Transactions on Mass-Data Analysis
of Images and Signals
Vol. 3, No. 1 (2011) 3-14
© 2011, ibai-publishing
ISSN: 1868-6451
ISBN: 978-3-942952-10-1

ibai Publishing
www.ibai-publishing.org

A Novel Method for the Interpretation of Spectrometer Signals Based on Delta-Modulation and Similarity Determination

Petra Perner, Anja Attig, and Oleg Machno

Institute of Computer Vision and Applied Computer Sciences, IBaI
PSF 30 11 14, 04251 Leipzig
pperner@ibai-institut.de, www.ibai-institut.de

Abstract. We describe in this paper our novel method for automatic spectra identification. The method is based on a featureless representation. The representation calculated based on delta modulation is a 0/1 sequences which can be used for comparison between different spectras. Three different similarity measures, the Hamming distance, the Levenshtein distance, and the Damerau-Levenshtein distance, have been studied. The pros and cons of these three distances are discussed in this paper. Finally, we present a novel and powerful method for spectrometer interpretation.

1 Introduction

Our specific interest concerned the development of a novel methods for the automatically analyze of spectra. We introduced a novel technique to smooth and code the spectrometer signal in a featureless way so that it can be easily interpreted by a similarity-based method. By using a featureless representation we avoid the burden of the knowledge-acquistion and feature extraction problem. This technique makes it easy to store a huge number of signals without the need of a large memory. The software is based on an incremental learning procedure so that knowledge about the spectral characteristics of various for e.g. proteins and ligands can be learned over time and used for automatic interpretation.

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, we will explain the architecture of the automated spectrometer signal analysis system in Section 2. The novel smoothing and coding technique based on the Delta Modulation is described in Section 3. The similarity measures are presented in Section 4. A comparison of different representations based on different delta modulation methods and different similarity methods will be given in Section 5. Finally we summarize our work in Section 6.

2 Architecture of the Automated Spectrometer Signal Program

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

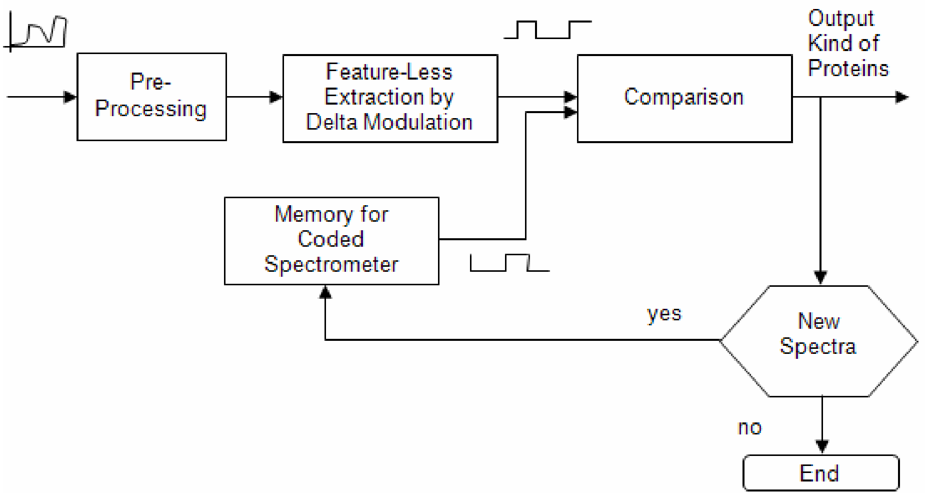
Pre-Processing
Feature-Less Extraction by Delta Modulation
Comparison
Output Kind of Proteins
Memory for Coded Spectrometer
yes
New Spectra
no
End

Fig. 1. Architecture of the Spectrum Interpretation System

After preprocessing the spectrometer signal, the signal is coded into a featureless representation by a delta modulator. The delta modulator codes the signal into a sequence of 1 and 0. While this is done, the signal is smoothed by a stepwise function. The featureless 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 analyzed for identity, similarity of the whole sequence, or for partial identity or similarity. This enables identification of part-spectra, special single peaks, or peak combination within a spectrum.

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

When no similar sequence is found in the data base, the input spectrum is stored in the data base after it has been coded by the delta modulator. This data collection is necessary since the appearance of the spectra for different proteins is not known yet.

In this special case, the pre-processing of the RAMAN spectra is a baseline correction [1], a Fourier transformation to eliminate the influence of the special system device and its parts [2], and the calculation of the difference between the spectrum of the buffer and the spectrum of the liquid in the buffer.

3 Featureless Representation of the Spectra by Delta Modulation

The delta modulator compares the actual signal value *s(i)* with an estimated signal value *r(i)* of the coder. This difference *e(i)* is coded by only one bit. Three different methods exist to estimate and predict the signal value: linear delta modulation (LDM) [3], constant factor delta modulation (CFDM) [4], and continuously variable slope delta modulator (CVSD) [5]. We study how these methods work on spectrometer signals and what kind of similarity measure is appropriate for the resulting representation of the spectrometer signal.

3.1 Linear Delta Modulation

In case of the linear delta modulation (LDM), the difference *e(i)* between the actual value *s(i)* and the estimated value *r(i)* at sampling point *i* is formed, see Fig. 2.

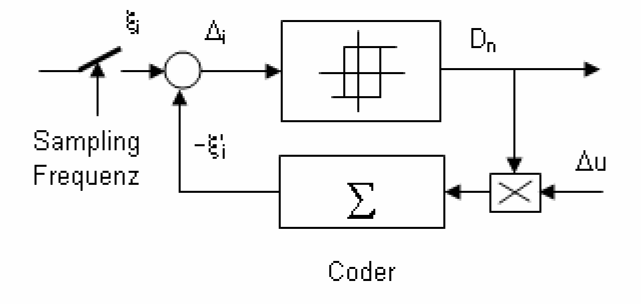

Fig. 2. Linear Delta Modulator

Coding is done only in the direction of the modification (*D*.. = “1” positive, *D*.. = “0” negative). This binary signal *D(i)* is stored in the memory. At the same time, the magnitude of the signal expected at the next sampling point *i* is estimated based on it. The corresponding rule is:

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

$$s(i) \le r(i) \ \ D_n = 1 \ \ r(i) = r(i-1) - \Delta u \tag{2}$$

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

For reproducing the signal (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 and the approximated signal is the approximation error $\varepsilon(i) = s(i) - s'(i)$, see Fig. 3.

When process dynamics change, the linear delta modulator is no longer optimally adjusted anymore and the reconstruction error increases 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.

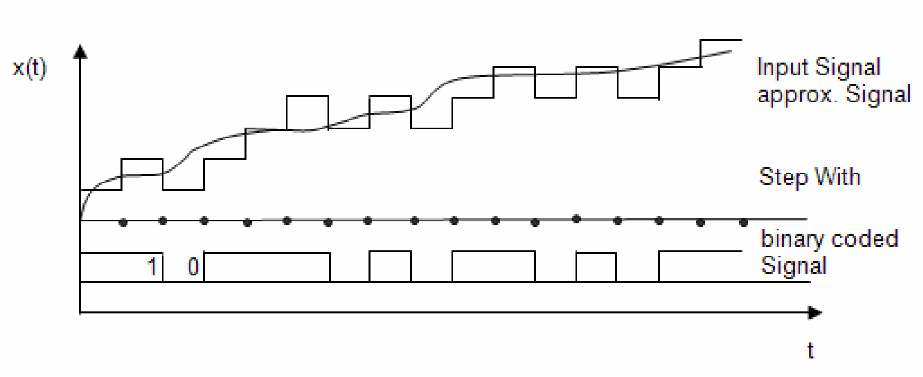

Fig. 3. Diagram with Input Signal, approximated Signal, and Binary Coded Signal

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 $\{D_n, D_{n-1}\}$ by which factor *k* the preceding increment size has to be multiplied:

$$\begin{aligned} \Delta u_i = \Delta u_{i-1} \cdot k \ \ D_n = D_{n-1} \ \ k = P \\ D_n \ne D_{n-1} \ \ k = Q \end{aligned} \tag{3}$$

A requirement is that $P \cdot Q = 1$ in order to observe the stability condition. For speech signals and road data of a robot [6] the values $P = 1.5$ and $Q = 0.66$ are known from the literature; according to our study these values also perform well in case of the spectrometer signal application.

3.3 Continuously Variable Slope Delta Modulator

The syllabic compander, also called continuously variable slope delta modulator (CSVDM), pursues, in contrast to the instantaneous value compander, the tendency of the signal course. Only when the same state has been recorded three times in a coincidence register $\{D_n = D_{n-1} = D_{n-2}\}$, the syllabic compander increases its increment size. It is therefore more inert than the instantaneous value compander. The rule for syllabic companding is:

3 bit coincidence $$k = 1 \;\; \Delta u_i = \Delta u_{i-1} + 1 \qquad (4)$$

no coincidence $$k = 0 \;\; \Delta u_i = \Delta u_{i-1} - 1 \text{ until } \Delta u_i = 0 \qquad (5)$$

As Δu must not become zero, a minimum increment size $u_{\min}$ larger than 1 must be added. A standard value $u_{\min} = \sqrt{\delta_{\Delta s}}$ can be assumed wherein $\delta_{\Delta s}$ is the variance of the first-order difference of the signal $\Delta s = s(i) - s(i-1)$.

4 Similarity Determination between the Featureless Spectra Representations

The spectra are represented by 1/0 sequences. To compare different spectra we need a distance measure that can work on such a representation. Different measures are known from text comparison and DNA sequence analysis. We choose for this work the Hamming distance [7], the Levenshtein distance [8], and the Levenshtein-Damerau distance [9] and compare the performance of these measures on a data set of Raman spectrometer signals.

4.1 Hamming Distance

Table 1. Representation of a Spectrum and the Coincidence Bits

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

The representation of a spectrum A and spectrum B is illustrated in Table 1.

We assume that all spectra have the same length *i* and that the peaks are stable at there position (wavelength) in the spectra.

Then we have to compare two binary vectors *A* and *B*. The distance *d* between these two binary representations is the number of bits in which the two vectors differ. This is the well-known Hamming Distance:

$$d(A,B) = \|A-B\| = \sum_{i=1}^{n} |A_i - B_i| \tag{6}$$

4.2 Levenshtein Distance

Let $d_L(A,B) = D_{m,n} / n$ be the Levenshtein-Distance between the binary vectors A and B with $m = |A|$ and $n = |B|$. The Levenshtein Distance between the binary vectors A and B is defined as the minimum of modifications needed to transform vector A in vector B. The allowed operations are substitution, inseration or delation of a single signal value.

Then the similarity between A and *B* is computed as follows:

$$D_{0,0} = 0$$

$$D_{i,0} = i \text{ for } 1 \le i \le m$$

$$D_{j,0} = j \text{ for } 1 \le j \le n$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} + 0 & \text{if } A_i = B_j \\ D_{i-1,j-1} + 1 & \text{(Substitution)} \\ D_{i,j-1} + 1 & \text{(Insertion)} \\ D_{i-1,j} + 1 & \text{(Deletion)} \end{cases} \quad \text{for } 1 \le i \le m, 1 \le j \le n \tag{7}$$

4.3 Damerau-Levenshtein Distance

Let $d_L(A,B) = DL_{m,n} / n$ be the Damerau-Levenshtein Distance between the binary vectors A and B with $m = |A|$ and $n = |B|$. The Damerau-Levenshtein Distance between the binary vectors A and B is defined as the minimum of modifications needed to transform vector A in vector B. The allowed operations are substitution, inseration or delation of a single signal value or the transposition of two adjacent single signal value. The Damerau-Levenshtein Distance in compared to the Levenshtein Distancs has the translation as additional operation.

The matrix is compute as follows:

$$D_{0,0} = 0$$

$$D_{i,0} = i \text{ for } 1 \le i \le m$$

$$D_{j,0} = j \text{ for } 1 \le j \le n$$

$$DL_{i,j} = \min \begin{cases} D_{i-1,j-1} + 0 & \text{if } A_i = B_j \\ D_{i-1,j-1} + 1 & \text{(Substitution)} \\ D_{i,j-1} + 1 & \text{(Insertion)} \\ D_{i-1,j} + 1 & \text{(Deletion)} \end{cases} \text{for} \begin{array}{l} (1 \le i \le 2, 1 \le j \le n) \\ \vee (1 \le i \le m, 1 \le j \le 2) \end{array} \tag{8}$$

$$DL_{i,j} = \min \begin{cases} D_{i-1,j-1} + 0 & \text{if } A_i = B_j \\ D_{i-1,j-1} + 1 & \text{(Substitution)} \\ D_{i,j-1} + 1 & \text{(Insertion)} \\ D_{i-1,j} + 1 & \text{(Deletion)} \\ D_{i-2,j-2} + 1 & \text{if } \mathrm{A}_i = B_{j-1} \wedge A_{i-1} = B_j \\ & \text{(Transpositon)} \end{cases} \tag{9}$$

$$\text{for } 3 \le i \le m, 3 \le j \le n$$

5 Evaluation and Results

Based on a data set of 30 spectrometer signals, each of the spectrometer signals has been processed and coded by delta modulation. The achieved results are presented in Section 5.1.

The final outcome is a string of 0/1 bits. We first calculated the three distance measures: Hamming distance, Levenshtein distance, and the Damerau-Levenshtein distance.

We evaluated our method by calculating the pairwise distance between the spectra and used the single-linkage clustering method to show the distance between the spectra based on the similarity measures.

5.1 Representation of the Spectrometer Signal by Delta Modulation

An exemplary representation of the real signal by the approximated signal of the delta modulator is 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. The CVSDM has not been tested yet. Future work will address this method.

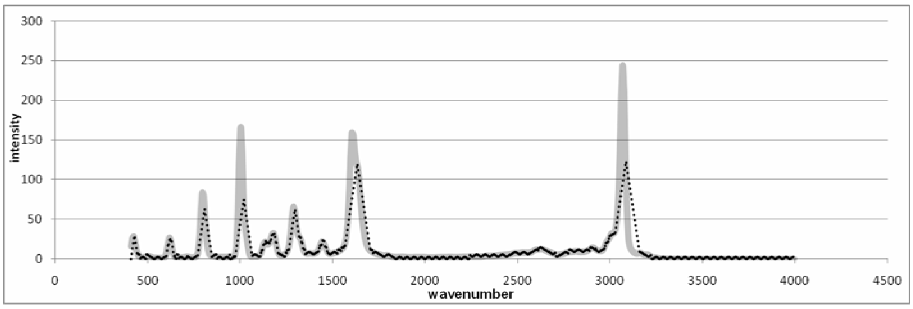

Fig. 4. Representation of Benzoic Acid Using LDM

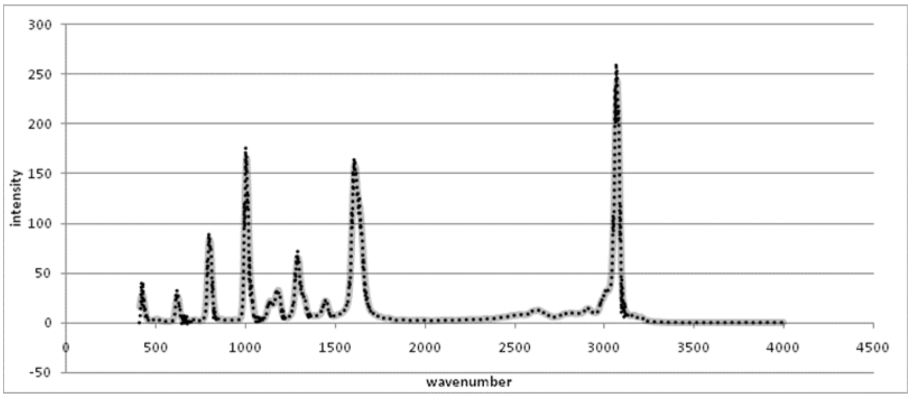

Fig. 5. Representation of Benzoic Acid Using CFDM

Table 2 . Binary Representation of the Spectrum of Benzoic Acid

LDM	… 1 0 0 1 0 1 0 0 1 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 1 0 0 1 0 1 0 …
CFDM	… 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 1 0 0 1 0 0 0 1 0 0 0 0 …

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

	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

5.2 Results for Similarity between the Spectra

In Section 5.1, we have demonstrated that the **CFDM** delta modulator gives the best result for calculating the featureless representation of the signal. The results for the different similarity measures between the 30 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). A much better representation of the similar groups is provided by the Levenshtein (Fig. 7) and Damerau-Levenshtein similarity (Fig. 8).

6 Conclusion

The featureless 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, the signal is also smoothed by a stepwise 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 presented above, we used the CFDM delta modulation method instead of the linear delta modulator. An even better method might be the continuously variable slope delta modulator. Constructing such a modulator for this kind of signals is the aim of 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 calculate, the latter two distances seem to represent the similar groups of spectras very well. However, these two distance measures are more computationally complex than the Hamming distance. The advantage of the Levenshtein and the Damerau-Levenshtein distances is that these distances can compare strings with different numbers of bits and can delete and substitute bits; it is therefore possible to eliminate small differences in the sequence caused by noise or by the behavior of the delta modulation.

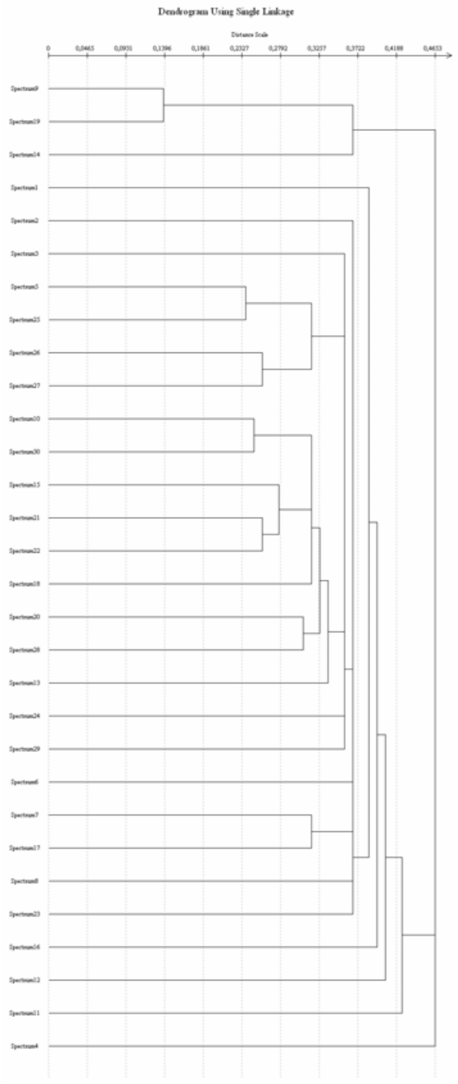

Fig. 6. Dendrogram Using Hamming Distance in CFDM

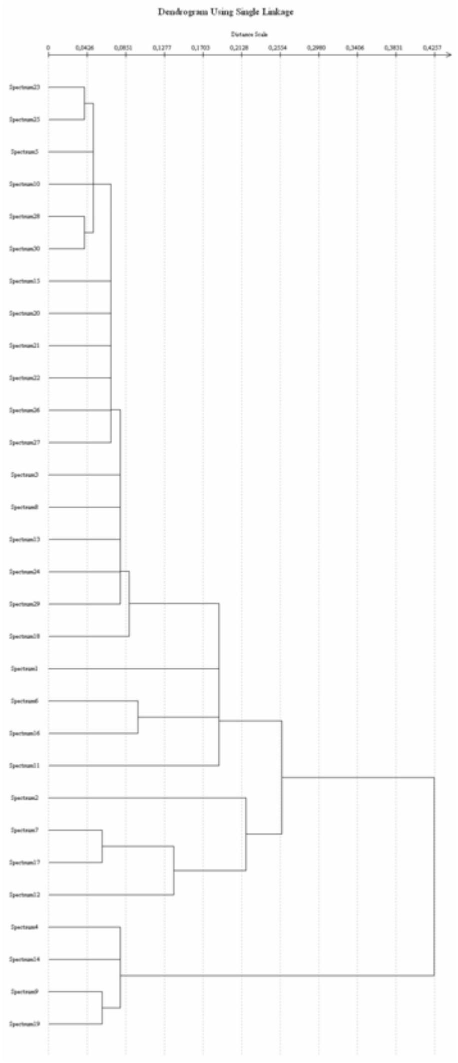

Fig. 7. Dendrogram Using Levershtein Distance in CFDM

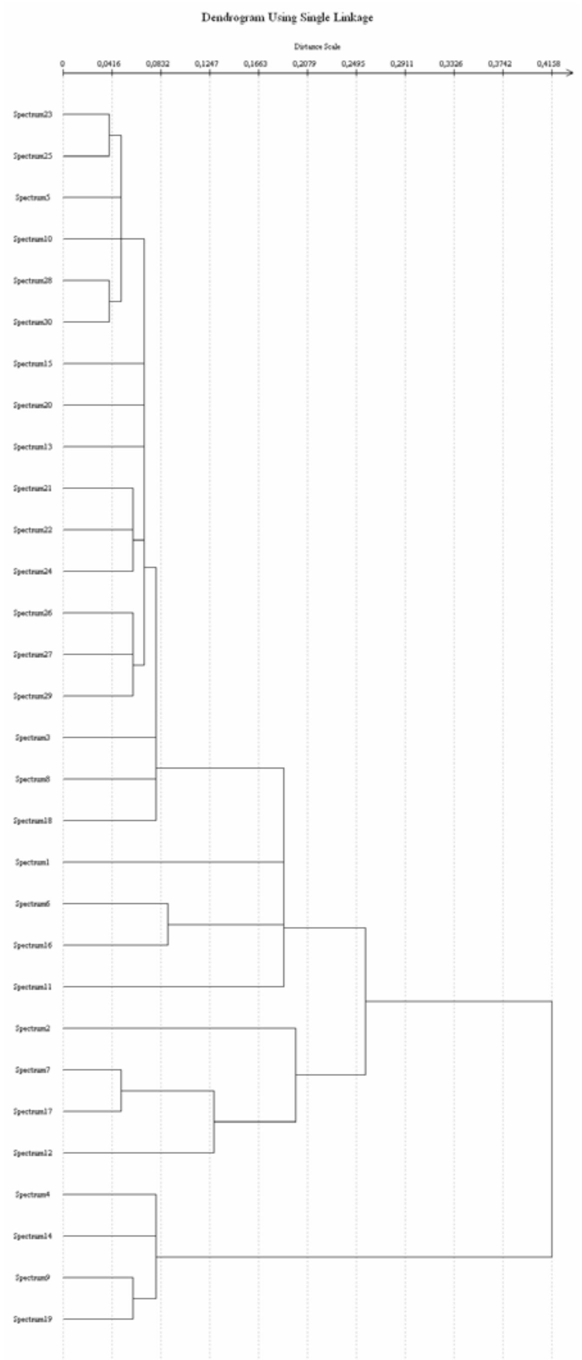

Fig. 8. Dendrogram Using Damerau-Levershtein Distance in CFDM

References

1. Savitzky, A., Golay, M.J.E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 36(8), 1627-1639 (1964)
2. Zhao, J., Carrabba, M.M., Allen, F.S.: Automated Fluorescence Rejection Using Shifted Excitation Raman Difference Spectroscopy. Applied Spectroscopy 56(7), 834-845 (2002)
3. 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)
4. Jayant, N.S.: Adaptive Delta Modulation with one-bit Memory, The Bell System Technical Journal 49(3), 76-80 (1970)
5. Jayant, N.S.: Adaptive Delta Modulation with one-bit Memory, The Bell System Technical Journal 49(3), 76-80 (1970)
6. Tazaki, S., Osawa, H., Shigematsy, Y.: A Useful Analytical Method for Discrete Adaptive Delta Modulation. IEEE Trans. on Communications 25 (2), 195-199 (1977)
7. Perner, P. Datenreduktionsverfahren für technologische Industrierobotersteuerungen mit direkter Teach-in-Programmierung. 2nd, unrevised edition, ISBN: 978-3-940501-16-5, ibai-publishing, Leipzig (2010)
8. Hamming, R. W.: Error detecting and error correcting codes. Bell System Technical Journal 29 (2), 147–160 (1950)
9. Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
10. Damerau, F.: A technique for computer detection and correction of spelling errors. Communications of the ACM 7(3), 171–176 (1964)