Transactions on
Case-Based Reasoning
Vol. 5, No. 1 (2016) 19-31

ISSN: 1867-366X ,
ISBN: 978-3-942952-46-0

ibai Publishing
www.ibai-publishing.org

A New Similarity Measure for Sequences and Time-Series Applied to Spectrometer Signal Analysis

Petra Perner

Institute of Computer Vision and applied Computer Sciences,
Kohlenstrasse 2, 04107 Leipzig, Germany
pperner@ibai-institut.de

Abstract. Different spectrometer methods exist that have been developed over time to practical applicable systems. Researchers in different fields try to apply these methods to different applications especially in the chemical and biological area. One of these methods is RAMAN spectroscopy for protein crystallization or Mid-Infrared spectroscopy for biomass identification. For the applications are required robust and machine learnable automatic signal interpretation methods. These methods should take into account that not so much spectrometer data about the application are available from scratch and that these data need to be learnt while using the spectrometer system. We propose to represent the spectrometer signal by a sequence of 0/1 characters obtained from a specific Delta Modulator. This prevents us from a particular symbolic description of peaks and background. The interpretation of the spectrometer signal is done by searching for a similar signal in a constantly increasing data base. The comparison between the two sequences is done based on a syntactic similarity measure. We describe in this paper how the signal representation is obtained by Delta Modulation, the similarity measure for the comparison of the signals and give results for searching the data base.

Keywords: Computational Methods, Delta Modulation, Feature Extraction, Incremental Knowledge Acquisition, Spectrometer signal analysis, Similarity-based Signal Interpretation, Case-Based Reasoning.

1 INTRODUCTION

Different spectrometer methods exist that have been developed over time to practical applicable systems. Researchers in different fields try to apply these methods to different applications especially in the chemical and biological area. One of these methods is RAMAN spectroscopy for protein crystallization [1], [2] or Mid-Infrared spectroscopy for biomass identification [3]. For the applications are required robust

and machine learnable automatic signal interpretation methods. These methods should take into account the sparse available data for the application and that new data need to be acquired while using the spectrometer system. We propose a novel spectrometer analysis method based on Delta Modulation and similarity determination. We represent the spectrometer signal by a sequence of 0/1 characters obtained from a specific Delta Modulator. While doing this we preprocess the signal by smoothing at the same time. This prevents us from the extraction of a specific symbolic description of peaks and background from the basic spectrometer signal based on signal-theoretic methods [4]. The interpretation of the spectrometer signal is done by searching for a similar signal in a constantly increasing data base. The two 0/1 sequences of the spectrometer signal are compared based on a syntactic similarity measure.

The proposed new method has been tested on RAMAN spectrometer signals for screening of bio-molecular interactions but the method can be used for all kinds of spectrometer signals. With the aid of Raman spectroscopy, the vibrational spectrum of molecules can be examined. Functional groups like amino, carboxyl or hydroxyl groups can be identified through characteristic vibrational frequencies.

In this paper the architecture of the spectrometer-signal analysis system is described in Section 2. The calculation of the signal representation obtained by Delta Modulation is explained in Section 3 for three different kinds of delta modulation methods. Then we describe three different syntactical dissimilarity measures used for this study in Section 4. Finally, we give results in Section 5 for the three Delta Modulation methods and select the best one. This method is used for further studying the best dissimilarity measure. We show how good these measures can group similar spectra. At the end we use a prototype-based classifier to show how good we can classify the spectra based on the chosen representation and with the three different dissimilarity measures. In Section 6 we give conclusion.

2 Architecture of the Automated Spectrometer Signal Program

The architecture of the automatic spectrometer identification system is shown in Fig. 1.

After preprocessing the spectrometer-signal, the signal is coded into a 0/1 sequence by the delta modulator. While doing that the signal is step-wise smoothed by a linear function. The representation makes it unnecessary to develop special high-level features that describe all interesting properties of the spectra. The sequence itself can be interpreted in different ways. It can be ask for identity, similarity of the whole sequence or for partial identity or similarity. That allows identifying part-spectra, special single peaks or peak combinations within spectra.

This sequence is compared to sequences of reference spectra stored in a memory. The name of the spectrum where the coded sequence gives the highest similarity is

given as output to the user. A side effect of the coding is also that the spectra is not stored with its real values but instead it is stored as 0/1 sequence. This saves memory capacity and makes it possible to implement the method into a special purpose processor.

When there is no similar sequence in the data base the input spectrum is stored into the data base after it has been coded by the delta modulator. The spectrum is labeled manually after it has been checked by other method what the spectrum is about.

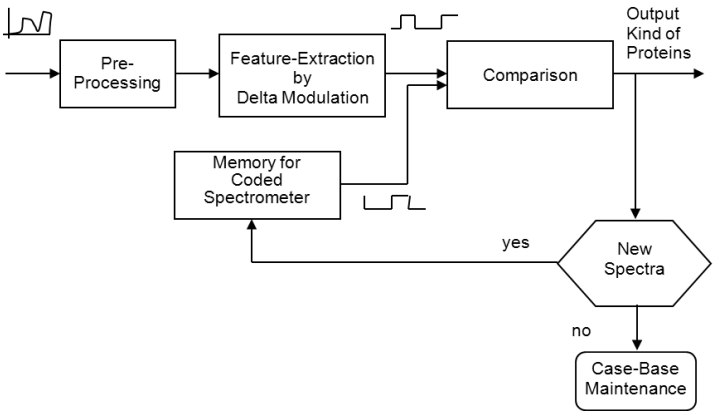

Fig. 1. Architecture of the Spectrum Interpretation System.

This data collection is necessary since the appearance of the spectra for different proteins is not known yet.

The pre-processing of the RAMAN spectra is in this special case a baseline correction [5], a Fourier transformation to eliminate the influence of the special system device and its parts [6], and the calculation of the difference between the spectrum of the buffer and the liquid in the buffer.

3 Representation of the Spectra by Deltamodulation

The delta modulator compares the actual signal value $s(i)$ with an estimated signal value $r(i))$ of the coder. The difference $e(i)$ between these two signals is coded by only one bit. It mainly represents if the signal was increased or decreased by a certain constant. Three different methods exist to estimate actual signal value: Linear Delta Modulation (LDM) [7], Constant Factor Delta Modulation (CFDM) [8], and Continuously Variable Slope Delta Modulator (CVSD) [9], [10], [11].

3.1 Linear Delta Modulation

In case of the Linear Delta Modulation, the difference $e(i)$ between the actual signal value $s(i)$ and the estimated signal value $r(i)$ at sampling point i is calculated, see Fig. 2:

$$e(i) = s(i) - r(i) \tag{1}$$

If the difference is positive then the code D is equal *"1"* and D is equal *"0"* if the difference is negative. This binary signal D_n is stored in the memory. At the same time the magnitude of the signal to be expected at the next sampling point i is estimated from it. The corresponding rule is:

$$s(i) > r(i).D_n = 0.r = r(i-1) + \Delta u \tag{2}$$

$$s(i) \leq r(i).D_n = 1.r = r(i-1) - \Delta u \tag{3}$$

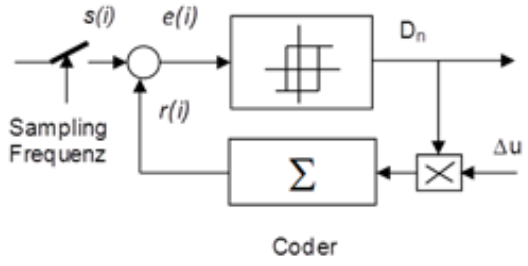

Fig. 2. Linear Delta Modulator.

The incremental size Δu is a constant value which has to be selected in function of the standard-deviation δ_Δ of the first-order difference signal: $\Delta(i) = s(i) - s(i-1)$.

On the reproduction side (which is not necessary here since we do not want to reconstruct the signal) an inversely functioning decoder then generates the original curve by means of the binary signal stored in the memory. This approximated signal is $s'(i)$. The difference between the original signal $s(i)$and the approximated signal $s'(i)$is the approximation error$\varepsilon(i) = s(i) - s'(i)$, see Fig. 3.

When process dynamics change, the linear delta modulator is not adjusted optimally anymore and the reconstruction error is increasing strongly. The adaptive delta modulators compensate this disadvantage. They dispose of a function block which takes over the control of the incremental size Δu in accordance with process dynamics. In the literature different adaptive delta modulators are known, two of which are presented in the following section.

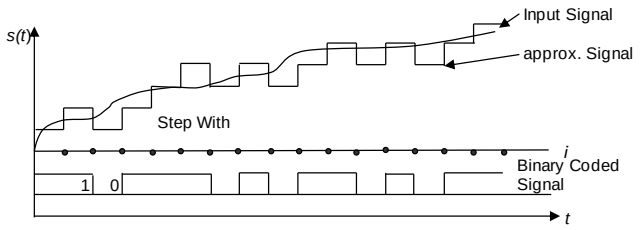

Fig. 3. Diagram with Input Signal, approximated Signal, and Binary Coded Signal.

3.2 Constant Factor Delta Modulation

The instantaneous-value compander, also called "Constant Factor Delta Modulator" (CFDM), changes its increment size at each sampling point.

An adaptation-logic decides based on the input signal sequence $\langle D_n, D_{n-1}\rangle$ by which factor k the preceding increment size has to be multiplied:

$$\Delta u_i = \Delta u_{i-1} * k \tag{4}$$

with $D_n = D_{n-1}$ then $k = P$ and $D_n \neq D_{n-1}$ then $k = Q$.

It needs to be $P * Q = 1$, in order to observe the stability condition. For speech signals are the values $P = 1.5$ and $Q = 0.66$ known from the literature that have also been shown good performance in case of the application presented in this paper.

3.3 Continuously Variable Slope Delta Modulator CVSDM

The syllable compander, also called Continuously Variable Slope Delta Modulator (CVSDM), pursues, in contrast to the instantaneous-value compander, the tendency of the signal. Only when the same state has been recorded three times in a coincidence-register $\langle D_n = D_{n-1} = D_{n-2}\rangle$, the syllable compander increases its increment size. It is therefore more inert than the instantaneous-value compander. The rule for syllable companding is:

$$\text{3 bit coincidence } \; k = 1;\; \Delta u_i = \Delta u_{i-1} + 1 \tag{5}$$

$$\text{no coincidence } \; k = 0;\; \Delta u_i = \Delta u_{i-1} - 1 \text{ until } \Delta ui{=}0 \tag{6}$$

As Δu must not become zero, a minimum increment size u_{min} larger than 1 needs to be added. As standard value u_{min} can be assumed as $u_{min} = \sqrt{\delta_\Delta}$.

4 Similarity Determination Between Two SPECTRA

The spectra are represented by *0/1* sequences. To compare different spectra we need a distance measure that can work on such kind of representation. Different measures are known from text comparison and DNA sequence analysis. We choose for this work the Hamming distance [12], the Levenshtein distance [13] and the Levenshtein-Damerau distance [14].

4.1 Hamming Distance

The representation of a spectrum *A* and spectrum *B* is illustrated in Table 1. We assume that all spectra have the same length n and that the peaks are stable at their position (wavelength) in the spectra.

Then we have to compare two sequences *A* and *B*. The distance *d* between these two binary representations is the number of bits in which the two vectors are different.

Table 1. Representation of two Spectra, Sampling Points, and XOR connection.

Spectrum A	1	0	0	0	1	1	0	0	0	…
Spectrum B	1	1	1	0	1	1	0	0	0	…
Sampling Points i	1	2	3	4	5	6	7	8	9	…
A **XOR** B	1	0	0	1	1	1	1	1	1	…

That is the well-known Hamming Distance:

$$d(A,B) = \|A - B\| = \sum_{i=1}^{n} |A_i - B_i| \tag{7}$$

4.2 Levenshtein Distance

Let $d_L(A,B) = D_{m,n}/n$ be the Levenshtein-Distance between the two 0/1 sequence *A* and *B* with $m = |A|$ and $n = |B|$. The Levenshtein distance is defined as the minimum number of modifications needed to transform the sequence *A* into *B*. The allowed operations are substitutions, insertions, and deletions. The dissimilarity in $D_{0,0}$should be $D_{0,0} = 0$. Then the dissimilarity is calculated as follows:

$$D_{i,0} = i, 1 \leq i \leq m$$
$$D_{o,j} = j, 1 \leq j \leq n$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} \quad +0 \quad if \quad A_i = B_j \\ D_{i-1,j-1} \quad +1 \quad Substitution \\ D_{i,j-1} \quad +1 \quad Insertion \\ D_{i-1,j} \quad +1 \quad Deletion \end{cases} \tag{8}$$
$$for\ 1 \leq i \leq m, 1 \leq j \leq n.$$

4.3 Damerau-Levenshtein-Distance

Let $D_{DL}(A,B) = D_{m,n}/n$ be the Damerau-Levenshtein-distance between the two 0/1 sequences A and B with $m = |A|$and $n = |B|$. The Damerau-Levenshtein distance is defined as the minimum number of modifications needed to transform the sequence A into B. Besides substitution, insertion, and deletion of a single character are allowed exchange of two adjacent single characters. The dissimilarity in $D_{0,0}$should be $D_{0,0} = 0$. Then the dissimilarity is calculated as follow:

$$D_{i,0} = i, 1 \leq i \leq m$$
$$D_{0,j} = j, 1 \leq j \leq m$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} +0 \quad if \quad A_i = B_j \\ D_{i-1,j-1} \quad +1 \quad Substitution \\ D_{i,j-1} \quad +1 \quad Insertion \\ D_{i-1,j} \quad +1 \quad Deletion \end{cases} \tag{9}$$
$$for\ (1 \leq i \leq 2, 1 \leq j \leq n)\ or\ (1 \leq i \leq m, 1 \leq j \leq 2)$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} \quad +0 \quad if \quad A_i = B_j \\ D_{i-1.j-1} \quad +1 \quad Substitution \\ D_{i,j-1} \quad +1 \quad Insertion \\ D_{i-1,j} \quad +1 \quad Deletion \\ D_{i-2.j-2} + c \quad Exchange \quad if \\ \quad A_i = B_{j-1} and\ A_{i-1} = B_j \end{cases} \tag{10}$$

$$for\ 3 \leq i \leq m, 3 \leq j \leq n.$$

5 Evaluation and Results

We have a data set of 30 different spectrometer signals. Each of the spectrometer signals have been preprocessed according to the methods described in Section II, and afterwards processed and coded based on the delta modulation (see Sect. III). The final outcome is a 0/1 sequence. The achieved results for the representation are presented in Section V.A.

We calculated the pairwise distances between the thirty signals based on three distance measures: Hamming distance, Levenshtein distance, and the Damerau-Levenshtein distance.

We used the single-linkage clustering method to evaluate the goodness of the measures in Section V.B.

5.1 Representation of the Spectrometer Signal by Delta-Modulation

The representation of the real signal by the approximated signal of the delta modulator is exemplary shown in Fig. 4 for Linear Delta Modulation and in Fig. 5 for Constant Factor Delta Modulation. The binary coded signal for both methods is shown in Table 2. It can be seen that the coded signal is different depending on the used delta modulation method. Table 3 shows the mean and maximum approximation error between the input signal and the approximated signal by the delta modulator. As expected the CFMD method shows the best result. The mean error is *1.677* increments and the maximum error is *16.04* increments. In the recent settings the CVSDM gave the worst results. It is left for further work to improve this method.

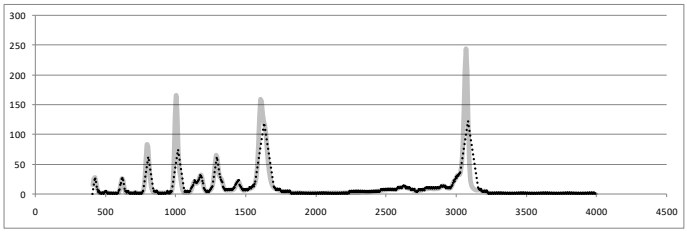

Fig. 4. Representation of Benzoic acid using LDM.

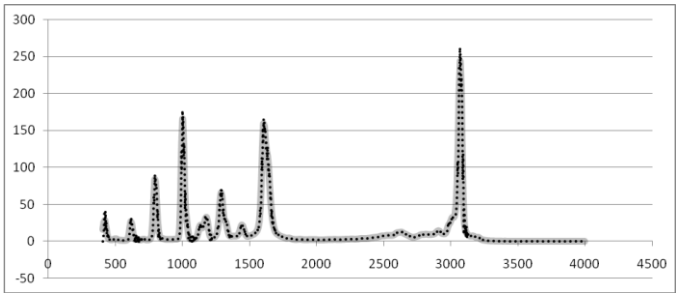

Fig. 5. Representation of Benzoic acid using CFDM.

Table 2. Binary Representation of the Spectrum of Benzoic acid.

Name of Compander	Sequence of Spectrum
LDM	... 1 0 0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 1 0 1 0 1 0 1 0 1 0 0 1 ...
CFDM	... 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 1 0 1 1 0 0 1 1 1 1 0 1 1 1 0 1 1 0 0 ...

Table 3. Mean and Maximum Approximation Error between Input Signal and approx. Signal.

Substance	Name of Delta Modulator			
	Linear Delta Modulator		CFDM	
	mean ε	max ε	mean ε	max ε
Acetone	1,74955958	4,642862	0,45178963	5,275991
Ascorbic acid	2,19114882	13,715031	1,06728145	10,356945
Benzamide	1,7514159	4,282954	0,40339027	2,977274
Benzoic acid	16,8602393	147,708368	4,78830105	45,56784
.				
mean	5,6380909	42,5873038	1,6776906	16,045

In this study, we chose the CFMD method for the representation of the spectrometer signal.

5.2 Results for Similarity Between Two Spectra

It has been shown in Section 5. A that the **CFDM** delta modulator gives the best result for calculating the 0/1 sequence of the signal. The dendrogram for the different similarity measures between the thirty different spectra are shown in Fig. 6-8. The Hamming distance shows the highest differences in similarity but does not represent the similar groups well (see Fig. 6). The similarity measure will be sensitive to small changes in the spectra that might be caused by noise. Much better are represented

similar groups in case of the Levenshtein (Fig. 7) and Damerau-Levenshtein similarity (Fig. 8). Both dendrograms show similarity in the group structure. They only slightly differ in the representation of the large group at the top of the dendrogram but in general the group structure is preserved.

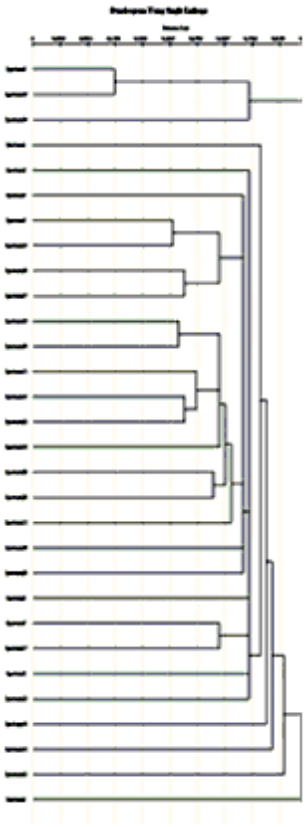

Fig. 6. Dendrogram using Hamming Distance using CFDM

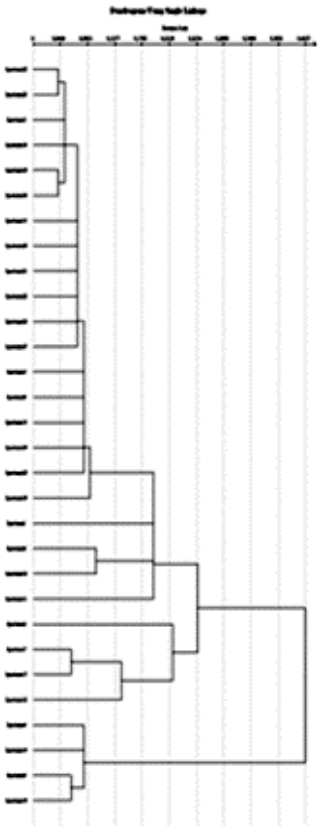

Fig. 7. Dendrogram using Levenshtein Distance using CFDM

5.3 Accuracy of the Classification

We enlarged the data base by ten samples from the same spectrum. The final data base consists of three hundred samples. Our prototype-based classifier PROTOCLASS [15] was used for classification where 299 samples were the prototypes and one sample was classified against the 299 samples by searching for the three nearest neighbors. Crossvalidation was used for calculating the error rate. The results are show in Table 4.

Table 4. Accuracy of prototype-based classifier for the different similarity measures.

Distance Measure	Accuracy in %
Hamming	85,2
Levenshtein	90,5
Damerau-Levenshtein	91,2

The best results we have got for the Damerau-Levenshtein distance followed by the Levenshtein distance. The worst result we have got for the Hamming distance.

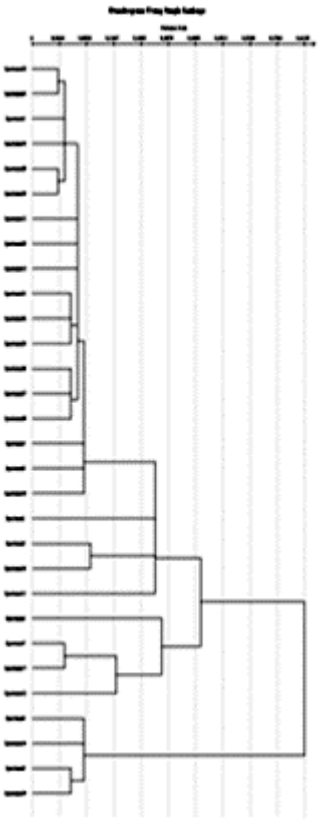

Fig. 8. Dendrogram using Damerau-Levenshtein Distance using CFDM.

6 Conclusion

The representation of the spectra by a 0/1 sequence is a good representation for a spectrometer signal. While coding the signal in a 0/1 sequence it also smoothing the signal by a step-wise linear function. To keep the approximation error between the original signal and the coded signal small an adaptive delta modulator has to be selected. In the experiment above we used the CFDM delta modulation method instead of the linear delta modulator. A better method than this might be the continuously variable slope delta modulator. To construct such a modulator for this kind of signals is left for further work.

Three different similarity measures have been used: Hamming distance, Levenshtein distance, and the Damerau- Levenshtein distance. While the Hamming distance is very fast and simple to calculated, the latter two distances seem to represent the similar groups of spectra´s very well. However, these two distance measures are more computationally expensive as the Hamming distance. The advantage of the Levenshtein and the Damerau-Levenshtein distance is that this distance can compare strings with different number of bits and since these measures can delete and substitute bits small differences in the sequence caused by noise or the behavior of the delta modulation can be eliminated. Finally, we tested the methods with our prototype-based classifier. We obtained good classification results for the Damerau-Levenshtein distance and the Levenshtein distance. The worst result we obtained for the Hamming distance.

In general we can say that the proposed novel method is a good method to represent spectrometer signals and that the similarity-based classification works very well. The proposed method allows us to extend our database of spectrometer signals very easily in the timely sequence the spectrometer signals occur and at the same time immediately to use the new acquired spectra for classification in daily work without going into a heavy update of the system parameters and functions.

We have tested it on data from RAMAN spectroscopy. However the method is not only applicable to RAMAN spectra. The method can be used for other spectra as well.

References

1. Janzen Chr, Delbrück H., Perner P., MARAS – Marker Free RAMAN Screening for Molecular Investigation of Biological Interactions, Project Report 2006.
2. Altose M. D., Zheng Y., Dong J., Palfey B. A., CareyP. R. “Comparing protein–ligand interactions in solution and single crystals by Raman spectroscopy”, Proceedings of the national Academy of Science, Vol. 98, 6, 3006 – 3011 (2001).
3. Rammal A., Perrin E., Chabbert B., Bertrand I., Mihai G., Vrabie V., Optimal Preprocessing of Mid InfraRed spectra. Application to classification of lignocellulosic biomass: maize roots and miscanthus internodes, In: P. Perner (Eds.) Advances in Mass

Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry, ibai-publishing 2013, p. 66-76, ISBN 978-3-942952-21-7.
4. Bleghith A., Collet Ch., Armspach J.-P., A Unified Framework for Peak Detection and Alignment: Application to HR-MAS 2D NMR Spectroscopy, In: P. Perner (Eds.) Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry, ibai-publishing 2011, p. 106-118, ISBN 978-3-942952-02-6.
5. Savitzky, A., Golay, M.J.E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 36(8), 1627-1639 (1964).
6. Zhao, J., Carrabba, M.M., Allen, F.S.: Automated Fluorescence Rejection Using Shifted Excitation Raman Difference Spectroscopy. Applied Spectroscopy 56(7), 834-845 (2002).
7. Un, C.K., Lee, H.S.: A Study of Comparative Performance of Adaptive Delta Modulation Systems, IEEE Trans. on Communications 28 (1), 96-101 (1980).
8. Jayant, N.S.: Adaptive Delta Modulation with one-bit Memory, The Bell System TechnicalJournal 49(3), 76-80 (1970).
9. Jayant, N.S.: Adaptive Delta Modulation with one-bit Memory, The Bell System Technical Journal 49(3), 76-80 (1970).
10. Tazaki, S., Osawa, H., Shigematsy, Y.: A Useful Analytical Method for Discrete Adaptive Delta Modulation. IEEE Trans. on Communications 25 (2), 195-199 (1977).
11. Perner, P. Datenreduktionsverfahren für technologische Industrierobotersteuerungen mit direkter Teach-in-Programmierung. 2nd, unrevised edition, ISBN: 978-3-940501-16-5, ibai-publishing, Leipzig (2010).
12. Hamming, R. W.: Error detecting and error correcting codes. Bell System Technical Journal29 (2), 147–160 (1950).
13. Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966).
14. Damerau, F.: A technique for computer detection and correction of spelling errors. Communications of the ACM 7(3), 171–176 (1964).
15. Perner, P., Prototype-Based Classification, Applied Intelligence 28(3): 238-246 (2008).