Estimation of the Number of Volatile Compounds in Simple Mixtures †

Olivier Hotel *, Jean-Philippe Poli, Christine Mer-Calfati, Emmanuel Scorsone and Samuel Saada

CEA, LIST, 91191 Gif-sur-Yvette, France; jean-philippe.poli@cea.fr (J.-P.P.); christine.mer@cea.fr (C.M.-C.); emmanuel.scorsone@cea.fr (E.S.); samuel.saada@cea.fr (S.S.)

* Correspondence: olivier.hotel@cea.fr

Published: 11 August 2017

Abstract: In this paper, we address the problem of simple mixtures identification using a SAW sensor based e-nose. First, we propose a linearized form of the SAW sensor’s response to mixtures. This equation is then used to set up a linear least squares problem whose residuals are suitable to identify the number of compounds present in a mixture thanks to a supervised learning algorithm.

Keywords: SAW sensor; mixture identification; pattern recognition

1. Introduction

In this paper, we propose an approach to tackle the problem of simple mixture identification using an array of SAW sensors. These sensors are based on the propagation of mechanical waves produced by piezoelectric materials along a layer coated with chemically interactive materials. Volatile compounds are absorbed onto the surface of the sensitive material, changing its properties and yielding to a measurable frequency shift. In [1], authors established that the frequency shift is the superposition of a mass loading effect and of a viscoelastic contribution. Both can be modelled by first order linear differential equations

\[ \tau_m \frac{d F_i}{dt} + F_i = K_i c(t), \quad i \in \{m,v\} \]

where \( F_m \) and \( F_v \) are respectively the frequency shift due to the mass loading effect and to the viscoelastic contribution, \( \tau_m, \tau_v, K_m \) and \( K_v \) are the time constants and the gains of the contribution and \( c(t) \) is the concentration profile which is not known. The total frequency shift is then given by \( F = F_m + F_v \).

One can show, by discretizing the differential equations and solving the recurrence equations, that the expression of \( F[n] \) can be written as

\[ F[n] = \sum_{i=1}^{N} \gamma_i F_i[n] = \sum_{i=1}^{N} \gamma_i A_m T_{m,i}^n + \gamma_i A_v T_{v,i}^n + P_i[n] \]

(1)

This assumption can be verified experimentally. The selected e-nose was composed of 6 functionalized diamond coated SAW sensors. The sensors were exposed to \( \text{SO}_2, \text{H}_2\text{S}, \text{CH}_3\text{OH}, \text{C}_2\text{H}_6, \text{NH}_3 + \text{H}_2\text{S}, \text{H}_2\text{S} + \text{CH}_3\text{OH} \) and to the ternary mixtures \( \text{H}_2\text{S} + \text{NH}_3 + \text{SO}_2, \text{H}_2\text{S} + \text{NH}_3 + \text{CH}_3\text{OH}, \text{H}_2\text{S} + \text{CH}_3\text{OH} \) and to the ternary mixtures \( \text{H}_2\text{S} + \text{NH}_3 + \text{CH}_3\text{OH}, \text{H}_2\text{S} + \text{CH}_3\text{OH} \) . Several cycles exposition—purge were done. The assumption was tested by solving a least squares optimization problem to estimate the coefficients \( \gamma \), and by measuring the relative error between the true mixture signals and the one reconstructed from the pure signals. Tables 1 and 2 show the average relative error.
be easily be proved by induction that the term is a polynomial of degree \( d \), \( P_k(E) \) the set of subsets of \( E \) of cardinality equal to \( k \). \( S \) the cumulative sum and \( S^k_a = \left\{ \begin{array}{ll} a & \text{if } k = 0 \\ S(S^k_{a-1}) & \text{otherwise} \end{array} \right. \) is the \( k \)th iterated cumulative sum of the sequence \( a \).

2. Estimation of the Number of Simple Mixture’s Components

2.1. Notation and Definitions

In the rest of this paper, \( P^d[.] \) denotes a polynomial of degree \( d \). \( S^k_a \) the set of subsets of \( E \) of cardinality equal to \( k \). \( S \) the cumulative sum and \( S^k_a = \left\{ \begin{array}{ll} a & \text{if } k = 0 \\ S(S^k_{a-1}) & \text{otherwise} \end{array} \right. \) is the \( k \)th iterated cumulative sum of the sequence \( a \).

2.2. Linearisation of the Sensor’s Response

Prior to propose a linear form for the sensor’s response, we remind that a direct consequence of the Faulhaber’s formula allows to conclude that the \( k \)th iterated cumulative sum of a sequence whose general term is a polynomial of degree \( d \) is a polynomial of degree \( d + k \). Moreover, it can be easily be prooved by induction that the \( k \)th iterated cumulative sum of a sequence whose general term is a sum of geometric series \( g[n] = \sum_i a_i r^n_i \) can be written in the form \( S^k_	heta = \sum_i a_i \left( \frac{r_i^k - 1}{r_i - 1} \right) r^n_i + p^{k-1}[n] \).

Equation (1) can be rewritten \( F[n] = \sum_{i=1}^{2N_g} A_i T_i^n + P[n] \). For the sake of clarity, we note \( \mathcal{X} \) the set \( \{1, \ldots, 2N_g\} \). Let’s compute \( F[n] + \sum_{i=1}^{2N_g} (-1)^j S^i_{\mathcal{X}}[n] \sum_{j\in \mathcal{E}(\mathcal{X})\setminus \emptyset} \prod_{i\in j} \frac{T_j - 1}{T_j} \). The previously described results and the fact that the iterated cumulative sum operator is linear yield to:

\[
F[n] + \sum_{i=1}^{2N_g} (-1)^j S^i_{\mathcal{X}}[n] \sum_{i\in \mathcal{E}(\mathcal{X})\setminus \emptyset} \prod_{j\in i} \frac{T_j - 1}{T_j} = F[n] + \sum_{k=1}^{2N_g} (-1)^k \sum_{i=1}^{2N_g} A_k \left( \frac{T_k - 1}{T_k} \right) T_k^n \sum_{i\in \mathcal{E}(\mathcal{X})\setminus \emptyset} \prod_{j\in i} \frac{T_j - 1}{T_j} + \sum_{i=1}^{2N_g} (-1)^i S^i_{\mathcal{X}}[n] \sum_{j\in \mathcal{E}(\mathcal{X})\setminus \emptyset} \prod_{i\in j} \frac{T_j - 1}{T_j}.
\]

First, one can notice that the term \( \sum_{i=1}^{2N_g} (-1)^i p^{i-1}[n] \sum_{j\in \mathcal{E}(\mathcal{X})\setminus \emptyset} \prod_{i\in j} \frac{T_j - 1}{T_j} \) is a weighted sum of polynomials of degree up to \( 2N_g - 1 \) so it is a polynomial \( P^{2N_g-1} \) of degree \( 2N_g - 1 \). Secondly, the term \( \sum_{i=1}^{2N_g} (-1)^i S^i_{\mathcal{X}}[n] \sum_{j\in \mathcal{E}(\mathcal{X})\setminus \emptyset} \prod_{i\in j} \frac{T_j - 1}{T_j} \) is a sequence which will be considered as an unknown in the next section. For the sake of clarity, this term is denoted as \( P_i[n] \) in the rest of this article. Finally, one can remark that the term \( \sum_{i=1}^{2N_g} (-1)^i \sum_{k=1}^{2N_g} A_k \left( \frac{T_k - 1}{T_k} \right) T_k^n \sum_{j\in \mathcal{E}(\mathcal{X})\setminus \emptyset} \prod_{i\in j} \frac{T_j - 1}{T_j} \) is equal to 0. Indeed, this term can be simplified by remarking that the sum indexed by the set \( i \) can be split into two parts:

### Table 1. Binary mixture reconstruction error.

<table>
<thead>
<tr>
<th>Mixture</th>
<th>( SO_2 + C_7H_8 )</th>
<th>( NH_3 + CH_3OH )</th>
<th>( H_3 + C_7H_8 )</th>
<th>( H_2S + NH_3 )</th>
<th>( H_2S + CH_3OH )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative error</td>
<td>23.01%</td>
<td>10.89%</td>
<td>19.9%</td>
<td>11.47%</td>
<td>12.84%</td>
</tr>
</tbody>
</table>

### Table 2. Ternary mixture reconstruction error.

<table>
<thead>
<tr>
<th>Mixture</th>
<th>( H_2S + NH_3 + SO_2 )</th>
<th>( H_2S + NH_3 + CH_3OH )</th>
<th>( H_2S + C_7H_8 + SO_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative error</td>
<td>24.6%</td>
<td>55.62%</td>
<td>37.76%</td>
</tr>
</tbody>
</table>

Table 1 shows the assumption is plausible for the binary mixtures since the average error is near the variance of the successive experiments and that the assumption does not hold for the ternary mixtures. However, this is still informative since the objective consists in estimating the number \( N_g \) of compounds in simple mixture.
estimated by solving the optimization problem to grow unbounded, and the matrix variables is greater than the number of equations (eq = eq). Consequently, we have:

\[ \sum_{i=1}^{\text{eq}} (-1)^i A_k \left( \frac{T_k}{T_k - 1} \right)^i T_k^n \sum_{j \in I} \prod_{j \in I} T_j - 1 T_j = - \sum_{i=1}^{\text{eq}} A_k \left( \frac{T_k}{T_k - 1} \right)^i T_k^n \sum_{j \in I} \prod_{j \in I} T_j - 1 T_j + \sum_{i=1}^{\text{eq}} A_k \left( \frac{T_k}{T_k - 1} \right)^i T_k^n \sum_{j \in I} \prod_{j \in I} T_j - 1 T_j \]

Moreover, we have:

\[ \sum_{i=2}^{2N_g} (-1)^i \sum_{k=1}^{2N_g} A_k \left( \frac{T_k}{T_k - 1} \right)^i T_k^n \sum_{j \in I} \prod_{j \in I} T_j - 1 T_j = \sum_{i=1}^{\text{eq}} A_k \left( \frac{T_k}{T_k - 1} \right)^i T_k^n \sum_{j \in I} \prod_{j \in I} T_j - 1 T_j \]

since the series are telescoping. Hence, we have \( \sum_{i=1}^{\text{eq}} (-1)^i \sum_{k=1}^{2N_g} A_k \left( \frac{T_k}{T_k - 1} \right)^i T_k^n \sum_{j \in I} \prod_{j \in I} T_j - 1 T_j = 0 \).

Hence, by posing \( \alpha = (-1)^i \sum_{k=1}^{2N_g} A_k \left( \frac{T_k}{T_k - 1} \right)^i T_k^n \sum_{j \in I} \prod_{j \in I} T_j - 1 T_j \), Equation (1) can be rewritten in the linear form:

\[ F[n] = \sum_{i=1}^{2N_g} \alpha_i S_i^F[n] + P^{2N_g-1}[n] + P_n[n] \quad (2) \]

2.3. Optimization Problem Formulation

In this section, we assume that N samples \( F[0], \ldots, F[N-1] \) were digitized. We define the vector \( F \) and \( C \) of respective length \( N \) and \( 4N_g + N : F = [F[0], \ldots, F[N-1]]^T \), \( C = [\alpha_1, \ldots, \alpha_{2N_g}, P_0, \ldots, P_{2N_g-1}, P_0[0], \ldots, P_0[N-1]]^T \). And the \( (N) \times (2N_g) \) matrices \( S \) and \( P : S = [S_0^F[0], \ldots, S_0^F[N-1], \ldots, S_{2N_g}^F[0], \ldots, S_{2N_g}^F[N-1]] \) and \( P = [1, 0^T, \ldots, (N-1), (N-1)^{2N_g-1}] \) and we form the matrix \( X = [S, P, I_\alpha] \). With these definitions, Equation (2) can be rewritten as \( F = XC \). As the number of variables is greater than the number of equations \( (4N_g + N) \text{ vs. } N \) equations), we should add a regularization term to avoid overfitting. We define the matrix \( \Gamma = [I_{4N_g}, 0, 0_{4N_g,N}, D] \) where \( D \) is the second order differencing matrix. The identity matrix prevents the coefficients \( \alpha \) and \( p \) to grow unbounded, and the matrix \( D \) smoothes the unknown sequence \( P_s \). The parameters can be estimated by solving the optimization problem \( \hat{C} = \arg \min ||F - XC||^2 + \lambda ||\Gamma C||^2 \) whose solution is
The residuals are defined as \( R = F - X\hat{\mathbf{c}} \) and the mean squared error (MSE) as \( MSE = \frac{1}{n} R^T R \).

2.4. Experimental Results

In this section, we propose an approach, based on supervised learning, to estimate the number of compounds in simple mixtures. In particular, we compare the performances of three different sets of features. The first set is composed of the steady state amplitude of the response of each sensor whereas the second set is composed of the RMS error obtained from the least squares optimization problem for \( N_g = 1, 2, 3 \) and for each sensor. The third set is the concatenation of these two sets. Table 3 gives the results obtained over a 5-fold cross validation process using the LMNN algorithm [3]. It exhibits that the proposed features are only outperformed by the steady state when \( N_g = 3 \). It also shows the interest of using both features since in this case the performances are increased.

<table>
<thead>
<tr>
<th>Predicted ( N_g )</th>
<th>Predicted ( N_g )</th>
<th>Predicted ( N_g )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Actual ( N_g )</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>90.4%</td>
<td>8.1%</td>
</tr>
<tr>
<td></td>
<td>93.1%</td>
<td>6.9%</td>
</tr>
<tr>
<td></td>
<td>96.3%</td>
<td>3.2%</td>
</tr>
<tr>
<td></td>
<td>25.6%</td>
<td>72.0%</td>
</tr>
<tr>
<td></td>
<td>19.7%</td>
<td>76.7%</td>
</tr>
<tr>
<td></td>
<td>6.7%</td>
<td>92.8%</td>
</tr>
<tr>
<td></td>
<td>9.0%</td>
<td>9.0%</td>
</tr>
<tr>
<td></td>
<td>8.1%</td>
<td>11.1%</td>
</tr>
<tr>
<td></td>
<td>3.0%</td>
<td>3.0%</td>
</tr>
</tbody>
</table>

3. Conclusions

In this paper, we established that the response of SAW sensors to a simple mixture of volatile compounds can be linearized thank to iterated cumulative sums. This linear equation allows the formulation of a regularized linear least squares problem whose residuals, in conjunction with the steady state amplitudes of the signals and supervised learning algorithms, are suitable to estimate the number of compounds in a mixture (the estimation accuracy is higher than 90%).

Acknowledgments: This research was supported by the French trans-governmental CBRN-E R & D program.

Conflicts of Interest: The authors declare no conflict of interest.

References