Designation: E 2310 – 04

Standard Guide for Use of Spectral Searching by Curve Matching Algorithms with Data Recorded Using Mid-infrared Spectroscopy[1]

This standard is issued under the fixed designation E 2310; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 Spectral searching is the process whereby a spectrum of an unknown material is evaluated against a library (database) of digitally recorded reference spectra. The purpose of this evaluation is classification of the unknown and, where possible, identification of the unknown. Spectral searching is intended as a screening method to assist the analyst and is not an absolute identification technique. Spectral searching is not intended to replace an expert in infrared spectroscopy. Spectral searching should not be used without suitable training.

1.2 The user of this document should be aware that the results of a spectral search can be affected by the following factors described in Section 5: *(1)* Baselines, *(2)* sample purity, *(3)* Absorbance linearity (Beer's Law), *(4)* sample thickness, *(5)* sample technique and preparation, *(6)* physical state of the sample, *(7)* wavenumber range, *(8)* spectral resolution, and *(9)* choice of algorithm.

1.2.1 Many other factors can affect spectral searching results.

1.3 The scope of this document is to provide a guide for the use of search algorithms for mid-infrared spectroscopy. The methods described herein may be applicable to the use of these algorithms for other types of spectroscopic data, but each type of data search should be assessed separately.

1.4 The Euclidean distance algorithm and the first derivative Euclidean distance algorithm are described and their use discussed. The theory and common assumptions made when using search algorithms are also discussed, along with guidelines for the use and interpretation of the search results.

2. Referenced Documents

2.1 *ASTM Standards:* [2]

E 131 Terminology Relating to Molecular Spectroscopy

E 334 Practice for General Techniques of Infrared Microanalysis

E 573 Practices for Internal Reflectance Spectroscopy

E 1252 Practice for General Techniques of Qualitative Infrared Analysis

E 1642 Practice for General Techniques of Gas Chromatography Infrared (GC/IR) Analysis

E 2105 Practice for General Techniques of Thermogravimetric Analysis (TGA) Coupled with Infrared Analysis (TGA/IR)

E 2106 Practice for General Techniques of Liquid Chromatography—Infrared (LC/IR) and Size Exclusion Chromatography—Infrared (SEC/IR)

3. Terminology

3.1 *Definitions*—For general definitions of terms and symbols, refer to Terminology E 131.

3.1.1 *reference spectrum*—an established spectrum of a known compound or chemical sample.

3.1.1.1 *Discussion*—This spectrum is typically stored in retrievable format so that it may be compared against the sample spectrum of an analyte.

3.1.1.2 *Discussion*—This term has sometimes been used to refer to a background spectrum; such usage is not recommended.

3.1.2 *spectral searching*—the process whereby a spectrum of an unknown material is evaluated against a library of digital reference spectra. Each reference spectrum in the library is individually compared to the spectrum of the unknown, and assigned a numerical value as to the goodness of fit. To perform this comparison, each data point in the unknown spectrum is compared to each corresponding point in the reference spectrum.

3.1.3 *peak searching*—the process whereby the peak table of the spectrum of an unknown material is evaluated against a library of peak tables. Each reference spectrum in the library contains a peak table and the peak table is individually compared to the peak table of the unknown, and assigned a numerical value as to the goodness of fit.

3.1.4 *spectral library*—a collection of reference spectra stored in a computer readable form, also called a library, database, or spectral database.

[1] This guide is under the jurisdiction of ASTM Committee E13 on Molecular Spectroscopy and is the direct responsibility of Subcommittee E13.03 on Infrared Spectroscopy.

Current edition approved Feb. 1, 2004. Published Feb. 2004.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

3.1.5 *search algorithm*—the mathematical formula used to make a point-by-point comparison of two spectra.

3.1.6 *hit quality value*—the spectral search software compares each spectrum in the database to that of the unknown, and assigns a numeric value for each library entry demonstrating how similar the two spectra are.

3.1.6.1 *Discussion*—There are several methods for assigning Hit Quality values and either a high or low value can be assigned as the best match. Refer to the software manufacturers documentation.

3.1.7 *hit quality index (HQI)*—a table which ranks the library spectra in the database according to their Hit Quality values (see 7.5).

3.1.8 *Euclidean Distance algorithm*—the Euclidean Distance algorithm measures the Euclidean distance between each library spectrum and the unknown spectrum by treating the spectra as normalized vectors. The closeness of the match, or, HQI, is calculated from the square root of the sum of the squares of the difference between the vectors for the unknown spectrum and each library spectrum.

3.1.9 *First Derivative Euclidean Distance algorithm*—in the First Derivative Euclidean Distance algorithm the Euclidean distance is also computed, except the derivative of each spectrum is calculated prior to the Euclidean distance calculation.

3.1.10 *normalization*—the mathematical technique used to compensate for an intensity difference between two spectra (see 5.1).

4. Theory

4.1 *Beer's Law*—One of the basic principles that make spectral searching possible is Beer's Law (see Terminology E 131), which states that $A = abc$, where A is the absorbance, a is the absorptivity, b is the sample pathlength, and c is the concentration of the analyte of interest. As long as Beer's Law applies, two spectra of the same material recorded under similar conditions can be made to appear the same by normalization of the data.

NOTE 1—In an ideal case, this is true for transmittance spectra, but there are differences in the spectral peak intensities when reflectance spectra are compared to transmittance spectra.

5. Spectral Data Pre-Treatment

5.1 *Normalization*:

5.1.1 Normalization of spectra compensates for the differences in sample quantity (concentration or pathlength, or both) used to generate the reference spectra in the library and that of the unknown. The spectra are normalized over the complete spectral range of the library. When searching less than the full spectral range of the library, the spectra must be re-normalized over the new range before an accurate comparison can be made. Normalization of a spectrum for library searching is a two step process. First, the minimum absorbance value in the selected spectral range is subtracted from all the absorbances in the same range. The resulting values are then scaled by dividing by the maximum result value in the range. The end result is a spectrum (or a sub-range portion of a spectrum) where the minimum value is zero (0) and the maximum is one (1) absorbance. If the range chosen for normalization has only one or two strong bands and a few medium intensity bands, the range of the spectrum must be reselected or the spectrum will be dominated by the strong bands in the spectrum and the HQI will be insensitive to weaker fingerprint bands necessary for identification of a specific compound. Successful compound identification may require the spectral match exclude the strongest bands, then the normalization will be based on a medium intensity band, and weak fingerprint bands will be emphasized in the HQI.

5.2 *Data Point Matching*:

5.2.1 The algorithms used for searching a spectrum against a library use a calculation that mathematically compares the data points of the spectrum being searched to the data points of the spectra in the library. This requires that the data points in both the sample and library spectra occur at the same frequency. If the data points in the sample and library spectra are not aligned in this manner, then one of the spectra must be mathematically altered (interpolated) to make the data points match. Typically the unknown spectrum being searched is altered to match the data point spacing of the spectra in the library.

5.2.2 Data point matching is commonly accomplished using a linear data point interpolation method. In this method, the slope and offset of a line segment is calculated between the absorbances of every pair of data points in the spectrum. A new set of absorbances is calculated by locating the values that occur on the line segments at positions corresponding to the datapoint frequency of the library spectrum.

6. Conditions or Issues Affecting Results

6.1 Spectral quality is one of the primary conditions or issues that can affect search results. There is no substitute for a carefully recorded spectrum. There are several conditions or issues that affect spectral quality as pertains to spectral searching. These conditions or issues apply to both the spectra used to create the reference database and to the unknown spectrum.

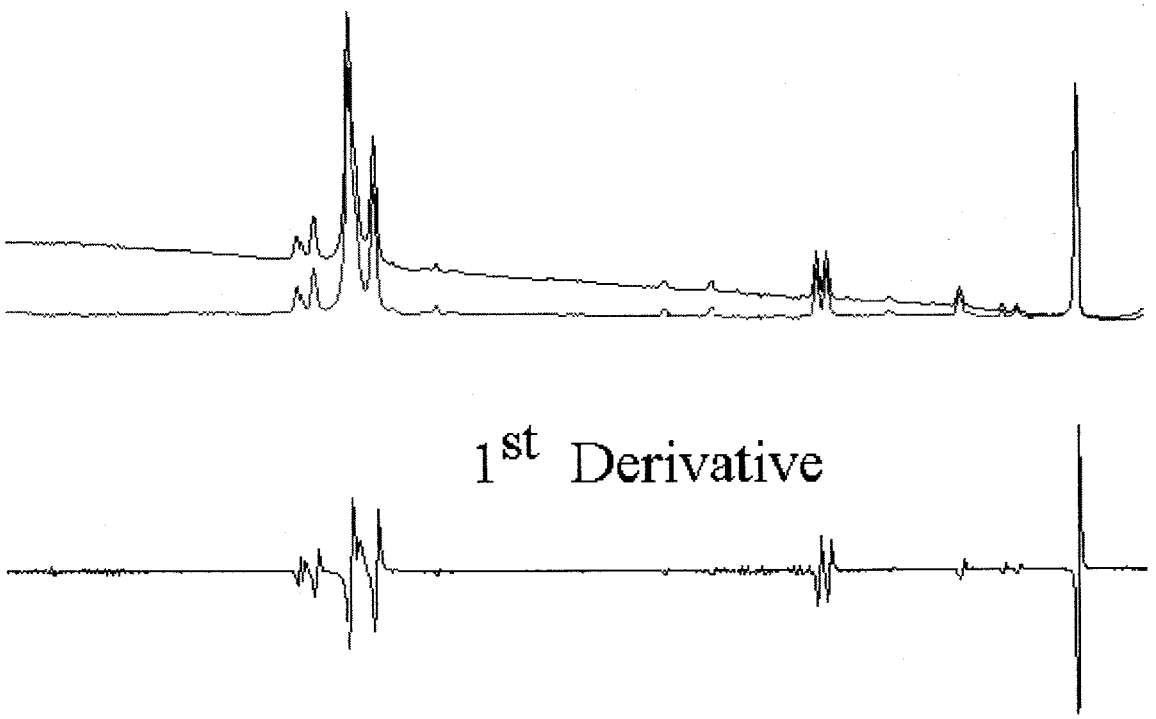
6.2 *Baselines*:

6.2.1 A flat baseline is preferred for the Euclidean distance algorithm as the Euclidean distance algorithm compares each data point in the unknown spectrum to the corresponding data point in the reference spectrum. The effect of an offset or slope in the baseline is interpreted as a difference between the two spectra. Therefore, when a spectrum with a sloping baseline or offset is evaluated using the Euclidean distance algorithm, a simple baseline correction should be used.

NOTE 2—Negative bands can also produce an offset in the baseline as a result of the data normalization process.

6.2.2 The first derivative Euclidean distance algorithm minimizes the effect of an offset or sloping baseline. In this algorithm, the comparison is made between the difference of a pair of adjacent points in the unknown spectrum to the difference between the corresponding pair of adjacent points in the reference spectrum. In effect, this causes the first derivative Euclidean distance algorithm to look only at the differences in the slope of adjacent data points between the two spectra. Fig. 1 shows how the two algorithms view the same two spectra.

NOTE 3—The first derivative algorithm converts a sloping baseline into an offset that is then eliminated by the normalization procedure.



The bottom two spectra demonstrate the results of the 1st derivative of a spectrum with a sloping baseline as compared to a spectrum with a flat baseline. The two spectra in the bottom trace are almost completely overlapped.

FIG. 1

6.3 *Sample Purity*:

6.3.1 The physical state of the sample should be as close as possible to the physical state of the reference materials used to obtain the library. For example, a pure liquid sample would ideally be searched against a library of spectra of only liquid reference materials. A sample which is probably a mixture, such as a commercial formulation, should be compared to a library of commercial formulations.

6.3.2 In some cases the nature of the sample may not be well understood. An unknown sample may be a pure material or a mixture. It may have additional contaminants that will affect its spectrum by adding spurious bands. In addition there are several other sources of spurious spectral features that may appear as either positive or negative bands. Several of these are listed below:[3]

6.3.2.1 Features due to variations in the carbon dioxide or water vapor levels in the optical path,

6.3.2.2 Bands from a mulling agent,

6.3.2.3 Halide salts used as window material and as the diluent for both pellets and diffuse reflection analysis often contain contaminants such as adsorbed water, hydrocarbon and nitrates. Always use dry halide salts and keep unused halide salts in a desiccator,

6.3.2.4 Water can alter the spectrum of the sample from its dry state. Spectra of inorganic samples with waters of hydration are particularly sensitive to adsorbed water,

6.3.2.5 Solvent bands from samples run in solution, and

6.3.2.6 Bands from solvents left over from an extraction or from casting a film from a solution.

NOTE 4—Retain spectra of any solvents used, so that bands due to the solvent can be identified in the spectrum of the unknown.

NOTE 5—If the solvent bands in a region of the spectrum cannot be removed from the spectrum (by either re-recording the spectrum, using an uncontaminated sample, or by spectral subtraction using the solvent reference spectrum), then that region of the spectrum should be excluded during a search. It is not sufficient to remove the offending bands digitally by drawing a straight line through the region before the search. The search algorithm will calculate a poor match in this region for any reference spectrum containing features in the region. It should be realized that the removal of the solvent bands may also remove underlying features in the sample spectrum.

6.4 *Absorbance Linearity (Beers Law)*:

6.4.1 A spectrum recorded using good practices (see Practices E 334, E 1252, E 1642, E 2105, and E 2106) should follow Beer's Law, and so maintain the relative absorbance intensities of its bands, independently of sample thickness. As long as this ratio between the bands is maintained, the spectra can be normalized and a good comparison between spectra can be made. For a spectrum to meet this requirement, each ray of light of a given frequency must pass through the same amount of sample. There are at least two general cases where this may not happen.

6.4.1.1 One case occurs when there is an uneven thickness of sample in the beam. For example, if the sample is wedge shaped in thickness, or irregular in shape, some rays of light pass through the thin part and some rays pass through the

[3] Coleman, Patricia B., *Practical Sample Techniques for Infrared Analysis*, CRC Press, FSBN# 0849342031: 8/26/93.

thicker part of the wedge. A similar concern arises when making KBr pellets for analysis. Unless the powder is carefully spread in the pellet die, the pellet can be pressed with a density gradient across the diameter. The sample must also be evenly distributed by thorough mixing of the sample and pellet matrix. This is of particular concern when the beam geometry is smaller than the sample diameter, and is a common problem when using a beam condensing accessory or an infrared transmitting microscope.

6.4.1.2 A second case is when the sample does not completely cover the entire beam cross-section. This occurs with a film that has a void in it, or when a spectrum of a liquid is recorded with an air bubble present in the sample. Some rays of light pass through the sample and some rays pass through the void. The net effect is that fewer molecules are measured than if all the rays passed though the sample resulting in a distortion of the observed relative band intensities.

6.4.2 In the example shown in Fig. 2, when there is no void in the sample and all of the rays pass through an equal thickness of sample, the ratio of intensities between bands A and B is 2.33. When half of the rays pass through the sample and half of the rays pass though a void with no sample, the ratio of bands A and B changes to 1.78. See Figs. 2-5.

6.5 *Sample Thickness*:

6.5.1 A spectrum which is acquired with too much sample cannot be properly normalized. When a sample band absorbs all of the available energy at a given frequency, this produces a transmittance value of zero. The resulting Absorbance value is infinity, and the normalization step becomes meaningless. In addition, the relative band intensities become highly distorted when normalized to the infinitely absorbing band.

6.5.2 A sample can also be run at a thickness that exceeds the linear range of the detector. Each detector is only linear to some absorbance value. When this is exceeded, the bands at low and high absorbance values will no longer maintain their ratios. As a guide, when practical, ensure that the strongest band of interest has an absorbance of no more than 1.0 absorbance units.

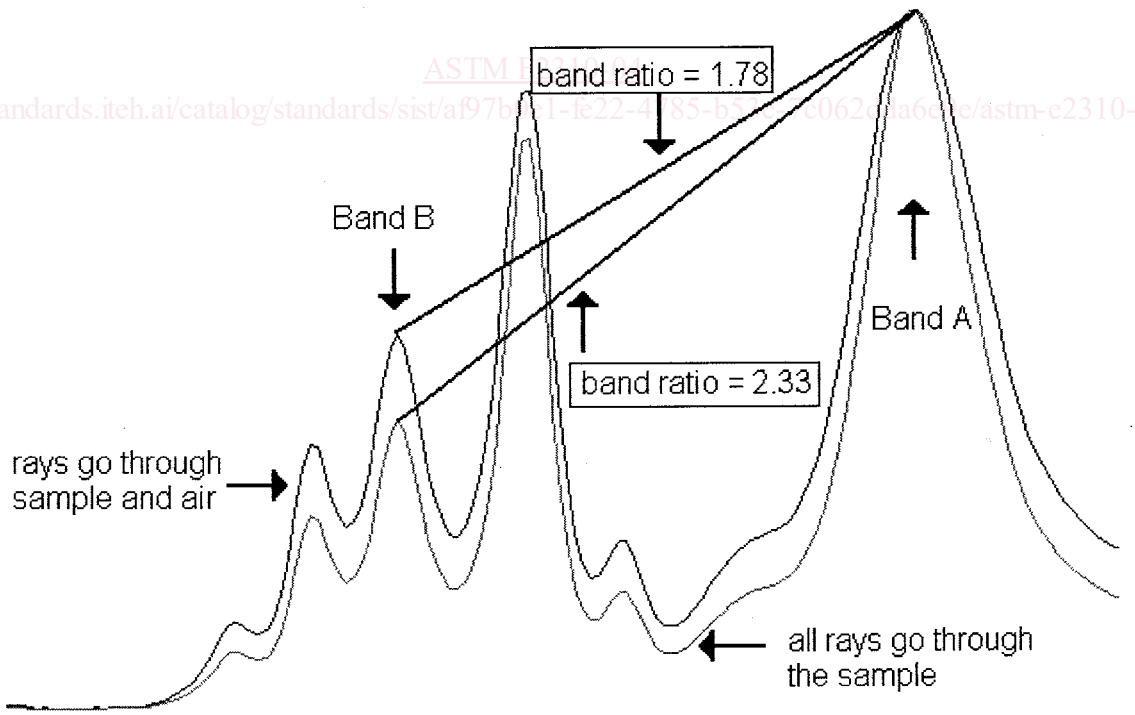
6.5.3 A spectrum that has too great an absorbance can often be detected by the presence of an apparently "flat" region associated with a peak. Often the rest of the spectral bands will still maintain their correct relative intensities. Therefore, eliminating the offending region from a search can produce better results.

6.5.4 Care should be taken not to use regions of the spectrum that are measured at frequencies outside the effective measurement range of the optical components installed in the spectrometer.

6.6 *Sampling Technique and Preparation*:

6.6.1 The analyst should be aware that sample preparation methods can alter the sample or the resulting spectrum. In addition, the measurement technique (optical accessory used) can have an effect on the spectrum as well. These effects can be very minor and have no impact on the search results, or they may be drastic enough to preclude a good match for a compound that is in the database. See Section 7.

6.6.2 The most severe effects will occur when trying to compare transmittance spectra, such as a KBr pellet, with



The ratio between bands A and B is different for a spectrum in which all of the rays pass through the same sample thickness compared to a spectrum where half the rays pass though air and half the rays pass though the sample.

The two spectra in Fig. 2 have been normalized to give a maximum of 1.0 absorbance at the strongest band, band A.

FIG. 2