectra of thousands of gas mixtures, which was further used [24] for the definition of the optimal spectral working range and even for identification of the single most informative wavenumbers within this range. In order to preliminarily test the proposed approach, the whole procedure has been applied to simulate the experimental spectra acquired with prototypes of the instrument on two sample mixtures.

## 2. DATASETS

### 2.1. Literature FT-IR Spectra

Two literature databases, namely PNNL [25] (Pacific Northwest National Laboratory) and HITRAN [26, 27] (HIGH-resolution TRANsmission molecular absorption database, Atomic and Molecular Physics Division, Harvard-Smithsonian Center for Astrophysics) were exploited for extraction of spectra. The most part of the spectral data was taken from the PNNL database, and consisted of FT-IR spectra acquired at 298 K, at wavenumbers ranging from 6500 to 600  $\text{cm}^{-1}$ , 0.0603  $\text{cm}^{-1}$  spectral resolution, corresponding to 97902 data. Moreover, a few spectra have been extracted from the HITRAN database, referred to 296 K, within a wavenumber range from 3174 to 749  $\text{cm}^{-1}$  at a resolution of 0.076  $\text{cm}^{-1}$ .

As a whole, for our specific needs, the spectra of 33 chemical species were considered, divided into three categories:

- 4 target spectra, namely 1-phenyl-2-propanone, acetic anhydride, ephedrine and safrole (concentration range: 0.02 - 1 ppm)
- 20 interfering species (pollutants)
- 9 air components.

Table 1 reports the list of the considered pollutants, together with the relevant upper concentration limits (the lower limit was set to 1 ppb for all of them), and Table 2 reports the air components, along with their typical and maximum concentration values, as inferred from literature [8, 28]. The list of pollutants was defined taking into account both the similarities of the spectral signatures with those of target gases and the most likely environmental conditions.

## 2.2. Spectra measured with EC-QCL-PAS prototype instrument

Real gas spectra have been recorded by a prototype of the EC-QCL-PAS instrument, in order to estimate the actual noise contribution and to collect preliminary experimental spectra suitable to test the validity of the developed procedure. The laser photoacoustic (LPAS) sensor (Figure 1) is composed by a cantilever enhanced photoacoustic cell (PA-cell) from Gasera Ltd. and by two different external cavity quantum cascade laser (EC-QCL) sources, i.e., a continuous wave version and a pulsed one. The continuous wave laser source was a 21095-MHF EC-QCL (Daylight Solutions) laser with about 50 mW average power over the tuning range (1020 – 1100  $\text{cm}^{-1}$ ). The pulsed laser source was a TLS-11096 (Daylight Solutions) laser with on average 5 mW effective power (effective duty cycle 5 %) over the tuning range (1030 – 1110  $\text{cm}^{-1}$ ). The EC-QCL system creates a collimated laser beam with relatively high optical output power, in the MIR region. Continuous wave QCL should be kept at a constant temperature; in particular, the considered device was thermostated at 23 °C by water cooling system. The laser power is modulated mechanically, by a rotating chopper wheel, at a frequency of 60 Hz. The collimated and modulated laser beam travels through the cylindrically shaped photoacoustic cell, which is 9.5 cm in length, 4 mm in diameter and is sealed by two windows at the ends. The sample gas absorbs the modulated infrared radiation and it heats and cools down periodically. The generated pressure wave is detected by the cantilever, whose position is measured by the laser interferometer. The amplitude of the cantilever oscillations at the modulation frequency corresponds to the intensity of the photoacoustic signal.

A first spectrum was measured using the continuous wave laser source, at 1 bar pressure, on 90 ppm  $\text{CH}_3\text{OH}$  diluted in nitrogen, and was used to estimate the noise structure of the EC-QCL-PAS signals. A second spectrum was measured on the same sample using the pulsed laser source. These two spectra were then used to estimate the corresponding values of the experimental correction factor, given by the cell response and absolute laser power coefficients, which is the constant multiplication factor necessary to convert the photoacoustic intensity into the corresponding

absorbance values. The two experimental correction factors were then used to convert two spectra used to perform the preliminary test of the proposed approach. They were measured on:

- a binary mixture composed by 1.6 ppm CH<sub>3</sub>OH and 1 ppm NH<sub>3</sub>, diluted in nitrogen at 1 bar pressure, measured with a spectral resolution equal to 0.1 cm<sup>-1</sup> with the continuous wave laser source;
- a quaternary mixture composed by H<sub>2</sub>O (10000 ppm), CO<sub>2</sub> (380 ppm), CH<sub>3</sub>OH (11 ppm) and safrole (11 ppm), diluted in nitrogen at 950 mbar pressure, measured with a spectral resolution equal to 0.5 cm<sup>-1</sup> with the pulsed laser source.

Due to the low intensity of both laser sources at the spectra extremes, only the 1038-1098 cm<sup>-1</sup> range was considered for further elaborations.

### 3. ALGORITHMS

All the calculations, both for the elaboration of the data and for the creation of the algorithms, were performed in Matlab 7<sup>®</sup> language, running on a desktop PC with Windows 7 – 64 bit<sup>®</sup>, equipped with an Intel Core<sup>®</sup> i7-2600 CPU @ 3.40 GHz processor and 4.00 GB RAM. Moreover, some of the subroutines available in the Wavelet Toolbox ver. 4.6 (The MathWorks, Inc.) and in the PLS-Toolbox ver. 7.0 (Eigenvector Research, Inc.) were employed.

#### 3.1. Importation and pre-processing of the literature spectra

The step by step procedure adopted for processing the spectra extracted from the literature databases is described in details hereafter.

##### 3.1.1. *Importing spectra*

Spectra stored in different databases and recorded with different instruments possess different characteristics. For this reason, an algorithm to import spectral data with varying wavenumber



ranges, resolutions and input file formats has been developed. All the imported literature spectra have been converted into a common format and structure. In order to denoise the database spectra at best, these have been imported using all the original points, exploiting in this manner the highest possible resolution.

### 3.1.2. Denoising

Different noise structures affect the literature spectra, and are also quite reasonably diverse from the noise structure of the spectra measured with the device under development. For this reason, it is mandatory to separate the useful spectroscopic information, that will be present also in the signals of the new instrumental device, from the noise contribution of the specific instrument used to record each literature spectrum. This goal has been achieved by using Wavelet Transform (WT) -based signal processing techniques, thanks to the ability of WT to map the analysed signal both in the original and in the relevant frequency domains at the same time. Furthermore, the use of various wavelets to decompose the signal into the WT domain allows a wide number of representations among which to choose the most effective one.

An *ad-hoc* developed Fast Wavelet Transform (FWT [21,23])-based algorithm operates on an individual signal by convolving it with two filters (called the High-pass and Low-pass decomposition wavelet filters, i.e., Hi\_D and Low\_D, respectively), and splitting it into two orthogonal subspaces: the vector of approximations, retaining only the low frequency content of the signal, and the vector of details, which collects the high frequency content, respectively. Being the two wavelet filters orthogonal to each other, the frequencies retained by Lo\_D are not brought by Hi\_D, and *vice versa*: they are fully complementary to each other, since the original signal can be perfectly reconstructed from the approximations and details vectors, by applying the proper couple of wavelet reconstruction filters. The decomposition procedure can be repeated to further decomposition levels, applying the same two filters to the approximations vector. In this way, sharp and coarse properties of the signal are disjointed and stored into different sub spaces (approximation

and detail vectors, at different levels of decomposition). In the present application, the reconstruction into the original domain employed the approximations vector at a user-defined level of decomposition, suitable to effectively remove the noise: in this way, the pure absorbance component of the spectra was obtained, i.e., the spectra cleaned from the instrumental noise. To this aim, a proper function was written that, through an interactive interface, allowed us to easily find the optimal values of the calculation parameters, i.e., wavelet type and decomposition level, thanks to the direct visualization of the resulting signals, i.e., the original spectrum, the low-frequency contribution, and the high-frequency contribution.

Different wavelet filters were considered, namely wavelet functions belonging to the Daubechies, symlets and coiflets families. The most part of the spectra were denoised using a Daubechies wavelet function (db3) at the 3<sup>rd</sup> level of decomposition. As an example, in Figure 2 a portion of the spectrum of benzene is reported, showing the original spectrum together with the denoised one (upper plot), and their difference (lower plot). The irregular shape of the signal reported in the lower plot of Figure 2 clearly confirms that the high frequency content extracted by FWT actually corresponds to a random noise affecting the spectrum, not bearing any useful information.

### 3.1.3. *Standardising spectra*

The next step consisted in the transformation of the denoised signals in order to obtain uniform datasets of standardized spectra at constant concentration (1 ppb), within the 1000-2500  $\text{cm}^{-1}$  spectral range. To this aim, the smoothed spectra are elaborated by a Matlab function that resamples the spectra at the desired resolution, into a spectral window defined by the user. First, in order to obtain an output signal resembling the hypothetical spectrum resulting from the instrument, the original spectrum is convolved with a Gaussian function with mean value of 1 and standard deviation derived from the Full Width at Half Maximum (FWHM) of the EC-QCLPAS laser line, according to the equation:

$$\sigma = \frac{\text{FWHM}}{2\sqrt{2\ln(2)}} \quad (1)$$

In this case FWHM was set to  $0.1 \text{ cm}^{-1}$ .

After convolution, the resulting signal is resampled at the desired wavenumber values by shifting a second order polynomial interpolating three subsequent points at a time.

### 3.2. Extraction of the Noise Structure of EC-QCL-PAS spectra

As already mentioned, the final version of the instrument is not available yet; however, extraction of the specific noise from a signal measured with an available prototype was possible. In order not to arbitrarily assume an homoscedastic nature of the noise, the dependence upon signal intensity was also computed. In this way, a proper noise contribution can be added to the preprocessed spectra of the mixtures.

The spectrum of  $\text{CH}_3\text{OH}$  at 90 ppm concentration was exploited to this aim, and analyzed using the same FWT-based procedure that was previously used to denoise the literature spectra.

The noise structure was defined by the following procedure:

1. identification of the optimal conditions, in terms of type of wavelet and decomposition level, to separate the informative signal ( $I$ ) from the corresponding noise ( $N$ ). In particular,  $I$  corresponds to the reconstructed approximation vector, while  $N$  can be obtained by subtracting  $I$  from the original spectrum;
2. sorting of  $I$  in ascending order ( $I'$ ), and of the noise signal  $N$  accordingly ( $N'$ );
3. subdivision of  $I'$  and  $N'$  in  $n$  intervals;
4. calculation of the mean of the intensity signal,  $I'_m$ , and of the standard deviation of  $N'$ ,  $N'_s$ , for each interval;

5. estimate of robust linear regression models of  $N'_s$  as a function of  $I'_m$ , both with and without inclusion of the intercept  $b_0$ ;
6. repetition of points 3 to 5, changing the number of intervals  $n$  (from 10 to 200, step by 10, for 20 iterations overall);

The plots of the robust regression models have been included as Supplementary Material, together with the results of the noise extraction from the spectrum of CH<sub>3</sub>OH.

Considering the values of the regression coefficients,  $b_0$  and  $b_1$ , of the relevant error estimates [ $s(b_0)$  and  $s(b_1)$ ], and of coefficients significance [ $p(b_0)$  and  $p(b_1)$ ] as a function of the different number of intervals (Figure 3), it is possible to observe that the intercept values are often not significant, and that the slope values of the models calculated by including the intercept are very similar to those calculated by setting  $b_0$  to 0. Moreover, it can be also noticed that the  $b_1$  values do not vary significantly at varying the number of intervals,  $n$ .

This led us to use the following equation to estimate the noise as a function of the signal intensity:

$$\hat{N}_s(i) = \bar{b}_1 \times I(i) \quad (2)$$

where, for each point  $i$  of the signal, the estimate of the standard deviation of noise  $\hat{N}_s(i)$  is obtained by multiplying the corresponding intensity of the denoised signal  $I(i)$  by  $\bar{b}_1$ , i.e. the mean value of the  $b_1$  coefficients of the regression models obtained for each number of intervals,  $n$ .

Then, the noise structure can be applied to a simulated spectrum  $P$  to give the corresponding final spectrum  $S$ , using the following equation:

$$S(i) = P(i) + P(i) \times \hat{N}_s(i) \times r(i) \quad (3)$$

where, for each point  $i$  of the signal,  $\hat{N}_s(i)$  is the noise estimated, by equation 2, from the intensity of the pure spectrum  $P(i)$ , and  $r(i)$  is a randomly generated number drawn from the standard normal

distribution. As an example, the result of the application of this noise structure to the smoothed spectra of the target molecules (1 ppm concentration) is reported in Figure 4.

### **3.3. Gas mixtures concentration**

#### *3.3.1. Definition of the concentration domains*

In order to simulate a spectral dataset suitable to take into account the complex variability of the composition of gaseous mixtures in real environments, Experimental Design techniques, along with an adequate randomization strategy, have been exploited. In this work more than 30 different gases have been considered, including i) target compounds, ii) interfering species (pollutants) and iii) typical air components. This required us to create a concentration matrix accounting for the composition of the mixtures, going through the separate elaboration of 3 different concentration matrixes, *viz.* one for each class of components, that were subsequently merged. In details, variable amounts of 4 target molecules and of 20 interfering species were considered to be possibly present, in addition to the 9 gases typically found in the atmosphere that are detectable by MIR spectroscopy.

As for the target molecules, in order to simulate an homogeneous set of possible gas mixtures, an approximately balanced number of mixtures containing 1, 2, 3 or 4 target gases has been formulated. To address this issue, all the possible combinations, *i.e.*, one single species, all the possible couples of 2 targets, all the possible terns of 3 targets, and the 4 targets altogether, were considered. For each combination a full Factorial Design (FD) was adopted to simulate a set of possible concentration values. The number of concentration levels of each FD was purposely managed, as reported in Table 3, leading to 499 mixtures of target species on the whole.

In view of the high number of considered pollutants, however taking into account that the simultaneous presence of a high number of them is unrealistic, only mixtures containing from 1 to 3 interfering species were considered. In order to realize a balanced design, 3 FDs have been

developed, considering 13, 4 and 3 concentration levels for each combination of 1, 2 and 3 interfering species, respectively, which leads to the number of mixtures reported in Table 4. The whole number of mixtures of pollutants results equal to 34080. The consideration of all these mixtures is actually impractical, so that a subsampling procedure, which will be described in the following section, was adopted.

In order to properly define experimental domains for both targets and pollutants, it was important to cover quite wide concentration ranges, but at the same time it was also appropriate to consider a relatively high number of mixtures containing low concentrations, since low values are most likely to be found in real scenarios. To this aim, the experimental designs with the lower number of concentration levels (from 1 to 5) were built using a non-uniform spacing criterion for the concentrations, by defining a dyadic sequence such that, for example, for a 5 level FD in the range from 0 to 1, the non-uniform spacing leads to the following levels: 0, 1/8, 1/4, 1/2, 1.

For the air components a different planning scheme has been adopted, since in this case all the 9 components must be always included. Therefore, considering that the relevant concentrations follow lognormal distributions, their values were randomly generated from similar distributions, defined for each single species. To this aim, starting from the maximum ( $MAX_{LIT}$ ) and the typical ( $TYP_{LIT}$ ) environmental concentration values, as inferred from literature [28] (Table 2), a function was written to generate the corresponding lognormal frequency distribution function. Such a procedure was followed to obtain 1000 concentration values for each air component. Moreover, in order to avoid the risk to obtain unrealistic results, the random generation of the concentration values was repeated until the final dataset satisfied the following constrains:

- in order to guarantee that the most part of the range between  $TYP_{LIT}$  and  $MAX_{LIT}$  is covered, the highest concentration value falling within this range is forced to result  $\geq 0.9 \times MAX_{LIT}$ ;
- concentration values higher than  $10 \times MAX_{LIT}$  are discarded;

- in order to simulate anomalous though realistic situations, such as very high CO<sub>2</sub> concentrations that could be found, for instance, within a sealed shipping container, 2 to 5 concentration values  $> \text{MAX}_{\text{LIT}}$ , and one value  $> 5 \times \text{MAX}_{\text{LIT}}$  have been considered;
- unrealistic H<sub>2</sub>O concentration values ( $> 10^8$  ppb) are not considered.

### 3.3.2. *Final mixture concentration matrix*

The final mixture concentration matrix was generated according to the following scheme (Figure 5):

1. 501 rows of zeros were added at the end of the matrix relative to the 499 target mixtures, to give a target concentration matrix of size  $\{1000 \times 4\}$ , composed by 1000 mixtures of the 4 target molecules;
2. 250 mixtures with one pollutant, 250 mixtures with 2 pollutants, and 250 mixtures with 3 pollutants were randomly selected from the whole set of 34080 pollutant mixtures, and the resulting matrix was added with 250 rows of zeros; the 1000 rows were finally shuffled randomly, to give a pollutants concentration matrix of size  $\{1000 \times 20\}$ , composed by 1000 mixtures of the 20 pollutants;
3. in order to guarantee that all the 20 pollutants are sufficiently represented, step 2 was repeated until each pollutant is present in 65 mixtures at least;
4. the 1000 air concentration values for each air component were merged, to give an air components concentration matrix of size  $\{1000 \times 9\}$ , consisting of 1000 mixtures of the 9 air components;
5. the 3 matrices were merged, to give the final mixture concentration matrix of size  $\{1000 \times 33\}$ , bearing the concentration values of each single component (33 columns) for a set of 1000 gas mixtures (rows).

This procedure was iterated 5 times, in order to build a final data matrix with 5000 different mixtures, which is supposed to give efficiently account for a real scenario.

### 3.4. Spectra multiplication

#### 3.4.1. Sigmoidal Weighting Function

Once the matrices of the denoised literature spectra and of the concentrations of the mixtures are ready, they can be used to build up the matrix of the spectra profiles for the gas mixtures. Under the quite reasonable assumption of a linear relationship between the concentration of a given pure chemical species and the intensity of the relevant spectral bands, the simple multiplication of each unit concentration spectrum by the relevant concentration level within each gas mixture, gives unrealistic results. By such a procedure, in fact, the background of the unit concentration spectrum is forced to change accordingly to the corresponding concentration value, and the correct signal shape is therefore not preserved. For this reason, an algorithm based on a Sigmoidal Weighting Function (SWF) was developed to multiply correctly the spectra ascribed to the actual sample under exam by the corresponding concentrations, preserving the background intensity and shape, hence only operating on the true absorption bands.

In detail, given a literature spectrum,  $S$ , consisting of  $1 \leq i \leq p$  data points, the corresponding spectrum  $R$ , resulting from multiplication by a concentration value  $c \geq 1$  ppb, and weighted by the SWF, is obtained for each point  $i$  by the equation:

$$R(i) = \left[ 1 + \frac{1}{1 + e^{-a \left( \frac{S(i) - S_{MIN}}{S_{MAX} - S_{MIN}} - m \right)}} \times (c - 1) \right] \times S(i) \quad (4)$$

where  $S_{MIN}$  and  $S_{MAX}$  are the minimum and maximum intensity values of signal  $S$ , respectively, and  $a$  and  $m$  are two adjustable parameters used to define the shape of the SWF. Using this function, the lowest intensity values of the signal  $S$  are multiplied by 1, while the highest values are linearly multiplied by the proper concentration value,  $c$ . Intermediate intensity values are multiplied by a factor ranging from 1 to  $c$ , which depends on the values of the parameters  $a$  and  $m$ . In particular,  $m$



corresponds to the percentile of the signal intensities at which the multiplication factor is equal to  $(c+1)/2$ , while  $a$  defines the slope of the sigmoidal function. SWF, whose effects are shown in Figure 6 for the high (Figure 6.a) and low (Figure 6.b) intensity values, needs being suitably tuned by properly setting the values of  $m$  and  $a$ . In particular, in Figure 6.b the difference between the adoption of a linear multiplication of the whole spectrum by a constant concentration factor (dotted lines) and the use of the appropriate SWF (solid lines) is well evident.

#### 3.4.2. *Simulated spectral data matrix*

Each mixture spectrum of the simulated spectral data matrix is obtained by:

- using equation (4) to multiply each unit concentration spectrum by the corresponding value in the concentration mixture matrix;
- summing up all these spectra, based on the additive property of absorbance;
- adding the noise estimated by equation (3).

The 5000 mixture spectra of the final dataset include air, pollutants, and targets, at the chosen concentration levels, in the proper mixing proportions.

## 4. TEST OF THE PROPOSED APPROACH

A preliminary experimental test was performed in order to verify the validity of the proposed procedure to simulate the spectra expected from the instrument going to be realized. The signals recorded with the EC-QCL-PAS prototype instruments on a binary mixture composed by CH<sub>3</sub>OH and NH<sub>3</sub>, and on a quaternary mixture composed by H<sub>2</sub>O, CO<sub>2</sub>, CH<sub>3</sub>OH and safrole, were used to this purpose. The two different mixtures were tested with two different laser setups, as it was described in Section 2.2. Moreover, for each laser source a spectrum collected with the same EC-QCL-PAS prototype instrument on 90 ppm CH<sub>3</sub>OH was compared with the corresponding

simulated spectrum, in order to estimate the value of the constant multiplication factor necessary to convert the photoacoustic intensities into the corresponding absorbances. This constant multiplication factor was calculated as the ratio between the average of the absorbance values of the CH<sub>3</sub>OH simulated spectrum in the 1050-1060 cm<sup>-1</sup> range and the average of the EC-QCL-PAS intensity values in the same spectral range. This choice was justified by the highest intensity values of the CH<sub>3</sub>OH spectrum in this range. The constant multiplication factor was then used to convert the EC-QCLPAS mixture spectrum to an absorbance plot that, in turn, was compared to the corresponding simulated spectrum.

Figure 7 reports the comparison of the EC-QCL-PAS spectra (in black) with the corresponding simulated spectra (in gray), both for the binary NH<sub>3</sub> + CH<sub>3</sub>OH gas mixture (Figure 7.a) and for the quaternary H<sub>2</sub>O + CO<sub>2</sub> + CH<sub>3</sub>OH + safrole gas mixture (Figure 7.b). Both the simulated spectra show a satisfactory agreement with the experimentally measured ones, with respect to both the signal intensity and shape. In particular, the simulated spectrum of the binary mixture (Figure 7.a, spectral resolution 0.1 cm<sup>-1</sup>) is almost coincident with the corresponding EC-QCL-PAS spectrum, except for some very limited discrepancies in the intensity and position of the peaks with highest intensity (at about 1084.5, 1065.6 and 1046.4 cm<sup>-1</sup>) due to NH<sub>3</sub>. While the differences in the intensity values can be ascribed to the accuracy of the actual gas concentration and to the uneven distribution of the laser power curve, the slight shifts in the wavenumbers (max 0.2 cm<sup>-1</sup>) are likely due to the repeatability of the positioning of the laser source, which might cause some random changes in the absolute accuracy of the wavenumber values. A lower, though well acceptable performance was obtained for the simulation of the quaternary mixture (Figure 7.b, spectral resolution 0.5 cm<sup>-1</sup>). In this case, although the overall shape of the signal is quite well modeled, higher discrepancies are observed, especially in the range between 1060 and 1038 cm<sup>-1</sup>. While the absorbance values of the simulated and of the measured spectra are very close to each other, differences can be observed in terms of spectral resolution. This is likely due to the fact that the

laser used for the measurement of this gas mixture operated in pulsed mode, which typically broadens the linewidth of the laser itself. While the resolution used both for the acquisition of the experimental data and for the simulation of the spectra was equal to  $0.5 \text{ cm}^{-1}$ , the linewidth of the pulsed laser source is below  $1 \text{ cm}^{-1}$ : convolution of the laser linewidth with the actual spectrum occurs, which results in broadening of some sharp spectral features.

However, in general, also in view of the fact that the final instrumentation will employ a continuous wave laser source, the comparison between the simulated spectra and the corresponding EC-QCL-PAS experimental data confirms the validity of the proposed approach. This indicates that the dataset of 5000 simulated gas mixtures spectra constitutes a valuable tool to address properly the realisation of the final instrumentation.

## 5. CONCLUSIONS

In this work we faced the issue to build up a spectral database, specific of a given MIR instrument, relative to mixtures of a high number of gases. The starting point consisted of literature spectra of single species, taken from different databases, which had to be properly processed in order to possess equal features in terms of abscissa scale and range, resolution, etc. Once the noise, differently affecting the single imported spectra, has been separated from the informative trace, the specific noise of the instrument to which the mixtures database is devoted has been added, and the reliable concentrations of the different components was accounted for by a suitable concentration matrix. A series of complex spectra of the possible mixtures of the different considered species was thus built, to constitute the database sought.

It should be emphasized that, despite the specificity of the case dealt with in the present article, the algorithms developed for the various stages of the procedure are of general validity and can be very simply adapted to a variety of different experimental situations.

Based on this first stage, the optimal choice of the spectral points will be identified, in order to achieve most efficient extraction of the information sought. Even better results will be gained by designing a learning procedure that allows the system to acquire new informative spectra and possibly discard less informative ones, depending on the progressive knowledge of the characteristics of the instrument with respect to noise, possible drift, etc., as well as of the specific environment in which it operates. The flexibility of the developed software allows one to address it easily even to quite different situations.

## **6. ACKNOWLEDGEMENTS**

EU is gratefully acknowledged for the financial support within the EU FP7 for the research project CUSTOM (Drugs and Precursor Sensing by Complementing Low Cost Multiple Techniques), of which this work is part. The authors are pleased to acknowledge also INSTM (Consorzio Interuniversitario Nazionale per la Scienza e Tecnologia dei Materiali) for the financial support and project management. Dr. Jaakko Lehtinen is gratefully acknowledged for technical support in the phase of EC-QCLPAS spectra acquisition.

## REFERENCES

- [1] V.A. Basiuk, IR spectra simulation as auxiliary tool for gas chromatography-Fourier transform IR spectroscopy-mass spectrometry identification of unknown compounds: Comparison between several semi-empirical methods, *Spectrochim. Acta - Part A* 55 (1999) 289-298.
- [2] V.A. Basiuk, IR spectra simulation as auxiliary tool for gas chromatography/Fourier transform IR spectroscopy/mass spectrometry identification of unknown compounds. 2. PM3, AM1, MNDO and MINDO3 simulations for simple nitriles, *Spectrochim. Acta - Part A* 55 (1999) 2771-2782.
- [3] C. Topaçli, A. Topaçli, Semi-empirical infrared spectra simulations of benzidine and its metal chloride complexes, *J. Mol. Struct.* 658 (1-2) (2003) 9-15.
- [4] L.M. Babkov, J. Baran, N.A. Davydova, J.I. Kukielski, S.V. Trukhachev, Vibrational spectra and structure model of 2-biphenylmethanol molecule, *J. Mol. Struct.* 661-662 (1-3) (2003) 41-48.
- [5] L.A. Gribov, V.A. Dementev, *Modelirovanie kolebatelnyh spektrov sloznyh soedineniy na EVM*, Nauka, Moscow, 1989 (in Russian).
- [6] E.B. Wilson, J.C. Decius, P.C. Cross, *Molecular vibrations. The Theory of Infrared and Raman Vibrational Spectra*, McGraw-Hill, New York, 1956.
- [7] P.C. Painter, *The Theory of Vibrational Spectroscopy and its Application to Polymeric Materials*, Wiley, New York, 1982.
- [8] M.E. Webber, M. Pushkarsky, C.K.N. Patel, Optical detection of chemical warfare agents and toxic industrial chemicals: simulation, *J. Appl. Phys.* 97 (11) (2005) 113101.
- [9] I. Dunayevskiy, A. Tsekoun, M. Prasanna, R. Go, C.K.N. Patel, High-sensitivity detection of triacetone triperoxide (TATP) and its precursor acetone, *Appl. Optics* 46 (25) (2007) 6397-6404.
- [10] Y. Huang, Y.H. Fang, W. Xiong, M.P. Shen, D.C. Li, D.M. Dong, Simulation for infrared spectra of pollutant gas and parameters setting, *Guangdian Gongcheng/Opto-Electronic Engineering* 33 (6) (2006) 61-64.
- [11] Y. Sulub, G.W. Small, Spectral simulation methodology for calibration transfer of near-Infrared spectra, *Appl. Spectrosc.* 61 (4) (2007) 406-413.
- [12] J. Bak, Rapid method for simulating gas spectra using reversed PCR temperature calibration models based on Hitran data, *Appl. Spectrosc.* 53 (11) (1999) 1375-1381.
- [13] J. Bak, Modeling of gas absorption cross sections by use of principal-component-analysis model parameters, *Appl. Optics* 41 (15) (2002) 2840-2846.

- [14] M.G. Gao, W.Q. Liu, T.S. Zhang, J.G. Liu, Y.H. Lu, J. Zhu, Y. Lian, F. Lu, Passive remote sensing of VOC in atmosphere by FTIR spectrometry, *Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy and Spectral Analysis* 25 (7) (2005) 1042-1044.
- [15] C. Corsi, A. Dundee, P. Laurenzi, N. Liberatore, D. Luciani, S. Mengali, A. Mercuri, A. Pifferi, M. Simeoni, G. Tosone, R. Viola, D. Zintu, Chemical Warfare Agents Analyzer Based on Low Cost, Room Temperature, and Infrared Microbolometer Smart Sensors, *Advances in Optical Technologies* (2012) 808541.
- [16] J. Uotila, J. Lehtinen, T. Kuusela, S. Sinisalo, G. Maisons, F. Terzi, I. Tittonen, Drug precursor vapor phase sensing by cantilever enhanced photoacoustic spectroscopy and quantum cascade laser, *Proc. SPIE* 8545 (2012) 85450I.
- [17] A. Secchi, A.M. Fiorello, S. D'Auria, A. Varriale, A. Ulrici, R. Seeber, J. Uotila, V. Venditto, P. Estensoro, F. Colao, Drugs and precursor sensing by complementing low cost multiple techniques: overview of the European FP7 project CUSTOM, *Proc. SPIE* 8545 (2012) 85450G.
- [18] EU FP7 project CUSTOM - Drugs and Precursor Sensing by Complementing Low Cost Multiple Techniques, <http://www.custom-project.eu/site/index.php>
- [19] Y. Sun, K.Y. Ong, *Detection Technologies for Chemical Warfare Agents and Toxic Vapors*, CRC Press, Boca Raton, 2005.
- [20] M. Calderisi, A. Ulrici, L. Pigani, A. Secchi, R. Seeber, Experimental design-based strategy for the simulation of complex gaseous mixture spectra to detect drug precursors, *Proc. SPIE* 8545 (2012) 85450B.
- [21] B. Walczak, *Wavelets in Chemistry*, Elsevier, Amsterdam, 2000.
- [22] M. Cocchi, R. Seeber, A. Ulrici, WPTER: Wavelet packet transform for efficient pattern recognition of signals, *Chemom. Intell Lab. Syst.* 57 (2) (2001) 97-119.
- [23] M. Cocchi, R. Seeber, A. Ulrici, Multivariate calibration of analytical signals by WILMA (wavelet interface to linear modelling analysis), *J. Chemom.* 17 (8-9) (2003) 512-527.
- [24] A. Ulrici, R. Seeber, M. Calderisi, G. Foca, J. Uotila, M. Carras, A.M. Fiorello, A feature selection strategy for the analysis of spectra from a photoacoustic sensing system, *Proc. SPIE* 8545 (2012) 85450K.
- [25] S.W. Sharpe, T.J. Johnson, R.L. Sams, P.M. Chu, G.C Rhoderick, P.A. Johnson, Gas-phase databases for quantitative infrared spectroscopy, *Appl. Spectrosc.* 58 (12) (2004) 1452-1461.
- [26] L.S. Rothman, I.E. Gordon, A. Barbe, D.C. Benner, P.F. Bernath, M. Birk, et al. The HITRAN 2008 molecular spectroscopic database, *J. Quant. Spectrosc. Radiat. Transfer* 110 (9-10) (2009) 533-572.

- [27] L.S. Rothman, I.E. Gordon, Y. Babikov, A. Barbe, D.C. Benner, P.F. Bernath, M. Birk, et al. The HITRAN2012 molecular spectroscopic database, *J. Quant. Spectrosc. Radiat. Transfer* 130 (2013) 4-50.
- [28] NIOSH - National Institute for Occupational Safety and Health, <http://wwwn.cdc.gov/pubs/niosh.aspx>, and therein cited references.

## BIOGRAPHIES

**Marco Calderisi** holds a Master Degree in Chemistry (2000) from the University of Pisa, Italy and a PhD in Metabolomics (2010) from the NMR Center of The Policlinico Le Scotte, University of Siena, Italy. He works as Chemometrics consultant since 2003 and has a professional experience in R&D, mainly in chemistry industry, and in environmental analysis and monitoring. Actually he has a post-doctoral fellowship at the Department of Life Sciences of the University of Modena and Reggio Emilia. He has more than 20 articles in peer-reviewed and technical journals, and has been presenter in number of international conference.

**Alessandro Ulrici** received his Master Degree in Chemistry in 1997 and his PhD Degree in Chemistry in 2001 from the University of Modena and Reggio Emilia. Currently he is Associate Professor of Analytical Chemistry at the Department of Life Sciences of the University of Modena and Reggio Emilia. He is co-author of more than 70 research articles published on international scientific journals and book chapters. His main research interests are related to the elaboration of original chemometric algorithms for the fast and nondestructive characterization of complex systems through the use of multivariate analysis of signals and images.

**Sauli Sinisalo** received his Master's degree from the Department of Physics and Astronomy, University of Turku in 2011 from the field of Optics and Spectroscopy. Before graduating, he studied also at University of Washington, Seattle, in 2008 and started working at Gasera Ltd. in 2010 with gas analyzer development, strongly specializing in the laser based photoacoustic spectroscopy (LPAS). While closely following the cutting-edge laser technology and by applying diode-, QC-lasers and OPO's to photoacoustic trace gas detection he has also been co-authoring in a few scientific publications. He is currently a Client Partner at Gasera, responsible for managing of several EU-project contributions, client projects as well as managing product development.

**Juho Uotila** received his Master Degree in Physics in 2003 and respectively PhD. degree in Physics in 2009 from the University of Turku (Finland). He has worked as research scientist in the University of Turku and Finnish Defence Forces Technical Research Center, product manager at Gasera Ltd. and currently as electro-optics specialist in Patria Aviation Oy. His main scientific interests are photoacoustic spectroscopy and imaging electro-optical systems. He has authored or co-authored 16 scientific publications regarding cantilever enhanced photoacoustic spectroscopy.

**Renato Seeber** received his Master Degree in Chemistry from the University of Padova (Italy). He worked in different universities and became Full Professor of Analytical Chemistry in 1986. He currently works at the University of Modena and Reggio Emilia. He is co-author of more than 180 scientific publications on international journals and book series. His present main scientific interests are in i) development and characterization of conductive nanostructured composites for amperometric sensors; ii) development of new procedures for electroanalysis; iii) elaboration of novel chemometric tools for experimental design and most effective treatment of signal and data of interest in different fields of sensing.



## CAPTIONS TO TABLES AND FIGURES

Table 1 List of pollutants (interfering species) together with the relevant upper concentration limits.

Table 2 Air components, with corresponding typical and maximum concentration values.

Table 3 Definition of the number of mixtures of target molecules.

Table 4 Definition of the number of mixtures of pollutants.

Figure 1 Scheme of the EC-QCL-PAS instrument: 1) Laser power meter; 2) Photoacoustic cell; 3) Cantilever; 4) Balance cell; 5) Tubes for gas outlet and valves; 6) Tubes for gas inlet and valves; 7) Window; 8) Laser; 9) CMOS detector; 10) Rotating chopper wheel; 11) Chopper frequency read-out; 12) Rotating plane; 13) Grating; 14) Quantum Cascade Laser; 15) External Cavity –lenses; 16) QCL beam.

Figure 2 Denoising literature spectra: the upper plot reports a sample portion of benzene spectrum (black) together with the denoised signal (wavelet approximations, red), while the lower plot reports the extracted noise (wavelet details).

Figure 3 Regression coefficients  $b_0$  and  $b_1$ , relevant errors  $s(b_0)$  and  $s(b_1)$  and significance values  $p(b_0)$  and  $p(b_1)$ , as a function of the different number of intervals,  $n$ . For  $b_1$ ,  $s(b_1)$  and  $p(b_1)$ , the blue circles are referred to the values obtained including the intercept, while the red circles are referred to the values obtained setting  $b_0 = 0$ .

Figure 4 Spectra of the four target molecules at 1 ppm concentration, added with the experimental noise.

Figure 5 Mixture merging scheme

Figure 6 Effect of SWF on the multiplication of a 1 ppm spectrum of acetonitrile by 3 different concentration levels (1, 5, and 10): a) whole spectra; b) zoom on the Y axis. The dotted lines represent the results of the simple multiplication by a constant value (const), while the solid lines represent the spectra obtained using the SWF.

Figure 7 Comparison between simulated spectra (gray) and EC-QCL-PAS spectra (black) relative to: a) binary mixture of  $\text{NH}_3$  (1 ppm) and  $\text{CH}_3\text{OH}$  (1.6 ppm); b) quaternary mixture of  $\text{H}_2\text{O}$  (380 ppm),  $\text{CO}_2$  (10000 ppm),  $\text{CH}_3\text{OH}$  (11 ppm) and saffrole (11 ppm).

**Table 1**

<b>Molecule</b>	<b>Upper limit [ppm]</b>	<b>Molecule</b>	<b>Upper limit [ppm]</b>
Toluene	2.382	Propylene	0.01
Formaldehyde	0.4	Acetic acid	0.092
Ammonia	0.022	Ethylene glycol	0.491
Acrylonitrile	0.011	Naphthalene	0.071
Benzene	0.034	m-xylene	0.649
Ethanol	0.146	p-xylene	0.649
Methanol	0.016	o-xylene	0.016
Chloroform	0.038	Styrene	0.014
Ethylene	0.01	1,3-butadiene	0.005
Butane	0.033	Acrolein	0.011

**Table 2**

<b>Molecule</b>	<b>Typical concentration, TYP<sub>LIT</sub> [ppm]</b>	<b>Maximum concentration, MAX<sub>LIT</sub> [ppm]</b>
CH <sub>4</sub>	1.745	10
CO <sub>2</sub>	379	1000
CO	0.2	9
H <sub>2</sub> O	10000	60000
N <sub>2</sub> O	0.314	1
NO <sub>2</sub>	0.06	0.1
NO	0.02	0.1
SO <sub>2</sub>	0.005	0.03
O <sub>3</sub>	0.004	0.1

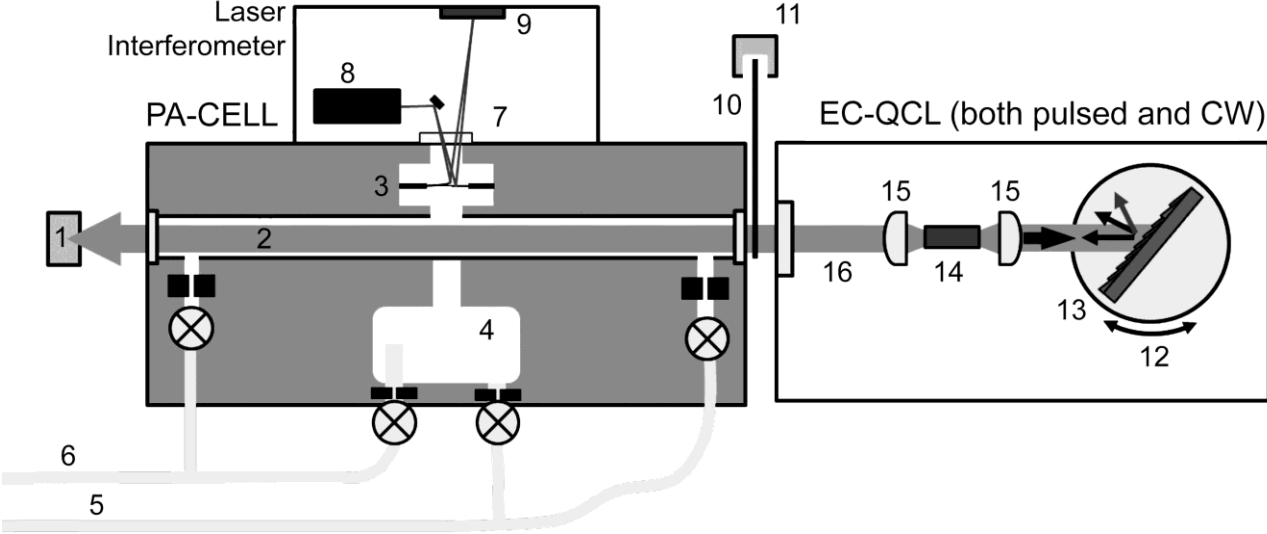
**Table 3**

<b>N° of target molecules per mixture (T)</b>	<b>Concentration levels (L)</b>	<b>N° of combinations of T targets (C)</b>	<b>Number of mixtures (N=C×L<sup>T</sup>)</b>
1	40	4	160
2	5	6	150
3	3	4	108
4	3	1	81

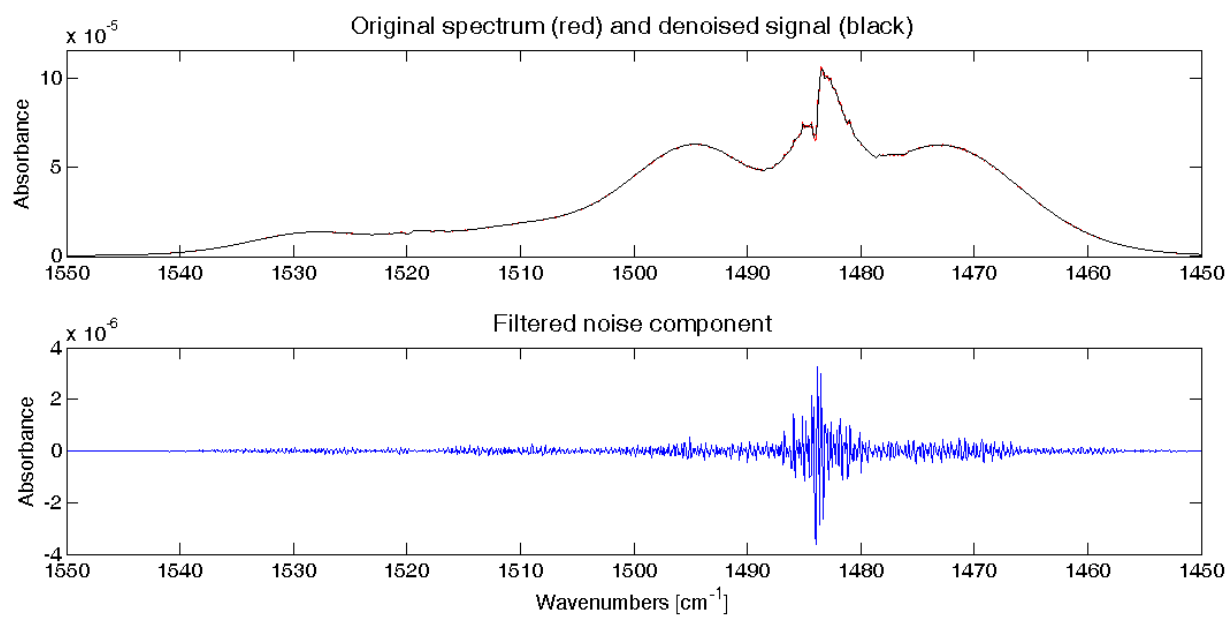
**Table 4**

<b>N° of Pollutants per mixture (P)</b>	<b>Concentration Levels (L)</b>	<b>N° of combinations of P pollutants (C)</b>	<b>Number of mixtures (N=C×L<sup>P</sup>)</b>
1	13	20	260
2	4	190	3040
3	3	1140	30780

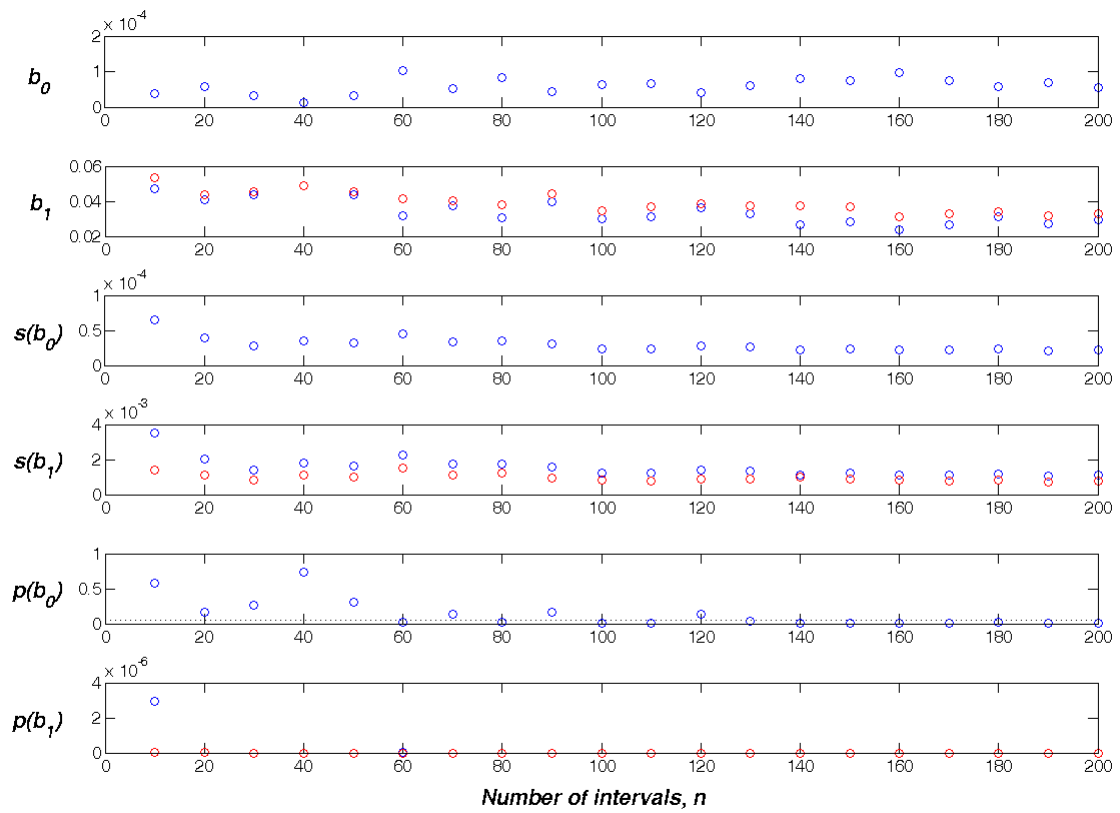
Figure 1



**Figure 2**

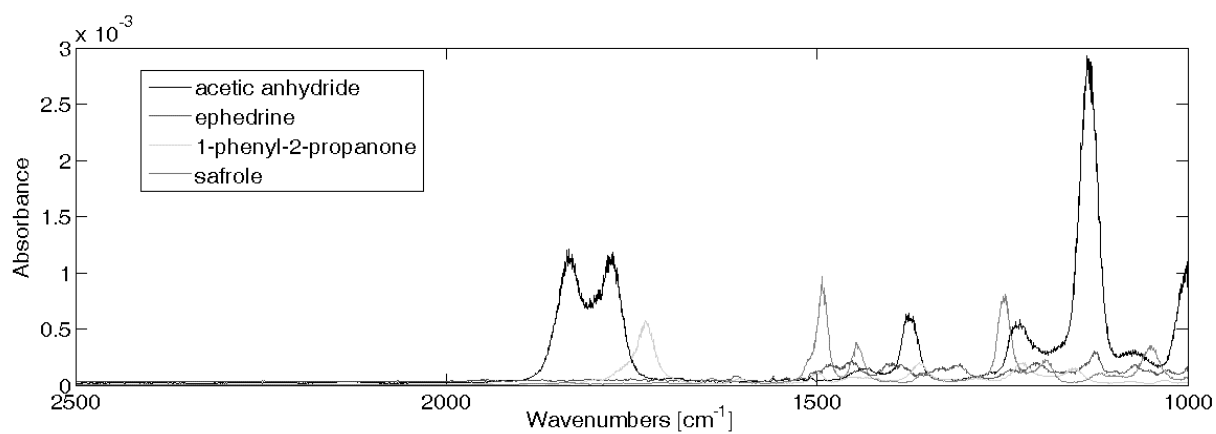


**Figure 3**

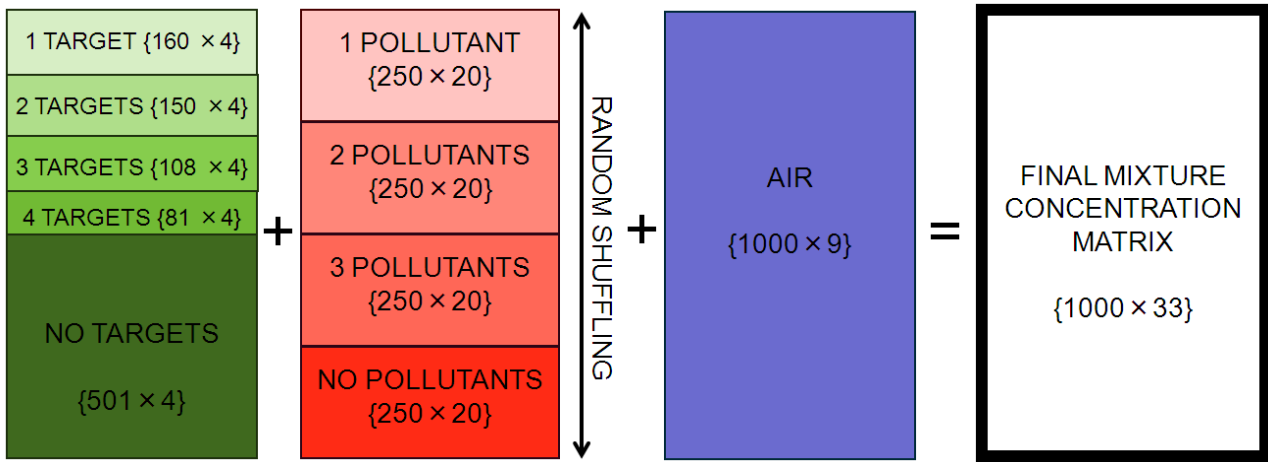




**Figure 4**



**Figure 5**



**Figure 6**

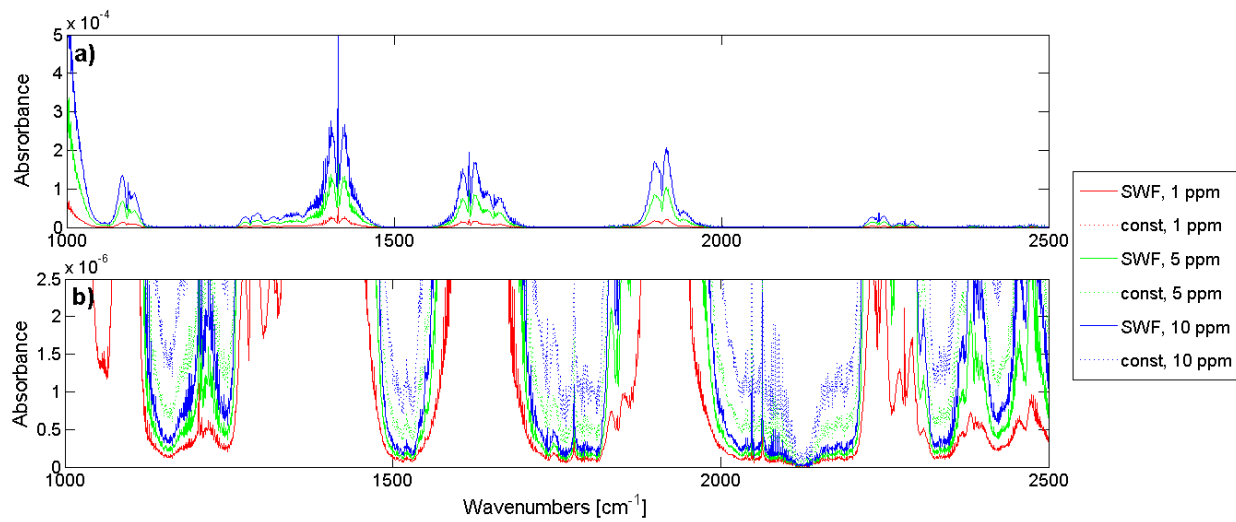
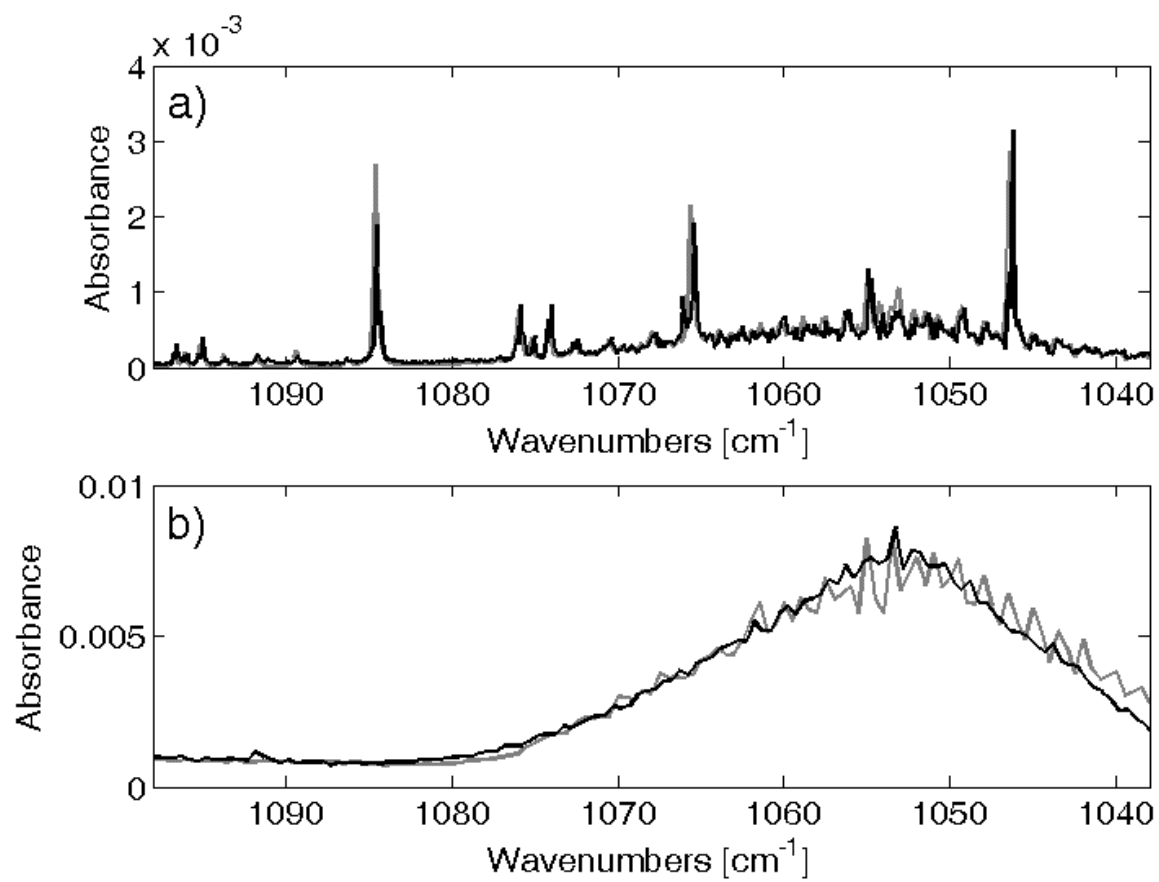


Figure 7



## SUPPLEMENTARY MATERIAL

Figure A1 Noise extraction from the spectrum of CH<sub>3</sub>OH experimentally measured with the continuous wave laser: a) original EC-QCL-PAS spectrum (black dotted line) and denoised spectrum (gray solid line); b) EC-QCL-PAS experimental noise. The best separation between instrumental noise and informative spectrum was gained using a sym2 wavelet at the first level of decomposition. The irregular shape of the signal reported in the lower plot clearly confirms that the high frequency content extracted by FWT actually corresponds to the noise component of the analyzed spectrum.

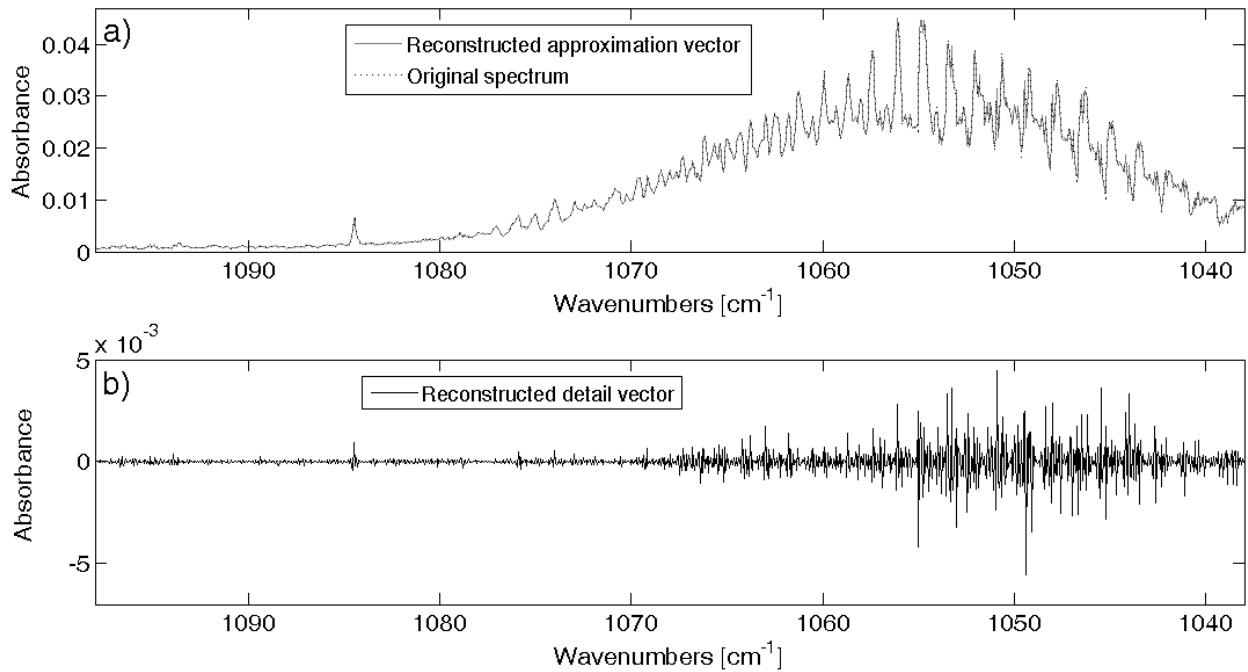


Figure A2 Estimation of the experimental LPAS noise: plots of the mean values of the intensity signal ( $I'_m$ ) as a function of the corresponding standard deviation values of the noise signal ( $N'_s$ ) calculated for the different number of intervals subdivisions,  $n$  (specified within each plot). The red lines correspond to the robust regression models calculated including the intercept, while the green lines are referred to the models with  $b_0 = 0$ .

