Experimental design-based strategy for the simulation of complex gaseous mixture spectra to detect drug precursors

Marco Calderisi* [a,b], Alessandro Ulrici[a,b], Laura Pigani[b,c], Alberto Secchi[d], Renato Seeber[b,c]
[a]Dipartimento di Scienze Agrarie e degli Alimenti, Università di Modena e Reggio Emilia, Padiglione Besta, Via Amendola 2, 42122 Reggio Emilia; [b] Consorzio INSTM, Via G. Giusti 9, 50121 Firenze, Italy; [c] Dipartimento di Chimica, Università di Modena e Reggio Emilia, Via Campi 183, 41125 Modena; [d]Selex-SI, Via Tiburtina, Km 12,400, 00131, Rome, Italy

ABSTRACT

The EU FP7 project CUSTOM (Drugs and Precursor Sensing by Complementing Low Cost Multiple Techniques) aims at developing a new sensing system for the detection of drug precursors in gaseous samples, which includes an External Cavity-Quantum Cascade Laser Photo-Acoustic Sensor (EC-QCLPAS) that is in the final step of realisation. Thus, a simulation based on FT-IR literature spectra has been accomplished, where the development of a proper strategy for the design of the composition of the environment, as much as possible realistic and representative of different scenarios, is of key importance. To this aim, an approach based on the combination of signal processing and experimental design techniques has been developed. The gaseous mixtures were built by adding the considered 4 drug precursor (target) species to the gases typically found in atmosphere, taking also into account possible interfering species. These last chemicals were selected considering custom environments (20 interfering chemical species), whose concentrations have been inferred from literature data. The spectra were first denoised by means of a Fast Wavelet Transform-based algorithm; then, a procedure based on a sigmoidal transfer function was developed to multiply the pure components spectra by the respective concentration values, in a way to correctly preserve background intensity and shape, and to operate only on the absorption bands. The noise structure of the EC-QCLPAS was studied using sample spectra measured with a prototype instrument, and added to the simulated mixtures. Finally a matrix containing 5000 simulated spectra of gaseous mixtures was built up.

Keywords: simulated spectra, experimental design, sigmoidal transfer function, denoising, Fast Wavelet Transform

1. INTRODUCTION

The increasing need of developing new miniaturised, low cost sensing systems, in order to be able to detect dangerous or illegal chemical substances that could be difficult to handle, often implies the need to work in absence of real experimental data, i.e. without possessing spectra measured with the newly developed instrumentation on real samples. Real data in fact, could be not available either due to the fact that the instrument it is not ready yet (at least in its final version), or because some of the chemical substances of interest are not available, due to their toxicity, hazards, or legal issues (e.g. in case of drugs or their precursors). Furthermore, the analysis of the expected spectra may be a necessary step in planning the detection system; this does hold in our case, with respect to the definition of the laser source. The challenging task, which is part of the EU FP7 project CUSTOM (Drugs and Precursor Sensing by Complementing Low Cost Multiple Techniques)[1], can be solved by following different strategies, depending on the specific problem to face. In this work we propose the procedure adopted by us for simulating spectra for the development of a new External Cavity-Quantum Cascade Laser Photo Acoustic Sensor (EC-QCLPAS)[2], to be used for the identification of vapours of drug precursors (targets).

In principle, in order to simulate the spectral response, it is possible to reconstruct them by means of physico-chemical simulations starting from the chemical structures, but this choice shows some drawbacks, mainly due to the approximations on which the calculations are based and to the noise issue, since a proper simulation should be suitable to represent not only the chemical information (ideal spectrum), but also the instrumental noise contribution to the final signal shape. Moreover, in case there is the need to simulate the spectra of many different substances, the excessive use of approximations could bring to spectral responses far from reality, not to mention the inherent computational load

m.calderis@chemiometria.it; http://www.dipsaa.unimore.it/eng/index.php?option=com_content&task=view&id=18&Itemid=27

of such an approach. In this case, since there was the need to account for a large number of chemical species mixed in varying amounts, a more feasible solution consisted in using spectra from literature databases. After proper processing, literature spectra of pure substances can be used as "building blocks" in order to simulate the spectral profile that would be obtained by the sensor under development. Of course, also this approach opens the gates to a lot of issues that must be addressed, before reaching the goal of a sufficiently realistic simulation. To this aim, a strategy based on various signal processing techniques has been developed, mainly involving preprocessing of spectral data and experimental design techniques for the simulation of the concentrations of the various gaseous species involved in the study.

The first step consists in the conversion of all the literature spectra of interest into a common format, since the literature spectra could have different resolutions, spectral ranges, and spectral domains with respect to the desired one. The spectra of the chemical species of interest are found, in fact, in different databases, which implies the need to work with different file types. Furthermore, the noise structure of the literature spectra is reasonably different from that of the spectra that will be obtained by the device under development. For this reason, it is mandatory to separate the useful spectroscopic information, that will be present also in the signal of the new instrumental device, from the noise contribution of the specific instrument used to measure a certain literature spectrum. This goal can be reached by using Wavelet Transform (WT)[3-5] -based signal processing techniques, thanks to the ability of WT to map the analysed signal both in the original and in the frequency domains at the same time, and to the possibility to obtain a wide number of different representations, by using different types of wavelet functions to decompose the signal into the WT domain.

In order to simulate a spectral dataset suitable to take into account the complex variability of gaseous mixtures composition, Experimental Design[6] techniques, along with an adequate randomization strategy, have been used. In this work more than 30 different gases have been considered including targets, pollutants, and typical air components, which required to develop a worked out process for the creation of the mixture concentration matrix, going through the separate elaboration of 3 different concentration matrixes, one for each class of components (targets, pollutants and air), that were subsequently merged. The final concentration matrix has been built by considering gas mixtures containing from 1 to 4 target species, 3 pollutants at most, and all the air components, at varying concentration levels.

When the denoised versions of the literature spectra and the matrix of the estimated concentrations of the different components considered are defined, they can be used to calculate the spectra profiles of the various gas mixtures. Even assuming that there is a simple linear relationship between the concentration of a given pure chemical species and the spectral intensity of its spectral bands, the simple multiplication of each unit concentration spectrum by the relevant concentration level within each gas mixture gives unrealistic results. In this case, in fact, also the background (which usually does not exhibit absorbance values exactly equal to zero) would vary according to the concentration of the chemicals. The correct signal shape would be therefore not preserved, since only the signal portions containing absorption bands should be taken into consideration in the multiplication operation. For this reason, proper nonlinear methods have been developed to multiply spectra of pure components by the desired concentration values.

Finally, the proper noise structure must be added, in order to obtain a realistic simulation of the output spectra that would be actually measured with the final device. In absence of experimental spectra acquired with the instrument under development, the only possible way to add the noise contribution consists in an approximate approach. In case that at least one spectrum has been acquired with a prototype version of the instrument, even if in a more restricted spectral range, assuming that the noise associated with this output is sufficiently similar to the noise of the spectra that will be measured with the final version of the instrument over the whole explored spectral range, it is possible to use these data for simulation of a realistic spectral response. Once the noisy component of the signal is extracted, it is possible to evaluate its frequency and amplitude and, in the case of heteroscedastic noise, to estimate also its dependence upon signal intensity. In this way, a proper noise contribution can be added to the denoised mixture spectra, in order to finally obtain a proper simulation of how the spectra acquired with the final version of the instrument would appear.

Briefly, the procedure proposed in this work in order to simulate in the most realistic and representative way the instrumental response can be summarized as follows:

- standardization of literature spectra imported from different literature databases, to obtain unit concentration (1 ppb) spectra over the desired spectral range, at the desired resolution;
- denoising of each single spectrum using WT, by choosing for each one of them the optimal conditions for wavelet decomposition (wavelet type and wavelet decomposition);

- assembling of the concentration matrix, by merging 3 different concentration matrices that were previously built for each gas class, i.e. targets, pollutants, and atmospheric components, separately;
- definition and optimisation of a sigmoidal transfer function for proper multiplication of the unit concentration spectra of each mixture component by the proper concentration level;
- estimate of the noise structure of sample spectra measured with a prototypal version of the EC-QCLPAS, by means of WT and of robust regression models to correlate noise amplitude with signal intensity;
- simulation of the final spectra by merging the denoised mixture spectra with the proper noise structure.

This approach allowed us to build a matrix containing the simulated EC-QCLPAS spectra of 5000 gas mixtures, which was further used[7] for the definition of the optimal working range and of the single wavenumbers where to perform the measurements with the final sensing device.

2. SELECTION OF CHEMICAL SPECIES

The first step of this work consisted in the selection of the interfering species (pollutants) and of the air components, to consider in addition to the chosen target molecules, i.e 1-phenyl-2-propanone, acetic anhydride, ephedrine, and safrole (common concentration range: 0.02 - 1 ppm). The pollutants list (Toluene, Propylene, Formaldehyde, Acetic acid, Ammonia, 2-Ethylene glycol, Acrylonitrile, Naphthalene, Benzene, m-Xylene, Ethanol, p-Xylene, Methanol, o-Xylene, Chloroform, Styrene, Ethylene, 1,3-Butadiene, Butane, Acrolein) was defined taking into account both the possible interference with target gases and the most likely environmental conditions. Their concentrations ranged from 1 ppb to maximum concentration values inferred from literature data[8]. The air components were CH_4, CO_2, CO, H_2O, N_2O, NO_2, NO, SO_2, O_3, and their typical concentration have been suitably chosen in order to give account for real situations

3. DATABASE BUILD UP

The spectra were modeled starting from FT-IR spectral data libraries PNNL and HITRAN. They have been previously imported at their original resolution, then they were transformed according to the expected prototype properties, in terms of sampling rate and spectral range. Finally, all the spectra were denoised by means of Fast Wavelet Transform (FWT).

3.1 Spectra importation

Spectra stored in different databases and coming from different sources may have different characteristics. For this reason, an algorithm to import spectral data with varying wavenumber ranges and resolutions has been developed, in order to obtain uniform datasets of standardized spectra at constant concentration (1 ppm). The importing algorithm is able to work with different input file format, namely txt and spc.

The algorithm keeps to the following working scheme:

1. importing spectra at full resolutions,
2. denoising,
3. converting smoothed spectra to the desired resolution, sampling rate and spectral window.

4. DENOISING

Database spectra have been denoised by means of a FWT-based algorithm developed *ad-hoc*, in order to consider only the relevant spectral information and to filter the not useful stochastic variation off.

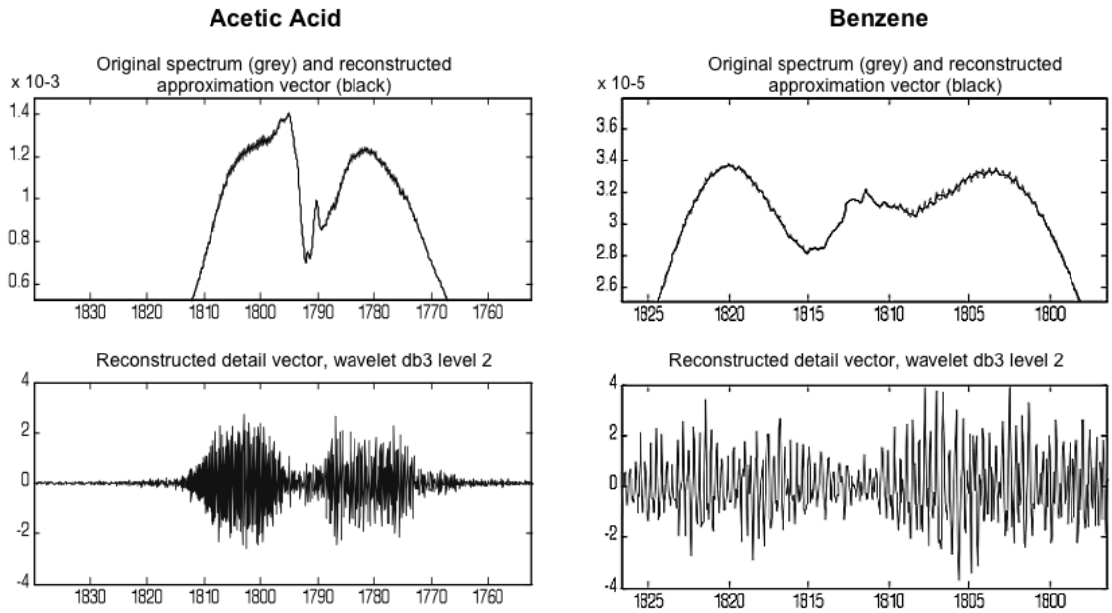

Figure 1 – Acetic Acid (left) and Benzene (right) denoising; detail of the signal.

The most part of the spectra were denoised using a Daubechies wavelet function (db3) at the 3rd level of decomposition. In Figure 1 portions of the Acetic Acid and of the Benzene spectra are reported, showing the original spectra, the denoised ones and the filtered noise components. As it can be seen, the noise intensity is somehow related to the intensity of the signal itself.

5. GAS MIXTURES CONCENTRATIONS

5.1 Concentration matrices

To build up the concentration matrices both Experimental Design[6] and randomization techniques have been used, in order to consider the complex variability of mixtures composition. The creation of the mixture concentration matrix is not straightforward at all, because of the high number of components, of the width of the concentration ranges, and of the different nature of the considered chemical species. For these reasons, as a first step, 3 separate concentration matrices, i.e. one for the targets, one for the pollutants, and one for the air components were built.

The concentration matrix of the target gases was built considering gas mixtures containing from 1 to 4 target species. In order to adopt a numerically balanced design, we used a 3 concentration levels Full Factorial Design (FFD) for mixtures with 3 and 4 targets (81 and 108 mixtures respectively), a 5 levels FFD for 2 targets mixtures (150 mixtures), and a 40 levels FFD for 1 target (160 mixtures), leading to a total of 499 mixtures of targets.

Concerning the interfering species, due to their high number and taking into account that the simultaneous presence of a high number of them would be unrealistic, mixtures from 1 to 3 species were considered. This led to the development of 3 FFDs with 13, 4 and 3 concentration levels for each combination of 1, 2, and 3 interfering species, respectively, leading to:

- 260 mixtures with 1 pollutant (at 13 concentrations levels)
- 3040 mixtures with 2 pollutants (at 4 concentrations levels)
- 30780 mixtures with 3 pollutants (at 3 concentrations levels)

for a total of 34080 mixtures. Of course, the use all these mixtures is not practical; therefore, a subsampling procedure was set up, which will be described in the following section.

Due to the wide concentration ranges, in order to have a higher number of low-concentration values for the mixtures components, the experimental designs with the lower number of concentration levels (from 1 to 5) were built using a non-uniform spacing for the concentration values.

A different planning scheme has been adopted for the air components, all the 9 components being always included. Furthermore, in view of the fact that the air components concentrations follow lognormal distributions, their concentration values were randomly selected after simulating this kind of distribution by means of an algorithm written *ad hoc*, able to take into account both the maximum and average concentration values of each component.

5.2 Final gas concentration mixture matrix

Concentration matrices relative to mixtures of target gases, pollutants, and air components have been merged together in order to obtain a final mixture matrix consisting of 1000 combinations, according to the following scheme (Figure 2):

- take all the target mixtures (499) and attach 501 rows of zeroes, in order to have a 1000 rows matrix
- randomly select 250 pollutants mixtures with one pollutant, 250 pollutants mixtures with 2 pollutants and 250 pollutants mixtures with 3 pollutants; 250 rows of zeroes are added, in order to a have a 1000 rows matrix
- create 1000 mixtures of air components as described previously
- merge the 3 matrices

This procedure was iterated 5 times, in order to build a final data matrix accounting for 5000 different mixtures including 4 target molecules, 9 air components and 20 pollutants in varying amounts.

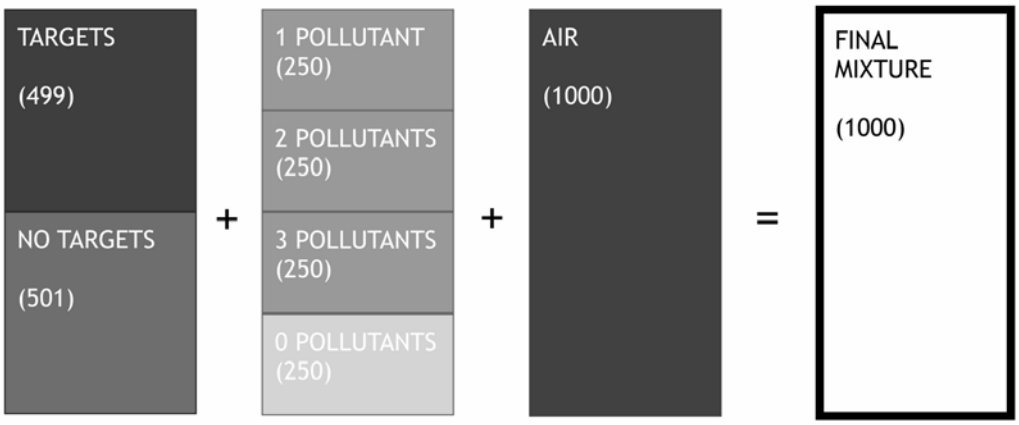

Figure 2 – Mixture merging scheme

6. ESTIMATION OF THE NOISE STRUCTURE

In order to estimate the noise structure, a spectrum of MeOH at 90 ppm concentration was used, acquired over the most significant spectral range by means of an ECQCL-PAS prototype instrument. This spectrum has been denoised using the same FWT-based procedure described before.

The noise structure was then defined following the procedure reported below:

- identify the optimal conditions, i.e. type of wavelet and decomposition level, to separate the intensity of the denoised signal (I) into the reconstructed approximation vector, and the corresponding noise signal (N) into the reconstructed detail vector
- sort I in ascending order and sort the noise signal N accordingly
- subdivide sorted I and sorted N in n intervals
- calculate the mean of the intensity signal, $\bar{I}$, and the standard deviation of N, N_{sd}, for each interval
- perform robust linear regression of N_{sd} as a function of $\bar{I}$, both with and without the intercept $b0$
- repeat the procedure by changing the number of intervals (from 10 to 200, step by 10)
- calculate $b0$ (intercept) and $b1$ (slope), and estimate their statistical significance

Considering the values of the regression coefficients $b0$ and $b1$, of their errors ($sb0$ and $sb1$, respectively), and of their significance [$p(b0)$ and $p(b1)$, respectively] as a function of the different interval subdivisions, it was possible to evidence that the intercept values were often not significantly different from zero, and that the slope values of the models calculated including $b0$ were very similar to those calculated with zero intercept. This led us to use a final noise model of the type

$$N_{sd} = b_1 \times \bar{I} \quad (1)$$

where the average of the b_1 values calculated for the different intervals was kept. The noise structure was then applied to the simulated spectra of gas matrices.

7. SPECTRA MULTIPLICATION

7.1 Sigmoidal transfer function

Given the matrix of concentrations and the matrix of the pure components spectra at unit concentration, the use of the simple matrix product to obtain the matrix of mixtures spectra leads to incorrect results, since also the baseline is incorrectly multiplied by the concentration value. For this reason an algorithm based on a sigmoidal transfer function was developed in order to multiply correctly spectra by concentration matrices, preserving the background intensity and its shape, and operating only on the actual absorption bands. The sigmoidal transfer function (Figure 3) needs to be suitably tuned to weight correctly the signal intensity.

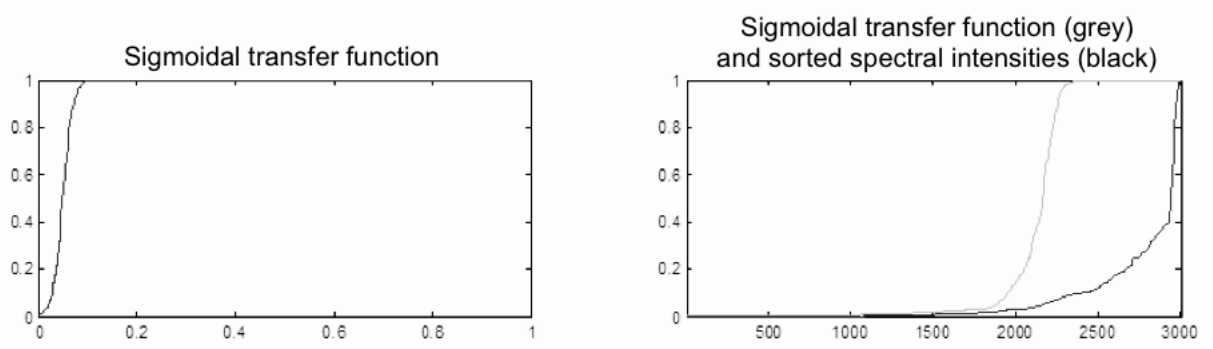

Figure 3- Sigmoidal transfer function for proper spectra multiplication

In Figure 4 the difference resulting from linearly multiplication of the whole spectra by a concentration factor and use of the appropriate sigmoidal transfer function is well evident.

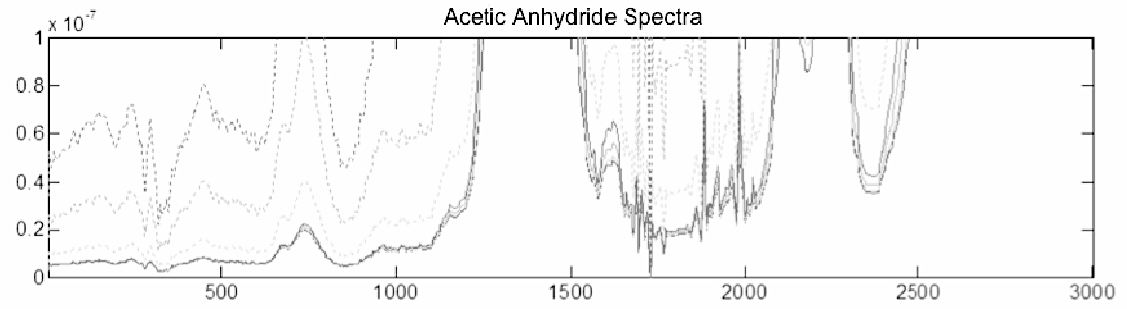

Figure 4- Effect of the sigmoidal transfer function for the multiplication of a unit concentration spectrum by 3 different concentration levels (2, 5, and 10) : a zoom of the low intensity (baseline) part of the spectrum. Dotted lines represent the results of the simple matrix product, solid lines the sigmoidal transfer function results.

7.2 Final simulated spectra data matrix

In order to build the final matrix of gas mixtures spectra, the last step consists in multiplying the final concentration mixture matrix by the database spectra, using the sigmoidal transfer function, and finally adding directly the estimated noise. The final gas mixture matrix contains all the gases (targets, pollutants, and air components) at the chosen concentration levels, in the suitable mixing proportions.

The final dataset of 5000 spectra of different gas mixtures includes mixtures with only air components, mixtures with air and target molecules, mixtures with air and pollutants, and mixtures with targets, pollutants, and air. It must be highlighted that the described procedure can be repeated to increase the number of simulated mixtures or to generate new datasets, whenever necessary. In the step described previously, in fact, pollutants and air components mixtures are randomly merged from a starting population that is newly defined each time the algorithm is used.

REFERENCES

[1] Uotila, J., Lehtinen, J., Kuusela, T., Sinisalo, S., Maisons,G., Terzi, F., Tittonen, J., "Drug precursor vapor phase sensing by cantilever enhanced photoacoustic spectroscopy and quantum cascade laser", Proc. SPIE 8545-8 (2012).

[2] Secchi, A., Fiorello, A. M., D'Auria, S., Varriale, A., Ulrici, A., Seeber, R., Uotila, J., Venditto, V., Estensoro, P., Colao, F., "Drugs and precursor sensing by complementing low cost multiple techniques: Overview of the European FP7 Project CUSTOM", Proc. SPIE 8545-17 (2012).

[3] Walczak, B. (ed.), [Wavelets in Chemistry], Elsevier, Amsterdam (2000).

[4] Cocchi, M., Seeber, R., Ulrici, A., "WPTER: Wavelet Packet Transform for Efficient pattern Recognition of signals", Chemom. Intell Lab. Syst. 57(2), 97-119 (2001).

[5] Cocchi, M., Seeber, R., Ulrici, A., "Multivariate calibration of analytical signals by WILMA (Wavelet Interface to Linear Modelling Analysis)", J. Chemometrics 17 (8-9), 512-527 (2003).

[6] Box, G., Hunter, J., Hunter, W., [Statistics for Experimenters: Design, Innovation, and Discovery], Wiley, New Jersey (2005).

[7] Ulrici, A., Seeber, R., Calderisi, M., Foca, G., Uotila, J., Carras, M., Fiorello, A. M., "A feature selection strategy for the analysis of spectra from a photoacoustic sensing system", Proc. SPIE 8545-18 (2012).

[8] Dunayevskiy, I., Tsekoun, A., Prasanna, M., Go, R., Patel, C.K.N., “High-sensitivity detection of triacetone triperoxide (TATP) and its precursor acetone”, Appl. Optics 46 (25), 6397-6404 (2007).