Application of Self-Organising Maps (SOM) to Chemical Spectra Analysis

H. Tokutaka*, K. Obu-Cann, K. Fujimura, K. Yoshihara[1] and Metal Materials Group of SASJ

Department of Electrical and Electronic Engineering, Tottori University, 4-101 Koyama, Tottori 680-8552, Japan
[1]*National Institute for Materials Sciences, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan*
*tokutaka@ele.tottori-u.ac.jp

(Received December 18, 2001; accepted May 8, 2002)

This paper reports on the application of SOM to chemical spectra analysis. The Self-Organising Map (SOM) method that was developed by T. Kohonen [1] was first applied to information processing. Currently, it has been applied to some problems of chemical spectra analysis using AES (Auger Electron Spectroscopy), XPS (X-ray Photoelectron Spectroscopy), and XRD (X-ray Diffraction) data. Using a 2-dimensional SOM, it became clear that the items that are described qualitatively by linguistic expressions could be explained more quantitatively by the position of the spectral data on the SOM together with a grey level expression. Also, the composition of an unknown sample can be determined very precisely by the SOM that has been constructed using the spectra from samples of known composition. Furthermore, this paper addresses the attempts to develop a SOM of all the elements of the periodic table. Currently, only 77 elements have been mapped out.

1. Introduction

During the1970s, Kohonen (Kohonen et al. [1]) developed the Self-Organising Maps (SOMs) that simulate the brain function. SOMs are a kind of visible neural network that has made significant impact in many research fields [2,3]. One of the characteristics of his network is its ability to transform multidimensional data into a 2-dimensional SOM. SOM was first used as an information-processing tool in the fields of speech and image recognition. The visualisation characteristics of SOM have attracted many researchers in other fields. Walker has done a pioneering work using neural networks for chemical multi-image analysis, where SOM is concluded to be more suitable in handling large dimensional problems than the usual back-propagation method (Walker et al. [4]). Zupan et al. [5] have discussed the application of SOM in their book where Fig. 18-10 corresponds to the original input data and Fig. 18-11 corresponds to the interpolated results. With a 2-dimensional SOM, it is possible to distinguish the difference between similar and dissimilar information by using a technique that illustrates the magnitude of the distance between the units on the map.

SOM has been applied to AES, XPS, and XRD data. For each kind of spectroscopy the SOM was built using the abscissa of AES, XPS and XRD spectra as the multi-dimensional description of the shape of each spectrum.

In section 2, the SOM algorithm is briefly reviewed with a typical example. In sections 3, 4 and 5 the results of AES, XPS, and XRD applications of SOM are demonstrated. Finally, section 6 discusses the attempts to develop a SOM of all the elements of the periodic table.

2. The SOM Algorithm

A brief introduction

Based on the functions of a neuron cell of a living thing especially the information processing ability of the human brain, (Kohonen et al. [1]) the following equation was developed:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + \alpha(t)[\boldsymbol{x}(t) - \boldsymbol{m}_i(t)] \qquad (1)$$

Consider the information processing ability that the present neuron cell (node) i possess at time t as $\boldsymbol{m}_i(t)$ and $\boldsymbol{x}(t)$ an input signal as shown in Fig. 1. At time t, the cell learns this input signal.

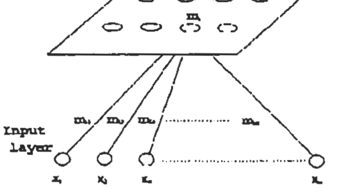

Fig. 1 The relationship between the multi-dimensional input vector x and the reference vector $\boldsymbol{m}_i$ of (unit i).

During the next time $t+1$, it has an information processing ability of $m_i(t+1)$ which is close to the input signal. If $x(t)$ is an n-dimensional input vector as shown in Fig.1, then $x(t)$ is expressed as $x(t) = [\xi_1, \xi_2, .., \xi_n]$. The n-dimensional reference vector $m_i(t)$ is also expressed as $m_i(t) = [\mu_{i1}, \mu_{i2}, .., \mu_{in}]$. The learning coefficient factor $\alpha(t)$ has values between 0 and 1. The neighbouring units $N_c(t)$ which surround $m_i(t)$ also learn the input vector $x(t)$ following the same equation.Furthermore, $t = 0, 1, 2, ..$ are discrete time co-ordinates. If an n-dimensional input vector is presented to the SOM network, then the reference vector in the network, which is closest to the input vector, is selected as the best matching node "winner". Prior to learning, a large reference unit area that surrounds the winner is selected as a neighbourhood region. The reference unit vectors in this neighbourhood region learn the input vector $x(t)$ as well as the winner, following eq.1. This is a typical cycle of the learning process. Another cycle of learning starts when the next input vector is presented to the SOM network. The size of the neighbourhood region in this cycle is reduced as compared to the last learning cycle. Thus, learning continues until the very last learning, where only the winner is trained by the input vector. Kohonen was the first to introduce a 2-dimensional SOM map from multi-dimensional input data. This is considered in Table 1, which distinguishes various animals with 13 characteristics, each of which is regarded as a dimension. The characteristics have a value of one (1) if the animal has that particular characteristic and zero (0) if it does not. Every animal has a 13-dimensional input vector. When the table is projected to a 2-dimensional SOM with 10 x 10 neuron units, Fig.2 is obtained.

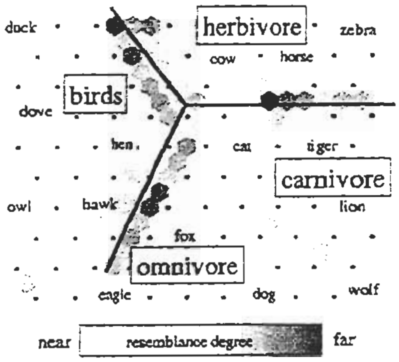

Fig. 2 After the network has been trained with inputs describing characteristic sets from Table 1, the map was calibrated by the columns of Table 1 and labelled correspondingly.

As shown in the Fig.ure, it is understood that the 16 animals are distributed with similar characteristics. For example, "birds" are on the left side of the figure, and "carnivores" like "tiger", "lion", and "wolf", are on the right. "Herbivores" like "zebra", "horse", and "cow" are placed in the upper part of the Fig.ure.

Table 1: Types of animals and their characteristics (attributes of 4, 11 and 16 are added to ref [1]).

attribute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	small	medium	big	nocternal	2legs	4legs	hair	hooves	mane	feathers	stripes	hunt	run	fly	swim	herbivore
dove	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0.5
hen	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0.5
duck	1	0	0	0	1	0	0	0	0	1	0.3	0	0	1	1	0.5
goose	1	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0.5
owl	1	0	0	1	1	0	0	0	0	1	0	1	0	1	0	0
hawk	1	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0
eagle	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0
fox	0	1	0	0.5	0	1	1	0	0	0	0	1	0	0	0	0
dog	0	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0
wolf	0	1	0	1	0	1	1	0	1	0	0	1	1	0	0	0
cat	1	0	0	0.5	0	1	1	0	0	0	0	1	0	0	0	0
tiger	0	0	1	0.5	0	1	1	0	0	0	1	1	1	0	0	0
lion	0	0	1	0	0	1	1	0	1	0	0	1	1	0	0	0
horse	0	0	1	0	0	1	1	1	1	0	0	0	1	0	0	1
zebra	0	0	1	0	0	1	1	1	1	0	1	0	1	0	0	1
cow	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	1

3. Applications of SOM to full scanned AES data

The results of Fig.2 are applied to the problem of chemical analysis shown in the AES graphs of Fig.3. In the case of Fig.2, there are 13-dimensional entities, but in AES the energy steps on the horizontal axis are considered as the dimensional units. The primary beam energy of 5 keV was used for all AES data. The AES data of Fig.3 are measured from 20 eV to 982 eV by 1 eV division. Therefore, each spectrum is a 963-dimensional input vector. The vertical axis shows the signal size. In Table 1, signal values were expressed as 0 and 1.

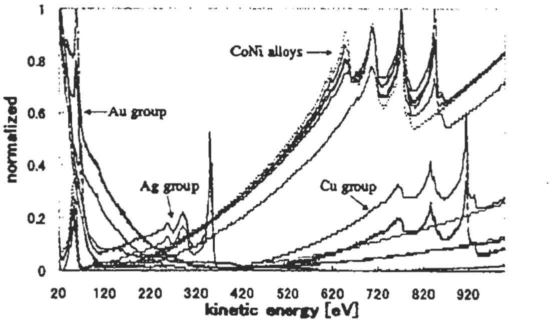

Fig. 3 AES data of 15 materials from 20 to 1000eV.

However, in Fig.3, the normalised signal between 0 and 1 become the signal magnitudes. The SOM for this particular case is shown in Fig.4 using a grey scale: the darker the grey the greater the distance between the nodes. As shown in the

Fig.ure, the results for CoNi alloys are systematically arranged from Ni 100% to Co 100%. Cu follows next though there is quite some distance between them as shown by the grey level. Au and Ag groups follow respectively. Conventionally, subtracting the background and/or separating the main peak from several smaller peaks have interpreted Auger spectra. The present method considers only the spectral shape as an information source. Using experimental AES data, the CoNi alloys are very well separated and in a systematic order as shown in Fig.4.

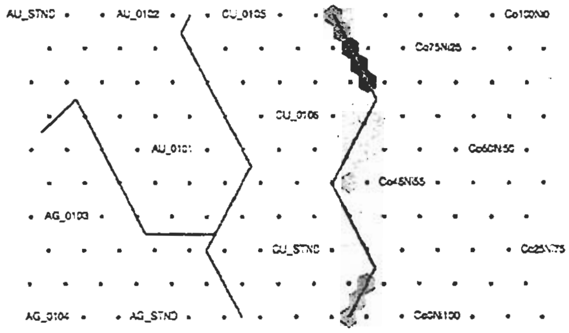

Fig. 4 2 dimensional SOM of 15 materials from AES data of Fig.3, where STND is the abbreviation of standard.

3.1 Quantitative application to chemical data mining

The SOM method is used for a more quantitative purpose using the AES data of Fig.3. In this case, the large backgrounds, which increase in the higher energy region, are subtracted linearly in order to raise the LMM signal sensitivity for CoNi alloys. The procedure for background subtraction is shown in Fig.5. All the normalised AES signals from 590 eV to 890 eV after background subtraction are shown in Fig.6. In the construction of the SOM in Fig.7, each energy step on the horizontal axis is taken as a single dimension. The compositions of the 6 CoNi alloys shown in Fig. 3 are considered as new dimensions between 0 and 1. CoNi50% alloy for example would have a dimension of 0.5. The final dimensions are composed of the composition label and the energy step of 1 eV of the AES spectra. The normalised composition values and the AES signals (AES signal −Min AES signal), which are normalised by the difference between the maximum and minimum AES signals, are taken as the new signal values. The SOM, which includes the composition labels, is shown as a 20x30 sized grid in Fig.7.

Five samples of CoNi alloys Ni 100, 75, 50, 25 and 0 % were used as input signals to the SOM.

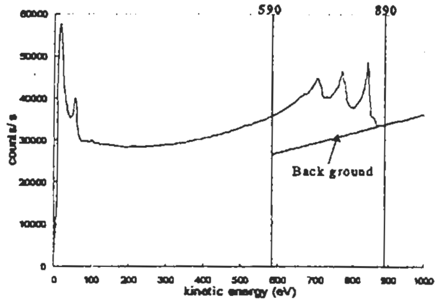

Fig. 5 The background is subtracted linearly between 590 and 890 eV using Ni 100 % AES spectra.

After the SOM learning, all the 20×30=600 units in Fig.7 are compared by the following error function (Err):

$$Err = \sum_{j=1}^{n} (x_j - m_{ij})^2 \qquad (3)$$

Where x_j and m_{ij} are the j-th component value of the n-th dimensional input data and i-th unit in Fig.7, respectively. Using eq.3, all the 5 input data are compared with all the 600 units.

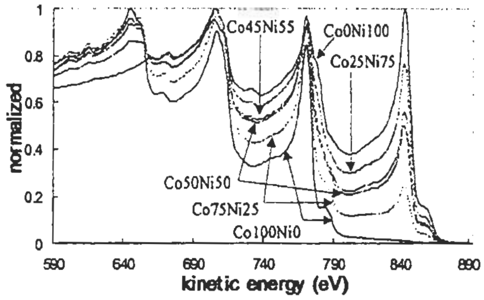

Fig. 6 6 CoNi AES spectra with backgrounds already subtracted linearly between 590 and 890 eV using the method described in Fig.5.

All labelled positions in Fig.7 are determined by the minimum values of the Error function of eq.3. For the purposes of data mining, the Ni 55 % spectrum data is used as test data and its composition is assumed to be unknown. Using eq.3, all the 600 units are compared one after the other with Ni 55 % spectra data (test data). The unit with the lowest value of the Error function of eq.3 is identified as the closest unit and marked by the large coloured circle, as shown in Fig. 7. For this experiment, the closest unit had a composition value of 55.26 %, which is very close to the true value of 55 %. The spectra of the original Ni 55 % and the learned unit are compared in Fig.8 with an error margin of 0.3. The labelled units of the alloys are arranged in a systematic order from 0 to 100 % as shown in

Fig.7. The shape of Ni 100 % is a little different from the other 5 samples. Therefore, the distance between Ni 100 % and the other alloys in the SOM is clearly shown in grey at the left upper position of Fig.7.

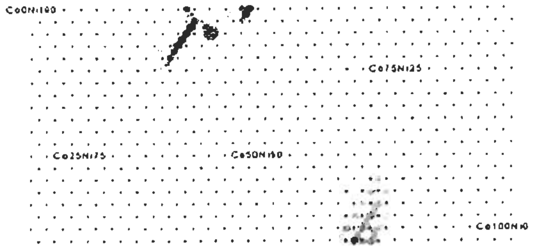

Fig. 7 2 dimensional SOM of the 5 CoNi alloys of AES data of Fig.6 where Co45Ni55 data is excluded from the input data in the construction of SOM. The large coloured circle marks the unit that corresponds to the input data spectra of Co45Ni55.

The above experiment was carried out using other Ni alloy compositions (Ni 25 %, 50 %, and 75 %). The composition error was less than 1 % in all cases. The spectra reproduction accuracy of the SOM method was compared with the usual synthesising method, where the spectra can be built up using Co 100 % and Ni 100 %. The accuracy estimation was carried out using the squared error method similar to eq.3, 0.008 for SOM and 0.025 for the usual synthesising method. The SOM is far better than the usual synthesising method, because any composition data and as many data as possible can be used for the construction of SOM.

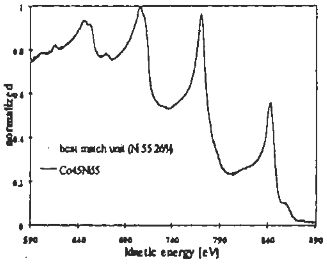

Fig. 8 AES spectra of the original Co45Ni55 input data and learned best match unit with the composition label of 55.26 %.

4. Characteristic analysis of High-Tc super- conducting oxides using SOM

SOM is applied to XPS data. The surfaces of single-crystal, ceramics and thin films of Bi based high-Tc super-conductors are cleaned by heating and the impurities are removed. The state of cleanliness is examined as a change of XPS signal of oxygen (O2). The result is shown in Fig. 9. In this case, the binding energy of the horizontal axis is considered as 300 points or 300 dimensions. The vertical axis is a normalized signal domain. The cleaved surfaces of single crystals are heated to 400 ℃ and are considered as a standard clean. Using this standard, SOM is constructed as shown in Fig. 10. There are three kinds of peaks in Fig. 10. These are right, split and left peaks. These three groups are clearly found with the dark grey valleys as shown in Fig. 10.

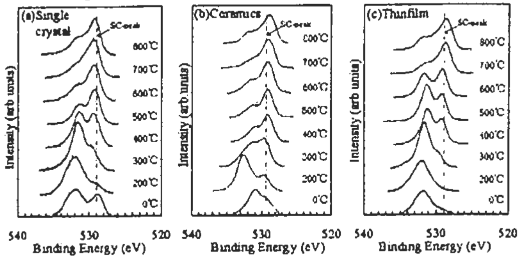

Fig. 9 Changes of O-1s XPS (X-ray Photoelectron Spectroscopy) spectra from Bi-based single crystals, ceramics and thin films during thermal annealing processes.

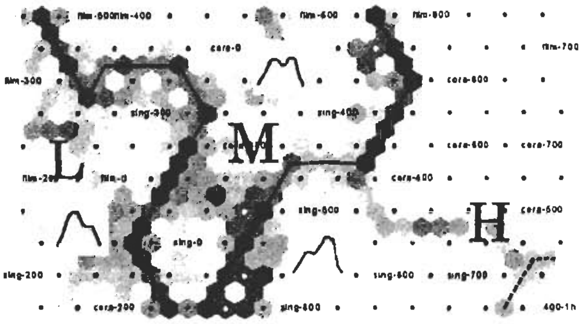

Fig. 10 SOM classification map of Fig.9, where O-1s XPS spectra from the air-cleaved Bi-based single crystal surface, which is annealed for 1h at 400℃, are used as a standard and indicated as 400℃-1. L: annealed at low temperature, M: annealed at medium temperature and H: annealed at high temperature. Cera represents ceramics, sing for single crystal and film for thin film.

Spectra with peaks to the right consist of the heat-treated clean surfaces. Single crystals are closest to the standard sample, which is cleaned at 400℃. Ceramics follow next. For thin films, the heat-treated samples at 700 and 800℃ are arranged next to the ceramics. However, spectra that have double peak and left peak form a cluster due to their shape. The clusters are not separated into single crystals, ceramics or thin films neither by their heat-treated temperatures. It is concluded that surface impurities affect SOM classifications more than material characteristics, because there are impurities left on the surfaces of the samples of the groups with a double peak or a left peak.

5. Application of SOM to XRD (X-Ray Diffraction) data

SOM is applied to other result of XRD from thin film. The experimental data of XRD is shown in Fig. 11. In the Fig.ure, Samples 1 and 2 are the usual experimental data from thin films. In the same Fig.ure, XRD results from 2212 (80K phase) and 2223 (110K phase) that are finely identified single crystal thin films are used as standard. It can be easily understood from the Fig.ure that Samples 1 and 2 will not be 2212 (80K phase), since the fourth peak (around 22.5°) does not fit these samples. The constructed SOM is shown in Fig. 12.

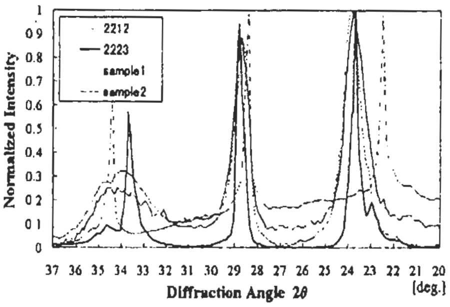

Fig. 11 XRD (X-Ray Diffraction) patterns of Bi-based thin films. In the experiments, the films, which are approximately the single-phases of 110K phase (2223) and 80K phase (2212), are used as a standard. Samples 1 and 2 are usually obtained.

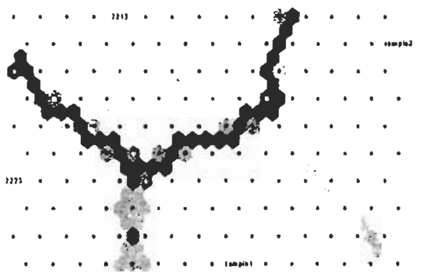

Fig.12. A SOM classification map of Fig. 11.

2212 (80K phase) is surrounded by a deep grey level. From this result, it can be stated that both Samples 1 and 2 are not 2212 (80K phase). However, as shown in the Fig.ure, both Samples 1 and 2 are connected to 2223 (110K phase) with a thin grey level, with Sample 1 closer to 2223 (110K phase) than Sample 2. Even though the resistance-temperature (R-T) experimental results have not been included in this article, it should be noted that interpretation from the SOM of Fig. 12 agrees quite well with the R-T experimental results.

6. SOM map using all the elements in the periodic table

It is intended to construct a SOM map using all the elements in the periodic table. Currently data for 77 elements in the periodic table have been obtained. These data are magnesium excited XPS data obtained from Ulvac-phi company. The data were measured from 1 eV to 960 eV by 1 eV division. Therefore, each spectrum is a 960-dimensional input vector. The vectors were processed as in section 3.1 before introduced to the SOM. Table 2 illustrates the 77 elements used for the SOM map of Fig. 13.

Table 2: List of the 77 XPS Mg excited elements in the periodic table (coloured) used for the SOM of Fig.13.

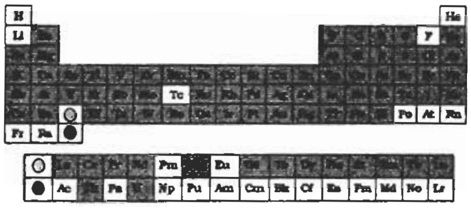

Figure 13 illustrates the SOM map for the Mg excited XPS signals of the 77 elements of the periodic table. The data from another database COMPRO6 is also processed as in section 3.1 and compared to the data of Ulvac-phi in the SOM map. The data for COMPRO6 were also measured from 1 eV to 960 eV by 1 eV division.

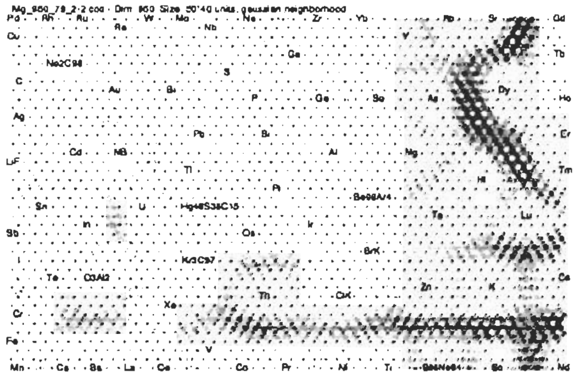

Fig. 13: SOM of the 77 XPS Al excited elements in the periodic table

As an example, the data for Nickel (Ni) from COMPRO6 is used as test data and its best match unit identified on the SOM map of Fig. 13, which was developed using data from ULVAC-PHI. The position of the BMU was found to be very close to that of Ni from ULVAC-PHI with a mean squared error of 0.84.

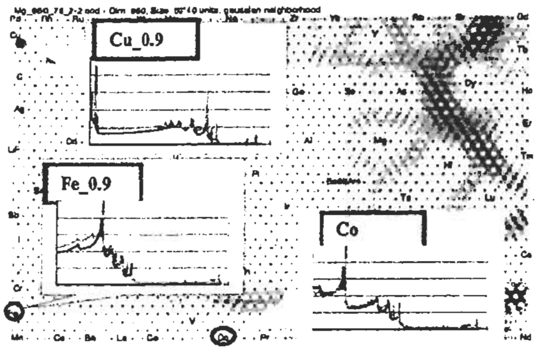

Fig. 14 Comparison of the signals from COMPRO6 and LVAC-PHIi companies.

Fig.14 illustrates the comparison of the signals from COMPRO6 and Ulvac-phi. Furthermore, Fig. 15 illustrates the two Ni signals from COMPRO6 and Ulvac-phi. The spectra for other elements were compared, for instance, the comparison for iron (Fe) resulted in a mean squared error of 1.62. Similar analysis was done for all the other elements of the SOM map of Fig. 13. Conventionally, Auger spectra have been interpreted by subtracting the background and/or separating the main peak from several smaller peaks. The present method considers only the spectral shape as an information source. If the aim of constructing a SOM for the periodic table is achieved, then the chemical analysis of the elements of the periodic table can be performed using SOM. If one is confronted with the spectrum of an unknown element, the signal values of this element can be used as test data and its best matching spectrum on the map can be identified and compared with the test data in order to identify the unknown element. For the case of composition analysis, a homepage is currently been developed for Iron-Nickel alloys. Here, the composition of an unknown spectrum can be analysed (after the necessary pre-processing has been done) by obtaining the best matching spectrum on the map to determine the various compositions that make up the alloy. When this is completed, it will provide researchers with another avenue for spectra analysis.

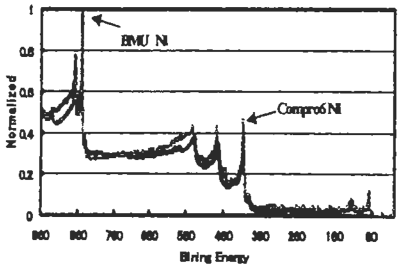

Fig. 15 Comparison of the two Ni signals from ULVAC-PHI and COMPRO6.

7. Conclusion

The SOM by Kohonen has been applied to some problems of chemical analysis as a preliminary examination. For CoNi alloys, the composition labels of unknown spectra can be estimated. Usually, results such as Fig. 3, are explained linguistically. However, the introduction of the SOM of Fig. 4, enables us to make the above results more visible and describe them more quantitatively. Thus, the position of each spectrum on the 2 dimensional SOM, depends on the similarity between the spectra in the input data. The introduction of this method has made more quantitative discussions possible to experimental results that are described linguistically. In chemical spectral data analysis, the details of the spectra and the composition analysis can be carried out using methods that consider the physical meaning of the spectra (Briggs et al. [6]). The analysis can also be done by methods such as smoothing by Savitzky-Golay, least square method, multivariate analysis, and principal component analysis, etc. (Sasaki et al. [7]). However, SOM is very useful as a first stage of pre-processing before obtaining details using the above mentioned sophisticated methods. To examine the SOM further, the SOM map of Fig.7 is being set up on our homepage. Any user, who has CoNi spectra data with the backgrounds already subtracted by the method of Fig.5, can browse through our homepage and download the map. Using eq.3, the user can the user can easily find the best-fit unit of his data among the 600 units in the SOM map and obtain an estimation of the composition of their spectra data. In the case of the traditional method such as a synthesising method, both pure 100 % Co and Ni signals are needed in order to find any composition spectra, using the following equation:

$$x\text{Co} + (1-x)\text{Ni},$$

Where Co and Ni are pure 100 % materials. Adjusting x, we can find the value of x which is the closest synthesised spectra to the input CoNi alloy signal. For the SOM method, we do not need pure 100 % Co and Ni materials, all one has to do is to follow the manual in order to subtract the backgrounds and then normalise the input signals. Thus, the introduction of SOM to chemical analysis allows us to explore a new development for future progress.

References

[1] T. Kohonen, Self-Organising Maps, Springer Series in Info. Sciences, Volume 30, 1995.
[2]Proc. of WSOM'97, Workshop on SOM, Univ.of Tech., Helsinki, Finland, June 4-6, 1997.
[3] Visual SOM, http://websom.hut.fi/websom/.
[4] C. G. H. Walker, SIA., **24**, 173, (1996).
[5] J. Zupan and J. Gasteiger, Neural Networks for Chemists, An intro. pp269 – 270 VCH Verlagsgesellschaft.mbh.(1993)
[6] D. Briggs and M. P. Seah, PSA, John Willey & Sons, (1983).
[7] K. Sasaki, S. Kawata, and S. Minam, Appl.. Opt., 23, 1955 (1984).