Estimating Gas Concentration using a Microcantilever-Based Electronic Nose

John Leis [a,*]

Weichang Zhao [b,c]
Lal A. Pinnaduwage [b,d]
Anthony C. Gehl [b]
Steve L. Allman [b]
Allan Shepp [c]
Ken K. Mahmud [c]

[a] *Department of Electrical, Electronic, and Computer Engineering, University of Southern Queensland, Toowoomba, Qld 4350, Australia*

[b] *Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6122*

[c] *Triton Systems, Inc., 200 Turnpike Road, Chelmsford, MA 01824*

[d] *Department of Physics, University of Tennessee, Knoxville, TN 37996*

Abstract

This paper investigates the determination of the concentration of a chemical vapor as a function of several nonspecific microcantilever array sensors. The nerve agent dimethyl methyl phosphonate (DMMP) in parts-per-billion concentrations in binary and ternary mixtures is able to be resolved when present in a mixture containing parts-per-million concentrations of water and ethanol. The goal is to not only detect the presence of DMMP, but additionally to map the nonspecific output of the sensor array onto a concentration scale. We investigate both linear and nonlinear approaches — the linear approach uses a separate least-squares model for each component, and a nonlinear approach which estimates the component concentrations in parallel. Application of both models to experimental data indicate that both models are able to produce bounded estimates of concentration, but that the outlier performance favors the linear model. The linear model is better suited to portable handheld analyzer, where processing and memory resources are constrained.

Key words: Pattern recognition, Information fusion

* Corresponding author.
Email address: `leis@usq.edu.au` (John Leis).

1 Introduction

The biological nose – and in particular, the canine nose – is a highly sensitive detector which is unrivalled its overall performance. Mass spectrometry is able to detect very small vapor concentrations, but such equipment is awkward, bulky and usually requires special operator skill and/or calibration. Emerging applications such as the war on terror, land mine detection, food product analysis and medical applications demand a smaller device which is able to quickly detect minute concentrations of chemical vapor.

Our electronic nose approach, using microcantilever sensor array and an artificial neural network for estimation of the concentrations, was initially reported in [5]. This paper further investigates the algorithmic aspects of the training and estimation phases, since the goal is to produce a reliable hand-held device which, necessarily, is constrained in processing and memory resources. In recent work, we have used selective cantilevers (sorbents on the cantilever surfaces) that make the cantilevers selective to explosive vapor molecules. Our detection of explosive vapors has been enhanced by cantilever selectivity, followed by pattern recognition techniques described here. To the best of our knowledge, this is the first successful application of such techniques to estimate three-component mixtures of such vapors.

The experimental setup is only briefly described here; more extensive details on the experimental procedures may be found in [5] and [6].

2 Sensor and Signal Characterization

We employ a microcantilever array which produces nonspecific and reversible response vectors to vapors to which the sensor array is exposed. The nonspecific nature of these sensors is both their strength and weakness [5,6]. They have a short recovery time, which means the measurement can be repeated after a short interval. However, since they are nonspecific, the exact pattern of response is not correlated to the analyte concentration in any obvious way. Thus, we have a signal processing challenge to determine whether such a mapping from sensor response to target concentration is feasible, repeatable and sufficiently accurate.

The sensor array consists of four Canti-4 piezoresistive microcantilever chips [1].

Each chip has two coated and two uncoated (reference) microcantilever sensors. Seven sensors were utilized for the results presented herein.
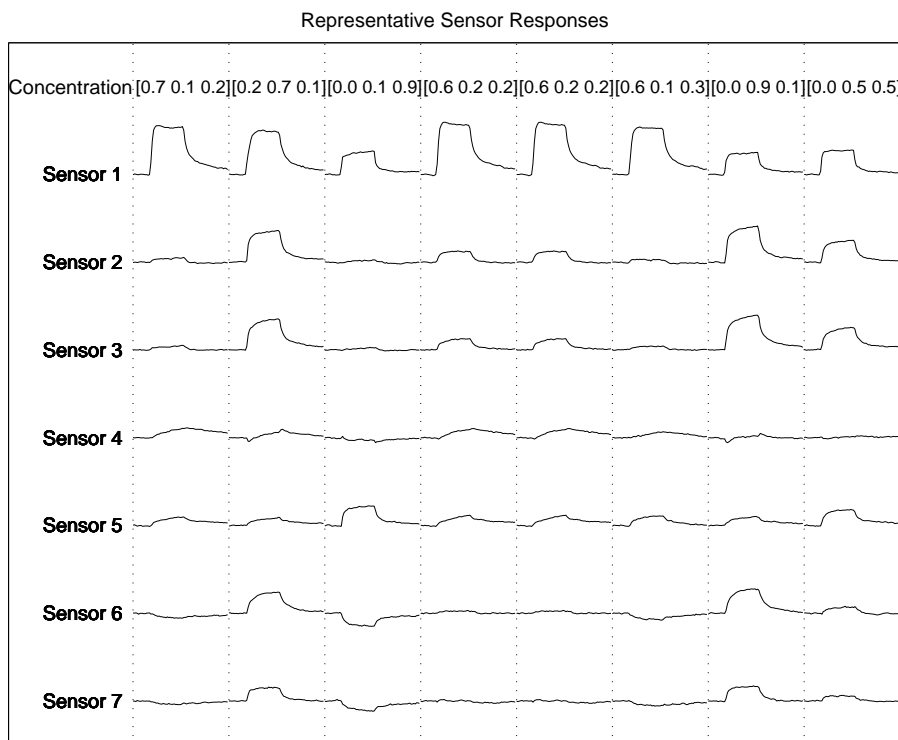
---

[1] http://www.cantion.com/

Fig. 1. Representative sensor response profiles at various concentration mixtures for the 7-sensor array. The relative concentration of the 3 vapors are shown. The data is for DMMP, water and ethanol mixtures with 100 Parts per Billion (PPB), 60 Parts per Million (PPM) and 60 PPM respectively (maximum concentration) for each individual vapor. Each pulse segment represents 10 seconds.

Representative sensor responses are shown in Figure 1. Clearly, some sensors responded similarly, and some differently, to each vapor mixture.

The approach taken is necessarily one of "training" the algorithm to detect particular patterns in a given input analyte vapor. Thus, the electronic nose approach is entirely different to the analytical chemistry approach where the analytes of a mixture are normally first separated and subsequently identified and quantified by comparison with analytical standards.

The linearity of the sensors is not a given; furthermore, the transient nature of the sensor responses may also yield useful classification and quantification information.

Representative response clusters are shown in Figure 2. This shows the potential for discrimination of the presence or absence of certain chemicals, and gives motivation to investigate the possibility of determining the concentrations present.

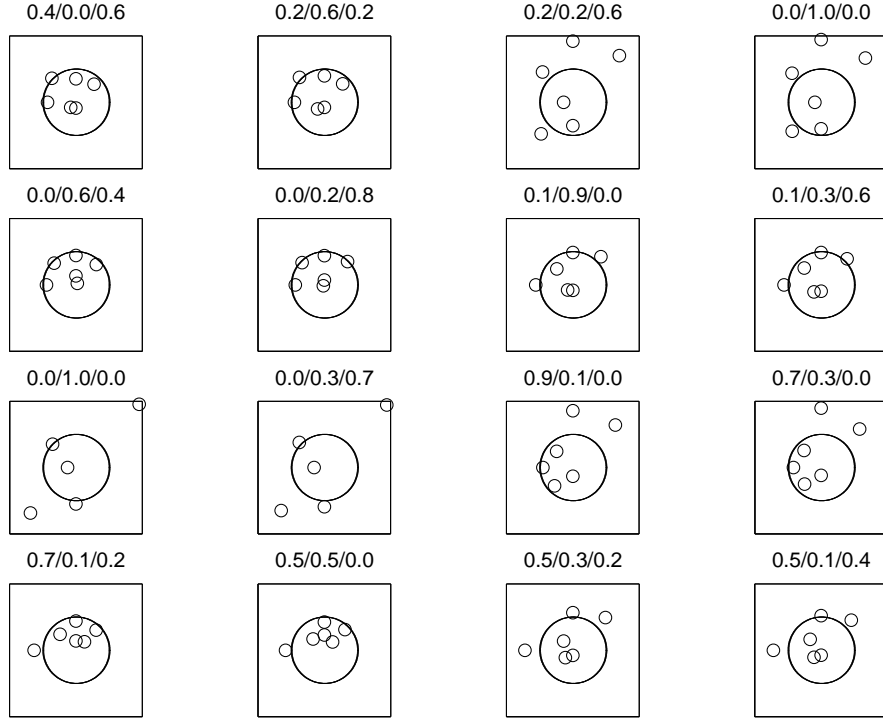Representative Sensor Responses for various input concentrations

Fig. 2. Representative sensor responses shown on radial axes. The magnitude of each sensor response is shown at a fixed angle. Clearly, different input mixtures produce different responses, and there is no obvious relationship between mixture concentration and sensor output.

3 Linear Estimation

The first 3 eigenvalues of each autocorrelation matrix revealed that there was significant correlation between the sensors. For the sensor array, we define each component i and develop a linear model from the experimental observations The output concentration y_c (where c is the index of the particular gas component) is assumed to be linearly related to sensor steady-state responses x_k.

80% of the available vectors were used for estimating the covariance matrix, with the remaining 20% used for testing, so as to not bias the result in any way. For N experiments and M sensors, we can formulate the linear response for each concentration $y_{c,n}$ using a linear model as

$$y_{c,1} = \hat{a}_{c,0} + \hat{a}_{c,1}x_{1,1} + \cdots + \hat{a}_{c,M}x_{1,M}$$
$$y_{c,2} = \hat{a}_{c,0} + \hat{a}_{c,1}x_{1,2} + \cdots + \hat{a}_{c,M}x_{2,M}$$
$$\vdots$$
$$y_{c,N} = \hat{a}_{c,0} + \hat{a}_{c,1}x_{N,2} + \cdots + \hat{a}_{c,M}x_{N,M}$$

We can write this as

$$\begin{pmatrix} y_{c,1} \\ y_{c,2} \\ \vdots \\ y_{c,N} \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,M} \\ 1 & x_{2,1} & \cdots & x_{2,M} \\ \vdots & & & \\ 1 & x_{N,1} & \cdots & x_{N,M} \end{pmatrix} \begin{pmatrix} \hat{a}_{c,0} \\ \hat{a}_{c,1} \\ \vdots \\ \hat{a}_{c,M} \end{pmatrix} \tag{1}$$

or

$$\mathbf{y}_c = \mathbf{X}\mathbf{p}_c \tag{2}$$

where the output concentration for component c is given by

$$\mathbf{y}_c = \begin{pmatrix} y_{c,1} \\ y_{c,2} \\ \vdots \\ y_{c,N} \end{pmatrix} \tag{3}$$

with sensor array response vector $\left(1 \mid \mathbf{x}_n^T\right)$ formed into the matrix $\mathbf{X}$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & & & & \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{pmatrix} \tag{4}$$

The parameter vector weights are thus

$$\mathbf{p}_c = \begin{pmatrix} \hat{a}_{c,1} \\ \hat{a}_{c,2} \\ \vdots \\ \hat{a}_{c,M} \end{pmatrix} \tag{5}$$

We seek the optimal solution to

$$||\mathbf{y}_c - \mathbf{X}_c \mathbf{p}_c|| \tag{6}$$

to yield the predictor parameters

$$\mathbf{p}_c^* = \begin{pmatrix} \hat{a}_{c,0} \\ \hat{a}_{c,1} \\ \vdots \\ \hat{a}_{c,M} \end{pmatrix} \tag{7}$$

The optimal solution which minimizes the least-squares distance between $\mathbf{y}$ and $\mathbf{Xp}$ is[2]

$$\widehat{\mathbf{p}}_c^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}_c \tag{8}$$

Empirical distributions for the predictor parameters can be established using the bootstrap estimate, as follows [7]. The predictor $\hat{\mathbf{p}}$ is estimated using the pseudoinverse as before. Then the estimated output and error are calculated as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{p}} \tag{9}$$
$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} \tag{10}$$

We then resample $\hat{\mathbf{e}}$ to give

$$\widehat{\mathbf{y}}^* = \mathbf{X}\widehat{\mathbf{p}} + \widehat{\mathbf{e}}^* \tag{11}$$
$$\widehat{\mathbf{p}}^{*(b)} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\widehat{\mathbf{y}}^* \tag{12}$$

---

[2] If the sensors are correlated, the inverse may not exist. In practice, this would need to be verified at an earlier stage of the process.
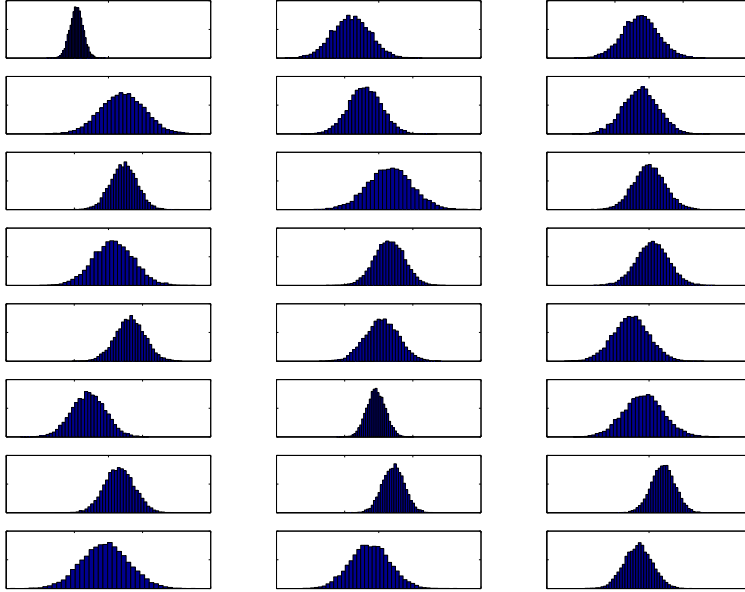
Fig. 3. Bootstrap estimates for the linear estimation parameters. Each column represents one of the mixture vapor components (DMMP-water-ethanol), with the distribution of each of the 8 predictor components (for 7 sensors plus a constant offset) shown.

The resulting bootstrap estimates are shown in Figure 3 for the $(M + 1) = 8$ sensors and $c = 3$ components.

4 Nonlinear Estimation

The above assumes that a linear prediction model is appropriate. That is, that the specific concentration of an analyte may be found as a weighted linear combination of all averaged sensor responses. The linearity issue is investigated in this section; the wisdom of averaging the sensor responses is also briefly investigated.

We employ a standard multilayer perceptron (MLP) configuration as indicated in Figure 4 [4,2]. In this configuration, the $M = 7$ sensor inputs are used to simultaneously predict the $c = 3$ concentration outputs in the ternary mixture.

Referring to Figure 4, defining x_i as the ith successive layer from input through hidden nodes, d_k as the target (desired) output, s_j as the weighted summation at each layer, weighted using w_{ij} from i to j. Each weighted output is subjected to a nonlinearity f. This so-called neuron activation function is a standard "sigmoid" function
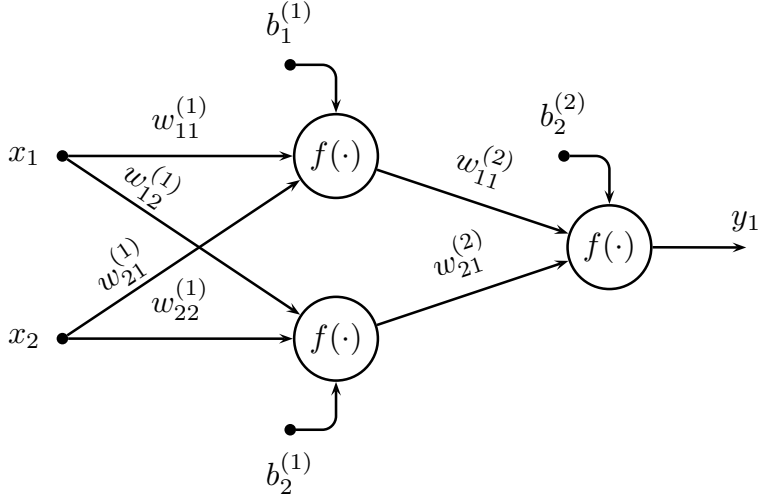
Fig. 4. Neural network architecture for the component estimation.

$$f(s, \theta) = \frac{1}{1 + e^{-(s+\theta)}} \tag{13}$$

with each neuron (node) performing the summation

$$s_j = \sum_k w_{kj} x_k \tag{14}$$

The training methodology is somewhat different to the linear case, which has a closed-form solution. Similar to the linear case, we have split the experimental data into 80/20 proportions so as not to bias the performance testing (ie the test data remains unseen). The learning algorithm is incremental; at each update we incrementally adjust the weights w_{ij} according to

$$w_{ij}(t + 1) = w_{ij}(t) + \eta \delta_j x_i \tag{15}$$

η is the empirically chosen learning rate, and affects the convergence of the training. In these experiments, $\eta = 0.1$ was utilized, though it is not critical. The internal parameter δ_j is necessary to estimate the gradient of the mean-square error with respect to the weight parameters, and is found for output nodes to be

$$\delta_j = y_j(1 - y_j)(d_j - y_j) \tag{16}$$

and for internal (hidden) nodes

$$\delta_j = x_k(1 - x_j) \sum_k \delta_k w_{jk} \tag{17}$$

8

The bias b is a special case of the weight w where the input is considered to be unity. To speed up the learning process, a "momentum" term α is utilized in the gradient descent

$$w_{ij}(t+1) = w_{ij}(t) + \eta \delta_j x_i + \alpha \left(w_{ij}(t) - w_{ij}(t-1) \right) \tag{18}$$

Because the gradient term $\delta \to 0$ as $x \to \pm 1$, the inputs x_i were scaled to be between 0.2 and 0.8. This scaling is reversed for the outputs y_i.

Tables 1 and 2 give some indicative results using this approach for the dimethyl methyl phosphonate (DMMP) and acetone-water-ethanol (AWE) mixtures for the test data set (selected to be outside the training data). Estimation of the output concentration is good in the majority of cases. One problem in practice is that the inverse scaling at the output of the network does not necessarily produce results which satisfy the conditions $0 \le y_k \le 1$ and $\sum y_k = 1$. This posterior-probability summation is required to ensure the concentrations sum to 100%.

Table 1
AWE estimation using ANN (20 nodes, 259 training patterns, 65 unseen test patterns, MSE 0.0118) for test dataset (outside the training dataset).

AWE Mixture					
Acetone		Water		Ethanol	
Actual	Predicted	Actual	Predicted	Actual	Predicted
0.00	0.00	0.00	0.00	1.00	0.99
0.10	0.03	0.00	0.00	0.90	0.97
0.20	0.18	0.00	0.04	0.80	0.76
0.40	0.37	0.00	0.00	0.60	0.62
0.60	0.72	0.20	0.21	0.20	0.09
0.40	0.43	0.40	0.43	0.20	0.11
0.20	0.03	0.40	0.45	0.40	0.50
0.20	0.15	0.40	0.42	0.40	0.39
1.00	0.80	0.00	0.00	0.00	0.20
0.50	0.41	0.00	0.01	0.50	0.58

Table 2
DMMP estimation using ANN (10 nodes, 646 training patterns, 162 unseen test patterns, MSE 0.0042) for test dataset (outside the training dataset).

DMMP Mixture					
DMMP		Water		Ethanol	
Actual	Predicted	Actual	Predicted	Actual	Predicted
0.70	0.70	0.10	0.12	0.20	0.15
0.70	0.70	0.0	0.04	0.30	0.26
0.50	0.42	0.40	0.46	0.10	0.11
0.50	0.42	0.0	0.06	0.50	0.54
0.30	0.25	0.70	0.73	0.0	0.01
0.30	0.27	0.60	0.53	0.10	0.19
0.30	0.25	0.50	0.51	0.20	0.24
0.30	0.19	0.0	0.04	0.70	0.77
0.10	0.08	0.90	0.88	0.00	0.03
0.10	0.08	0.0	0.02	0.90	0.88
0.60	0.46	0.30	0.44	0.10	0.08
0.00	0.02	0.10	0.13	0.90	0.83

5 Results

We have studied two vapor systems: one included the nerve gas stimulant dimethylmethyl phosphonate (DMMP) at parts-per-billion (ppb) concentrations and water and ethanol at parts-per-million (ppm) concentrations (DWE mixtures); the other system included acetone, water and ethanol all of which were at ppm concentrations (AWE mixtures). In both systems, individual, binary and ternary mixtures were detected.

Figure 5 shows a comparison of predicted concentration errors for the AWE experiment, for both the least-squares model, together with the MLP model using various numbers of hidden nodes. Each 3-bar represents the distribution of the 3 components in the mixture.

Figure 6 shows similar results for the DMMP experiment. It is evident that the MLP requires of the order of 10-20 hidden nodes to produce good estimates, especially for the outliers.

Several observations are in order. Firstly, the linear model produces surprisingly good predictions, though this was not unexpected given the preliminary

analysis of the covariance matrix of the experimental data. That is, some degree of linear dependence is present, and may be removed by principal component analysis. The linear model appears to function differently for the two vapor mixtures compared, and this is a cause for concern if the approach were to be generalized to other mixture types.

The neural network model is seen to give broadly similar performance for smaller network configurations, and improves with the larger (20-node) model parameterization. Care must be taken in this context, since too large a number of hidden nodes can lead to the well-known problem of network "over-training" – that is, where the output of each hidden node effectively "remembers" a particular training pattern, and does not generalize to unseen patterns.

Considering the above results for linear and nonlinear models, for various neural network parameterizations, it is suggested that a linear pre-processing stage may also be beneficial, in order to decorrelate the input sensor data, and leave only a nonlinear residual for analysis. It is not clear how this would be effected, however, and may require a larger number of sensors for exploration.

It is also noted that the above results utilize only the steady-state sensor output value, rather than the full transient response of each sensor. Whilst this reduces the computational complexity (further discussed below), it may be the case that the transient response profile (as shown at the beginning of the paper) includes important information, which would obviously be overlooked if only the steady-state sensor output is analyzed. Although some preliminary investigations into utilizing the entire time-response profile have been conducted, the amount of data available by sampling each sensor at a suitable rate presents computational problems. In particular, pre-processing stage would be required to smooth the transient response, before input to a neural network. Simply feeding the raw time-response samples as input to a very large (approximately 60 time samples per response) network produces results which are not meaningful, because the network effectively tries to model the sensor noise. Inspection of the time response profiles also indicates that smoothing and perhaps time-alignment of each sensor output would be required.

6   Computational Complexity

As stated at the outset, the goal is to produce a handheld device which is capable of estimating the vapor concentrations in real-time. In this context, "real-time" is envisaged to be of the order of tens of seconds to a minute for each sample. Since the sensor response times are fixed, the only variable is the computational time for the algorithm.
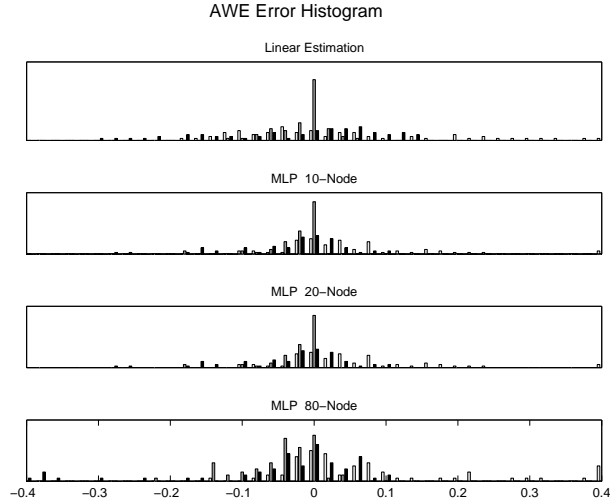
AWE Error Histogram

Linear Estimation

MLP 10–Node

MLP 20–Node

MLP 80–Node

Fig. 5. Error histograms for the AWE vapor mixture.

DMMP Error Histogram

Linear Estimation
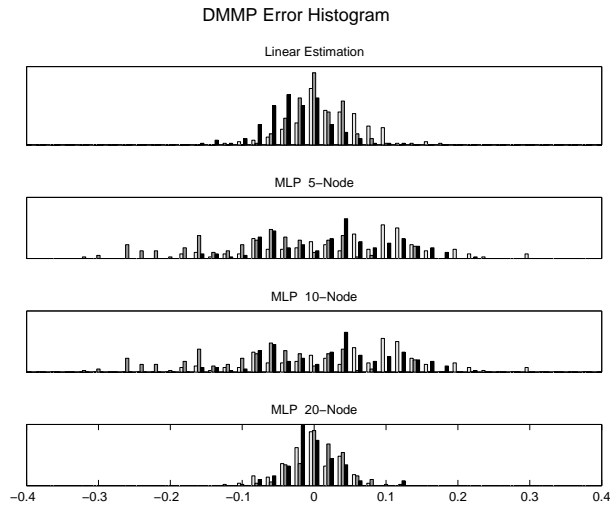
MLP 5–Node

MLP 10–Node

MLP 20–Node

Fig. 6. Error histograms for the DMMP vapor mixture.

Both the linear and nonlinear models presented here are highly asymmetric. That is to say, the determination of the parameter model in the off-line or training phase takes considerable processing, whereas the on-line estimation in practice using the predetermined parameters is considerably simpler. Furthermore, the parameter estimation in the case of the linear model is bounded, since the coefficients in the model are found by solving Equation (8). The most complex calculation in this case is the matrix inversion, but this is also relatively straightforward since (using the terminology defined previously) for for N experiments and M sensors, the linear response for each concentration $y_{c,n}$, the computation $(\mathbf{X}^T\mathbf{X})^{-1}$ requires an inversion of order M, since $\mathbf{X}$ is $N \times (M+1)$. In the present case, M is 7, so the inversion is relatively straightforward (though not trivial). The final computation performed in the hand-held device is a multiplication of each steady-state value by the precalculated coefficient (according to the model parameterization), and hence

each component is calculated according to the corresponding row of Equation (1). Again, since M is relatively small (equal to the number of sensors), this calculation is straightforward.

The nonlinear neural network model is more complicated in both parameter derivation and on-line calculation. Unlike the linear model, the training time for the parameters is not bounded, and depends on the convergence of the iterative training used. This calculation is done off-line in the training phase, and does not impact on the on-line performance of the device once trained. The on-line calculation for each vapor concentration estimate requires a neural network with M inputs, and H hidden nodes, and 3 outputs (in the present case). The computation proceeds according to Equation (14) for the coefficient multiplication and summation. The overall complexity is thus $M \times H + 3 \times H$, with a further H nonlinear activation functions computed according to 13. For the case of $M = 7$ and $H = 20$ hidden notes (as presented in the example), the complexity is of the order of 400 floating-point calculations (multiply-add) for all 3 estimations. This is well within the "real-time" design constraint, even for power-limited devices.

7   Conclusions

We have investigated linear and a multilayer perceptron estimators for the mapping of nonspecific electronic sensors into vapor mixture concentrations. For the present application, detection of the presence of a particular vapor was insufficient; concentration estimates were also required. Microcantilever sensors are utilized; these have the advantage of small size and fast response, but are not specific to individual vapor responses like other chemical sensors. It has been demonstrated that it is possible to estimate the concentration of each vapor in a ternary mixture with surprising accuracy. Although it would be assumed that the sensor responses are nonlinear, it has been demonstrated that a linear model produces good estimates.

The linear approach requires a separate weight parameter vector for each component, and the training phase comprises a one-pass calculation. The nonlinear approach utilizes one set of interconnecting weights to simultaneously estimate all three gas concentrations. However, it also requires a steepest-descent iterative weight parameter re-estimation, and determination of the terminating criteria can be problematic.

Further work requires the investigation of a dual linear-nonlinear hybrid approach, wherein the linear component is first removed, followed by a nonlinear estimator. Also, the investigation of the usefulness of the transient response profile of each sensor is promising, however initial linear filtering is required

because of the sensitive nature of the sensors themselves.

Acknowledgements

References

[1] Canti-4$^{\mathrm{TM}}$ sensor chips, Cantion A/S, A/S Skjernvej 4A DK9220 Aalborg Ø Denmark.

[2] R. Beale, T. Jackson, Neural Computing: An Introduction, IOP Publishing, Bristok, UK, 1990.

[3] J. W. Gardner, P. N. Bartlett, Electronic Noses: Principles and Applications, Oxford University Press, 1999.

[4] R. Lippmann, An Introduction to Computing with Neural Nets, in: C. Lau (ed.), Neural Networks – Theoretical Foundations and Analysis, chap. 1, IEEE Press, 1992.

[5] L. A. Pinnaduwage, W. Zhao, A. C. Gehl, S. L. Allman, A. Shepp, K. K. Mahmud, J. W. Leis, Quantitative Analysis of Ternary Vapor Mixtures using a Microcantilever-Based Electronic Nose, Applied Physics Letters 91 (4).

[6] W. Zhao, J. W. Leis, L. A. Pinnaduwage, A. C. Gehl, S. L. Allman, A. Shepp, K. K. Mahmud, Identification and Quantification of Components in Ternary Vapor Mixtures using a Microcantilever Sensor Array and a Neural Network, Journal of Applied Physics 103 (10).

[7] A. M. Zoubir, D. R. Iskander, Bootstrap Methods and Applications, IEEE Signal Processing Magazine 24 (4) (2007) 10–19.