Journal of Research of the National Bureau of Standards

Volume 90 **Number 5** **September-October**

ISSN 0160-1741 **Library of Congress Catalog Card No.: 63-37059**

The Journal of Research of the National Bureau of Standards features advances in measurement methodology and analyses consistent with the NBS responsibility as the nation's measurement science laboratory. It includes reports on instrumentation for making accurate and precise measurements in fields of physical science and engineering, as well as the mathematical models of phenomena which enable the predictive determination of information in regions where measurements may be absent. Papers on critical data, calibration techniques, quality assurance programs, and well characterized reference materials reflect NBS programs in these areas. Special issues of the Journal are devoted to invited papers in a particular field of measurement science. Occasional survey articles and conference reports appear on topics related to the Bureau's technical and scientific programs.

Issued six times a year. Annual subscriptions: domestic $17.00; foreign $21.25. Single copy, $3.00 domestic; $3.75 foreign.

United States Government Printing Office, Washington: 1985
Order all publications from the Superintendent of Documents
U.S. Government Printing Office, Washington, DC 20402

Journal of Research of the National Bureau of Standards
Volume 90, Number 5, September-October 1985

Note on Weighings Carried Out on the NBS-2 Balance

P. Carré
Bureau International des Poids et Mesures, Pavillon de Breteuil, F-92310 Sèvres, France

and

R.S. Davis
National Bureau of Standards, Gaithersburg, MD 20899

Accepted: August 28, 1985

The NBS-2 balance was designed and built at NBS and transferred to the BIPM in 1972. It is presently used for the comparison of national prototype kilograms with international standards. Excellent environmental conditions at the BIPM have resulted in a long-term standard deviation of 1 microgram (1×10^{-9}) for a comparison of two 1-kilogram standards. With this remarkable precision, one has begun to observe and quantify systematic biases of less than 5 micrograms. The nature of these biases is presented as well as the remedy adopted to eliminate their influence on both the final measurement results and the variance assigned to those results.

Key words: balance; errors; kilogram comparator; precision weighing; single-pan balance; weighing schemes.

The NBS-2 balance [1][1] permits mass intercomparisons of standards of nominal value $m_0=1$ kg by a substitution method. The features of this balance which are important to the present analysis are: 1) the balance has a single pan so that substitution weighing must be used; 2) the on-scale range of the balance is limited to about 40 mg; 3) a small "sensitivity weight" can be added or removed from the balance pan by remote control, thereby provided a means of calibrating the scale of the balance in mass units; 4) six weights can be placed on a table within the balance enclosure. The weight table may be raised, lowered, or rotated by remote control, combinations of these operations permitting any of the six weights to be placed on the balance pan; and 5) during the course of a day's measurements, the balance knives are kept in contact with their corresponding flats at full load. After measurements are completed, the balance is fully arrested (i.e., knife-flat contact is broken) until measurements are recommenced the next day.

In what follows, we use the term *weighing* to denote an ensemble of operations carried out successively and decomposable to a set number of elementary operations called *subweighings* in the course of which one determines the mass difference between two standards.

1. Subweighing

Let A and B be the standards, having masses m_A and m_B, which are used in a subweighing; let S be the sensitivity weight, of known mass m_S, which is used to determine the sensitivity of the balance.

One usually works with five operations (each yielding one position of balance equilibrium). The five operations consist of the balance being successively loaded with A, B, B and S, A and S, and finally A; one notes the corresponding positions of equilibrium, that is to say the readings λ_1, λ_2, λ_3, λ_4, and λ_5 taken from the balance scale.

Let us introduce the notation λ_0 to designate the reading which one would obtain if the balance were loaded with a standard of mass exactly equal to m_0. Because of continuous

---

About the Authors: P. Carré is a principal research physicist in the Bureau International des Poids et Mesures, and R.S. Davis is a physicist in the Length and Mass Division of NBS' Center for Basic Standards.

[1] Figures in brackets indicate literature references.

variations in ambient conditions, successive readings on a single weight can shift from one balance equilibrium to the next. For the moment, we assume that the balance readings have no random errors associated with them. To conserve *notational symmetry, we adopt the convention that* λ_0 corresponds to the third equilibrium observation and we introduce the following *drifts* from this reading:

d_2 between the first and second observation,
d_1 between the second and third,
d_1' between the third and fourth, and
d_2' between the fourth and fifth.

The readings which one would make with a mass standard exactly equal to m_0 would thus be, in succession,

$$\lambda_0-d_1-d_2,\ \lambda_0-d_1,\ \lambda_0,\ \lambda_0+d_1' \text{ and } \lambda_0+d_1'+d_2' \ .$$

Let a and b be the difference in readings corresponding to m_A-m_0 and m_B-m_0, respectively; let s be the difference in readings corresponding to m_S.

The five equilibrium positions are described by five equations:

Object on Pan	**Equations**
A	$\lambda_0-d_1-d_2+a=\lambda_1$
B	$\lambda_0-d_1+b=\lambda_2$
B and S	$\lambda_0+b+s=\lambda_3$
A and S	$\lambda_0+d_1'+a+s=\lambda_4$
A	$\lambda_0+d_1'+d_2'+a=\lambda_5$.

At this point, we recognize that the actual balance readings $\lambda_1, \ldots, \lambda_5$ are subject to random errors. One sees that λ_0 is always associated with either a or b, so that it is preferable to write

$$(\lambda_0+a)-d_1-d_2=\lambda_1$$

$$(\lambda_0+b)-d_1=\lambda_2$$

$$(\lambda_0+b)+s=\lambda_3$$

$$(\lambda_0+a)+s+d_1'=\lambda_4$$

$$(\lambda_0+a)+d_1'+d_2'=\lambda_5 \ .$$

To solve this system (5 equations, 7 unknowns) it is necessary to reduce the number of unknowns. The only way to accomplish this is to make hypotheses about the drift.

1.1 Classical Hypothesis of Drift: One postulates that the drift is the same during the 1st and the 3rd intervals between balance equilibria, so that $d_2=d_1'$, and that it is also the same for the 2nd and the 4th intervals between equilibria, so that $d_1=d_2'$.

The justification for this hypothesis could be that in the course of the 1st and 3rd interval there is an exchange of standards A and B, hence rotation of the weight table; whereas, in the course of the 2nd and 4th there is only manipulation of the sensitivity weight, S.

One is thus led to a system having five unknowns:

$$(\lambda_0+a)-d_1-d_1'=\lambda_1$$

$$(\lambda_0+b)-d_1=\lambda_2$$

$$(\lambda_0+b)+s=\lambda_3$$

$$(\lambda_0+a)+s+d_1'=\lambda_4$$

$$(\lambda_0+a)+d_1+d_1'=\lambda_5 \ .$$

The solution is:

$$(\lambda_0+a)=(\lambda_1+\lambda_5)/2$$

$$(\lambda_0+b)=(\lambda_2+\lambda_3-\lambda_4+\lambda_5)/2$$

$$s=(-\lambda_2+\lambda_3+\lambda_4-\lambda_5)/2$$

$$d_1=(-\lambda_2+\lambda_3-\lambda_4+\lambda_5)/2$$

$$d_1'=(-\lambda_1+\lambda_2-\lambda_3+\lambda_4)/2$$

By postulating that, after accounting for drift, the remaining *variations of balance readings are proportional to differences of mass*, one has

$$\frac{m_A-m_B}{m_S}=\frac{a-b}{s}=\frac{(\lambda_0+a)-(\lambda_0+b)}{s} \ .$$

Now $(\lambda_0+a)-(\lambda_0+b)=(\lambda_1-\lambda_2-\lambda_3+\lambda_4)/2$

from which

$$\boxed{\frac{m_A-m_B}{m_S}=\frac{\lambda_1-\lambda_2-\lambda_3+\lambda_4}{-\lambda_2+\lambda_3+\lambda_4-\lambda_5}} \ . \qquad (1)$$

If one assumes that the five independent readings λ_i have the same variance var(λ), then var($a-b$)=var(λ).

Remark: The hypothesis according to which variations of *balance reading are directly proportional to differences of* mass is, perhaps, not strictly verified. Let us introduce, therefore, a non-linearity in the form of a second-order term. In place of the differences in readings a, b, $b+s$ and, $a+s$, proportional to differences of mass m_A-m_0, m_B-m_0, $m_B+m_S-m_0$, and $m_A+m_S-m_0$, we write $a+ka^2$, $b+kb^2$,

$b+s+k(b+s)^2$, and $a+s+k(a+s)^2$. The five equations are thus written:

$$\lambda_0+a+ka^2-d_1-d_1'=\lambda_1$$

$$\lambda_0+b+kb^2-d_1=\lambda_2$$

$$\lambda_0+b+s+k(b+s)^2=\lambda_3$$

$$\lambda_0+a+s+k(a+s)^2+d_1'=\lambda_4$$

$$\lambda_0+a+ka^2+d_1+d_1'=\lambda_5 \;.$$

Then

$$\lambda_1-\lambda_2-\lambda_3+\lambda_4=2(a-b)[1+k(a+b+s)] \;,$$

$$-\lambda_2+\lambda_3+\lambda_4-\lambda_5=2s[1+k(a+b+s)] \;,$$

and it is still true that

$$\frac{\lambda_1-\lambda_2-\lambda_3+\lambda_4}{-\lambda_2+\lambda_3+\lambda_4-\lambda_5}=\frac{a-b}{s}=\frac{m_A-m_B}{m_S} \;.$$

Equation (1) therefore remains valid for the new hypothesis.

1.2 Linear Drift as a Function of the Sequence Number of the Equilibrium Observation: One thus has $d_2=d_1=d_1'=d_2'$ (Let us call it d) and the system of five equations and four unknowns is written:

$$(\lambda_0+a)-2d=\lambda_1$$

$$(\lambda_0+b)-d=\lambda_2$$

$$(\lambda_0+b)+s=\lambda_3$$

$$(\lambda_0+a)+s+d=\lambda_4$$

$$(\lambda_0+a)+2d=\lambda_5 \;.$$

One can solve this system by the method of least squares. The solution is:

$$(\lambda_0+a)=(3\lambda_1+\lambda_2-\lambda_3+\lambda_4+3\lambda_5)/7$$

$$(\lambda_0+b)=(-3\lambda_1/4+5\lambda_2+2\lambda_3-2\lambda_4+11\lambda_5/4)/7$$

$$s=(-\lambda_1/4-3\lambda_2+3\lambda_3+4\lambda_4-15\lambda_5/4)/7$$

$$d=(-\lambda_1+\lambda_5)/4 \;.$$

One extracts from this

$$a-b=(15\lambda_1-16\lambda_2-12\lambda_3+12\lambda_4+\lambda_5)/28 \;,$$

from which

$$\boxed{\frac{m_A-m_B}{m_S}=\frac{a-b}{s}=\frac{15\lambda_1-16\lambda_2-12\lambda_3+12\lambda_4+\lambda_5}{-\lambda_1-12\lambda_2+12\lambda_3+16\lambda_4-15\lambda_5}} \;.$$

With the same hypothesis as above, var$(a-b)=$ 55/56 var(λ).

1.3 Quadratic Drift as a Function of the Sequence Number of the Equilibrium Observation: It is sufficient to add to the total drift with respect to the third equilibrium position a second-order term: δ for the second and fourth equlibria and 4δ for the first and fifth.

The new system of five equations and five unknowns is:

$$(\lambda_0+a)\qquad -2d+4\delta=\lambda_1$$

$$(\lambda_0+b)\qquad -d+\delta=\lambda_2$$

$$(\lambda_0+b)+s\qquad =\lambda_3$$

$$(\lambda_0+a)\qquad +s+d+\delta=\lambda_4$$

$$(\lambda_0+a)\qquad +2d+4\delta=\lambda_5 \;.$$

The solution is:

$$(\lambda_0+a)=(-\lambda_1+4\lambda_2-4\lambda_3+4\lambda_4-\lambda_5)/2$$

$$(\lambda_0+b)=(-\lambda_1+3\lambda_2-\lambda_3+\lambda_4)/2$$

$$s=(\lambda_1-3\lambda_2+3\lambda_3-\lambda_4)/2$$

$$d=(-\lambda_1+\lambda_5)/4$$

$$\delta=(\lambda_1-2\lambda_2+2\lambda_3-2\lambda_4+\lambda_5)/4 \;,$$

from which one extracts

$$a-b=(\lambda_2-3\lambda_3+3\lambda_4-\lambda_5)/2 \;,$$

so that

$$\boxed{\frac{m_A-m_B}{m_S}=\frac{a-b}{s}=\frac{\lambda_2-3\lambda_3+3\lambda_4-\lambda_5}{\lambda_1-3\lambda_2+3\lambda_3-\lambda_4}} \;.$$

With the same hypothesis as above, var$(a-b)=5$var(λ).

Remark: At present, it is the first hypothesis [viz. I.1] which seems the best verified. In particular, the hypothesis of a quadratic drift is not verified.

2. Weighings

Classical weighing involves four standards, the mass differences of which are determined from pair-wise intercomparisons in all possible combinations. Let A, B, C, and D be standards having mass m_A, m_B, m_C, and m_D, respectively. In the *classic design* one determines successively:

$$m_A - m_B = m_1$$

$$m_A - m_C = m_2$$

$$m_A - m_D = m_3$$

$$m_B - m_C = m_4$$

$$m_B - m_D = m_5$$

$$m_C - m_D = m_6 \ ,$$

where each m_i $(i=1,\ldots,6)$ is the result of a subweighing. One considers m_A as known. Thus the differences $m_B - m_A$, $m_C - m_A$, and $m_D - m_A$, which we will designate m'_B, m'_C, and m'_D, constitute the three actual unknowns. We should also point out that each m_i is subject to random errors associated with the subweighing measurements. The six conditional equations are thus written:

$$-m'_B = m_1$$

$$-m'_C = m_2$$

$$-m'_D = m_3$$

$$m'_B - m'_C = m_4$$

$$m'_B - m'_D = m_5$$

$$m'_C - m'_D = m_6 \ .$$

One can solve this system by the method of least squares. The normal equations are:

$$3m'_B - m'_C - m'_D = -m_1 + m_4 + m_5$$

$$-m'_B + 3m'_C - m'_D = -m_2 - m_4 + m_6$$

$$-m'_B - m'_C + 3m'_D = -m_3 - m_5 - m_6 \ .$$

By addition, one obtains

$$m'_B + m'_C + m'_D = -(m_1 + m_2 + m_3) \qquad (2)$$

from which

$$\boxed{\begin{aligned} m'_B &= (-2m_1 - m_2 - m_3 + m_4 + m_5)/4 \\ m'_C &= (-m_1 - 2m_2 - m_3 - m_4 + m_6)/4 \\ m'_D &= (-m_1 - m_2 - 2m_3 - m_5 - m_6)/4 \end{aligned}}$$

By supposing the determinations of m_1, $m_2, \ldots, m_6$ to be independent and $\mathrm{var}(m_1) = \mathrm{var}(m_2) = \ldots = \mathrm{var}(m_6) = \mathrm{var}(m)$, one has immediately

$$\mathrm{var}\ (m'_B) = \mathrm{var}(m'_C) = \mathrm{var}(m'_D) = (1/2)\ \mathrm{var}(m)$$

$$\begin{aligned} \mathrm{covar}\ (m'_B, m'_C) &= \mathrm{covar}(m'_C, m'_D) \\ &= \mathrm{covar}(m'_D, m'_B) = (1/4)\ \mathrm{var}(m) \ . \end{aligned}$$

The residual deviations are:

$$g_1 = m_1 + m'_B \qquad = (2m_1 - m_2 - m_3 + m_4 + m_5)/4$$

$$g_2 = m_2 + m'_C \qquad = (-m_1 + 2m_2 - m_3 - m_4 + m_6)/4$$

$$g_3 = m_3 + m'_D \qquad = (-m_1 - m_2 + 2m_3 - m_5 - m_6)/4$$

$$g_4 = m_4 - m'_B + m'_C = (m_1 - m_2 + 2m_4 - m_5 + m_6)/4$$

$$g_5 = m_5 - m'_B + m'_D = (m_1 - m_3 - m_4 + 2m_5 - m_6)/4$$

$$g_6 = m_6 - m'_C + m'_D = (m_2 - m_3 + m_4 - m_5 + 2m_6)/4 \ .$$

3. Errors

Following numerous weighings, G. Girard [of BIPM] observed that the mass value found for a standard varies in a rather reproducible fashion with the position taken by this standard in the sequence of the four standards used during the weighing. The reference standard (of supposedly known mass) was, of course, always the same artifact. Another indication of a problem arose in the comparison of values obtained by a global treatment of the subweighings made in the course of the following six weighings involving six standards: (ABCD), (ABCE), (ABCF), (ABDE), (ABDF), and (ABEF) compared with those which one obtains from the following weighings: (ABCD), (BCDE), (CDEF), (DEFA), (EFAB), and (FABC). Each set of six weighings supplies 36 observations. Mass values of B, C, D, E, and F can be obtained for each set by the method of least squares assuming that the mass of A is known. The values obtained for the masses of B, C, D, E, and F using the two different ts of six weighings shown above differ significantly. (For ture reference, we refer to the first set above as Set I and e second set as Set II).

The subweighings are therefore tainted by errors, which re to some extent reproducible. These errors are compensated in the case of Set II (where all the standards play identical roles), but not in the case of Set I, nor in the case of a single weighing.

It was first supposed that all the subweighings were tainted by the same error, which could be caused by a non-linearity of the drift in equilibrium position of the balance at constant load. The global treatment of a number of subweighings has not confirmed this hypothesis, a result consistent with the remark made in 1.3.

One then supposed that the error depends on the rank of the subweighing—that is to say that the result of a subweighing of rank i ($i=1,2,\ldots,6$) gives a result tainted by an error ϵ_i.

Thus each m_i, in addition to random error, is biased by a systematic error ϵ_i. To notate this explicitly, consider the measured mass difference m_1. The quantity m_1, we now must admit, has two components which we separate in the following way:

$$m_1=m'_1+\epsilon_1$$

where m'_1 is an estimate of m_A-m_B subject only to random errors having a variance var(m), and ϵ_1 is the bias in the measurement. (As a first approximation, we suppose that the ϵ_i's do not themselves have a random component.) Thus, it is incorrect to treat m_1 as an unbiased estimate of m_A-m_B. We now assume that m_1 is in reality an unbiased estimate of $m_A-m_B+\epsilon_1$, i.e.

$$m_A-m_B+\epsilon_1=m_1.$$

The correct conditional equations corresponding to the classic design of the weighing (ABCD) are thus

$$m_A-m_B+\epsilon_1=m_1$$

$$m_A-m_C+\epsilon_2=m_2$$

$$m_A-m_D+\epsilon_3=m_3$$

$$m_B-m_C+\epsilon_4=m_4$$

$$m_B-m_D+\epsilon_5=m_5$$

$$m_C-m_D+\epsilon_6=m_6$$

If one now assumes that each ϵ_i keeps a constant value for weighings made on different days, but according to a schedule and a procedure as invariant as possible, one may make a global solution of a number of weighings and avail oneself of sufficient conditional equations to find the six unknown ϵ_i's thus introduced. (Note that the ϵ_i's cannot be uniquely determined from Set I but can be uniquely determined from Set II.)

The first such estimate gave[2]

$$\epsilon_2\approx\epsilon_3\approx\epsilon_4\approx\epsilon_5\approx\epsilon_6\approx 2.8\ \mu\text{g, denoted by } \epsilon;$$

$$\epsilon_1\approx-1.4\ \mu\text{g}\approx-\epsilon/2.$$

Introduction of the ϵ_i's into the treatment of results from Set II has, in addition, led to a notable reduction of the residual deviations. Typical data are presented in tables 1 and 2.

Let us take up again the conditional equations used for the classical weighing design (ABCD) in the form

$$m_A-m_B=m'_1+\epsilon_1$$

$$m_A-m_C=m'_2+\epsilon_2$$

$$m_A-m_D=m'_3+\epsilon_3$$

$$m_B-m_C=m'_4+\epsilon_4$$

$$m_B-m_D=m'_5+\epsilon_5$$

$$m_C-m_D=m'_6+\epsilon_6$$

Looking at the conditional equations in this form, we can estimate how the ϵ_i's would effect results calculated in ignorance of the bias which, in fact, exists.

It is very clear that the results obtained in section 2 are immediately applicable, by replacing m_i by ϵ_i, in order to find both the errors arising from the ϵ_i's as well as the contribution of the ϵ_i's to the residual deviations.

One obtains the following expressions and, taking account of the estimates given above for the ϵ_i's, the following numerical values:

$$\Delta m'_B=(-2\epsilon_1-\epsilon_2-\epsilon_3+\epsilon_4+\epsilon_5)/4\approx+\epsilon/4\approx+0.7\ \mu\text{g}$$

$$\Delta m'_C=(-\epsilon_1-2\epsilon_2-\epsilon_3-\epsilon_4+\epsilon_6)/4\approx-5\epsilon/8\approx-1.8\ \mu\text{g}$$

$$\Delta m'_D=(-\epsilon_1-\epsilon_2-2\epsilon_3-\epsilon_5-\epsilon_6)/4\approx-9\epsilon/8\approx-3.2\ \mu\text{g}$$

where $\Delta m'_B$, for instance, is the error in the calculated mass value of B which is incurred by ignoring the existence of bias.

---

[2] The results for the primary kilogram comparator used at NBS suggest that $\{\epsilon_i\}\approx\epsilon\approx 2.6$ μg.[2]).

Table 1. Results from a modified Set II design (only five weights were used). The column labeled "w" lists deviations to the least squares solution of the 11×30 design matrix shown. The column labeled "w/o" shows deviations to the least squares solution of the 5×30 design matrix in which the ϵ_i's are ignored.

A	B	C	D	E	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5	ϵ_6			
Constraint													
+											*0.1030 mg*		
Observations													
												w	w/o
+	−				+						−0.1792	−1.4 μg	−2.5 μg
+		−				+					−0.0562	−1.5	1.0
+			−				+				−0.2378	0.1	3.3
	+	−						+			0.1226	1.2	3.1
	+		−						+		−0.0589	2.7	5.5
		+	−							+	−0.1790	1.7	4.9
	+	−			+						0.1188	0.6	−0.7
	+		−			+					−0.0612	0.7	3.2
	+			−			+				0.1059	−0.8	2.2
		+	−					+			−0.1822	−0.4	1.7
		+		−					+		−0.0139	−0.9	1.9
			+	−						+	0.1702	−0.9	2.1
		+	−		+						−0.1861	−1.2	−2.2
		+		−		+					−0.0161	−2.8	−0.3
−		+					+				0.0613	1.0	4.1
			+	−				+			0.1703	0.3	2.2
−			+						+		0.2434	−0.4	2.3
−				+						+	0.0755	−0.6	2.5
			+	−	+						0.1698	2.9	1.7
−			+			+					0.2446	1.0	3.5
	−		+				+				0.0683	0.8	3.9
−				+				+			0.0735	−1.6	0.5
	−			+					+		−0.1018	−1.0	1.9
+	−									+	−0.1723	1.3	4.4
−				+	+						0.0711	−0.8	−1.9
	−			+		+					−0.0984	2.6	5.3
		−		+			+				0.0179	−1.0	2.1
+	−							+			−0.1741	0.5	2.6
+		−							+		−0.0548	−0.3	2.4
	+	−								+	0.1210	−1.5	1.5
				Σg_i^2:								60 μg²	260 μg²
				Variance of the Fit:								3 μg²	10 μg²
				Std. Dev. of the Fit:								1.7 μg	3.1 μg
				Degrees of Freedom:								20	26

The following set of Δg_i are the residual deviations to the least squares fit due only to the ϵ_i's.

$$\Delta g_1=(2\epsilon_1-\epsilon_2-\epsilon_3+\epsilon_4+\epsilon_5)/4 \quad \approx -\epsilon/4 \quad \approx -0.7\ \mu\text{g}$$

$$\Delta g_2=(-\epsilon_1+2\epsilon_2-\epsilon_3-\epsilon_4+\epsilon_6)/4 \approx +3\epsilon/8 \approx +1.0\ \mu\text{g}$$

$$\Delta g_3=(-\epsilon_1-\epsilon_2+2\epsilon_3-\epsilon_5-\epsilon_6)/4 \approx -\epsilon/8 \quad \approx -0.4\ \mu\text{g}$$

$$\Delta g_4=(\epsilon_1-\epsilon_2+2\epsilon_4-\epsilon_5+\epsilon_6)/4 \quad \approx +\epsilon/8 \quad \approx +0.4\ \mu\text{g}$$

$$\Delta g_5=(\epsilon_1-\epsilon_3-\epsilon_4+2\epsilon_5-\epsilon_6)/4 \quad \approx -3\epsilon/8 \approx -1.0\ \mu\text{g}$$

$$\Delta g_6=(\epsilon_2-\epsilon_3+\epsilon_4-\epsilon_5+2\epsilon_6)/4 \quad \approx +\epsilon/2 \quad \approx +1.4\ \mu\text{g}$$

$$\Sigma_i(\Delta g_i)^2 \approx 5\epsilon^2/8 \quad \approx 4.9\ \mu\text{g}^2.$$

Note that this result ignores cross terms which would be present if the random error is non-negligible. The cross terms, however, may be either positive or negative so that considerable cancellation occurs in their summation. Thus semiquantitative conclusions may still be drawn even though the cross terms are ignored.

Now, $\Sigma_i g_i^2$ is typically about 12 μg². For the classical weighing design with 6 observations and 3 independent unknowns the variance of an observation is $s^2=(1/3)\,\Sigma_i g_i^2=4$ μg², and the variance of a single result of the weighing is $s^2/2=2$ μg².

One sees that the ϵ_i's contribute in a modest (and, what is more, not directly detectable) way to the residual deviations, while they impose significant errors on the results of the weighing.

Let us pursue this analysis by now considering groups of weighings.

Table 2. The variance/convariance matrix for the 11×30 design shown in table 1 as well as least squares solutions to A, B, C, D, E, and the ϵ_i's. The mass values calculated for A-E would have been the same even if the ϵ_i's had not been included in the model. Note that, assuming cancellation of cross terms, we would estimate $\Sigma g_i^2(\mathrm{w/o})=\Sigma g_i^2(\mathrm{w})+\Sigma(\Delta g_i)^2$, where $\Sigma(\Delta g_i)^2=5(\Sigma\epsilon_i^2)=195$ μg². The data of table 1 verify this assumption.

Variance/Covariance Matrix

A	B	C	D	E	ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5	ϵ_6
0	0	0	0	0	0	0	0	0	0	0
0	2	1	1	1	0	0	0	0	0	0
0	1	2	1	1	0	0	0	0	0	0
0	1	1	2	1	0	0	0	0	0	0
0	1	1	1	2	0	0	0	0	0	0
0	0	0	0	0	3	0	0	0	0	0
0	0	0	0	0	0	3	0	0	0	0
0	0	0	0	0	0	0	3	0	0	0
0	0	0	0	0	0	0	0	3	0	0
0	0	0	0	0	0	0	0	0	3	0
0	0	0	0	0	0	0	0	0	0	3

Divisor=15

	Least Squares Results	**Std. Dev.**
A	0.1030 mg	0.0 μg
B	0.2797	0.6
C	0.1602	0.6
D	0.3441	0.6
E	0.1760	0.6
ϵ_1	−1.1 μg	0.7 μg
ϵ_2	2.6	0.7
ϵ_3	3.1	0.7
ϵ_4	2.0	0.7
ϵ_5	2.8	0.7
ϵ_6	3.1	0.7

For the Set II weighings described above (36 observations, 5 unknown weights), an examination of the least squares solutions with and without the ϵ_i's treated explicitly shows: the ϵ_i's do not impose any error on the results, which would be the same either for a treatment derived from starting equations such as $m_A-m_B=m_1$ or from equations such as $m_A-m_B+\epsilon_1=m_1$. The failure to consider the ϵ_i's explicitly, however, adds a residual deviation of ϵ_i to each subweighing of rank i. The sum of the squares of these contributions is

$$6\Sigma_i\epsilon_i^2\approx6(\epsilon^2/4+5\epsilon^2)=63\epsilon^2/2\ ,$$

a value which contributes to the variance of an observation $(63\epsilon^2/2)/(36-5)=1.02\epsilon^2\approx8$ μg². Once again, this is only an estimate because cross terms involving random error components have been neglected. Similar conclusions can be drawn from the Set II weighings shown in tables 1 and 2 (30 observations, 4 unknown weights).

Introduction of the six supplementary unknowns ϵ_i into the treatment appreciably reduces the variance of an observation. For the example shown in table 1, which is typical, the variance went from 9.6 μg² to 2.9 μg².

For the Set I weighings described above, numerical calculation gives

$$\Delta m_B'\approx+0.25\epsilon\approx+0.7\ \mu\mathrm{g}$$

$$\Delta m_C'\approx-0.58\epsilon\approx-1.6\ \mu\mathrm{g}$$

$$\Delta m_D'\approx-0.78\epsilon\approx-2.2\ \mu\mathrm{g}$$

$$\Delta m_E'\approx-0.98\epsilon\approx-2.7\ \mu\mathrm{g}$$

$$\Delta m_F'\approx-1.18\epsilon\approx-3.3\ \mu\mathrm{g}$$

and, for the variance of an observation, $0.15\epsilon^2\approx1.2$ μg². Once again, these are approximate errors which one would suffer through ignorance of the existence of a bias.

One thus sees how imprudent it would have been to choose the design which gives the smaller variance since, for this design, the ϵ_i's impose significant errors on the result of the weighings while, for the design which gives the greater variance, they impose no errors.

4. Remedies

It would not be judicious to retain the classic design for weighings, choosing a set of such designs which ensures that the contribution of the ϵ_i's to the final error of the results is zero. For such a case, the contribution of the ϵ_i's to the variance of an observation would lead to a serious over-estimation of the variance. One could, of course, include the ϵ_i's explicitly in the analysis. However, this approach would greatly increase the number of necessary observations over what had been previously required, and it would be based on the assumption that the ϵ_i's are constant throughout the many days required for a set of measurements.

It goes without saying that the solution lies in discovering the physical cause for the existence of the ϵ_i's and in eliminating it. But, in the meanwhile, one must carry out weighings and, since it is impossible to eliminate the cause, it is necessary to eliminate the effect; that is to say to ensure that the ϵ_i's compensate themselves as exactly as possible, or, put another way, to find an *unbiased* observation which will estimate m_A-m_B, for example.

The reader may have been struck by the fact that we have adopted for all the $\epsilon_i\,(i=2,3\ldots,6)$'s the same value ϵ, while $\epsilon_1=-\epsilon/2$. Without doubt, he or she has good reason to think that it would have been simple to obtain $\epsilon_1=\epsilon$ operationally; for example, by the addition of a preliminary subweighing.

The expressions derived above would be modified in this case to be:

$$\Delta m_B' = -\epsilon/2 = -1.4\ \mu g$$

$$\Delta m_C' = -\epsilon = -2.8\ \mu g$$

$$\Delta m_D' = -3\epsilon/2 = -4.2\ \mu g$$

$$\Delta g_1 = +\epsilon/2 = +1.4\ \mu g$$

$$\Delta g_2 = 0$$

$$\Delta g_3 = -\epsilon/2 = -1.4\ \mu g$$

$$\Delta g_4 = +\epsilon/2 = +1.4\ \mu g$$

$$\Delta g_5 = 0$$

$$\Delta g_6 = +\epsilon/2 = +1.4\ \mu g$$

$$\Sigma_i(\Delta g_i)^2 = \epsilon^2 = 7.8\ \mu g^2\ .$$

One would thus obtain both a worsening of the errors attached to the results and an increase in the residual deviations.

For Set II weighings the error attached to the results would, of course, remain zero but the contribution of the ϵ_i's to the variance would be of order $6\times6\epsilon^2/31 = 1.16\epsilon^2 \approx 9\ \mu g^2$.

For Set I, a numerical calculation indicates that one would then have

$$\Delta m_B' \approx -0.5\epsilon \approx -1.4\ \mu g$$

$$\Delta m_C' \approx -0.9\epsilon \approx -2.7\ \mu g$$

$$\Delta m_D' \approx -1.15\epsilon \approx -3.2\ \mu g$$

$$\Delta m_E' \approx -1.35\epsilon \approx -3.8\ \mu g$$

$$\Delta m_F' \approx -1.55\epsilon \approx -4.3\ \mu g$$

and, for the variance of an observation, $0.23\epsilon^2 \approx 1.8\ \mu g^2$.

Contrary to what one might intuitively think, the act of making the ϵ_i's equal would not automatically eliminate errors in the results.

At this juncture, a remark must be made. In the course of the subweighing of rank *i*, if one intercompares standards Y and X instead of X and Y, one replaces the equation

$$m_X - m_Y + \epsilon_i = m_{i,1}$$

with

$$m_Y - m_X + \epsilon_i = m_{i,2}$$

which is equivalent to $m_X - m_Y - \epsilon_i = -m_{i,2}$. The sign of ϵ_i is thus reversed.

One could hope that by changing the sign of certain, judiciously chosen, ϵ_i's one might eliminate, for each weighing, errors in $\Delta m_B'$, $\Delta m_C'$, $\Delta m_D'$. Unfortunately, there is nothing to be gained from this approach; in essence, one would then have (see eq (2)) $\Delta m_B' + \Delta m_C' + \Delta m_D' = \pm\epsilon_1 \pm \epsilon_2 \pm \epsilon_3$ an expression which cannot vanish as long as $\epsilon_1 = \epsilon_2 = \epsilon_3 \neq 0$.

But rather than replacing the intercomparison of X and Y by that of Y and X, it is quite clear that one must make both comparisons, which we will call "opposed comparisons," under as identical conditions as possible and, in particular, in subweighings of the same rank. The two equations written above give, assuming that ϵ_i keeps the same value during the two subweighings,

$$\epsilon_i = (m_{i,1} + m_{i,2})/2$$

and

$$m_X - m_Y = (m_{i,1} - m_{i,2})/2.$$

It is this last value which ought to be introduced into the conditional equations. This approach requires that the ϵ_i's remain essentially constant for only two successive days.

The problem is, therefore, to define a weighing design to which one may make a corresponding "opposed design." *A priori*, if one sticks to the principle of the classic weighing in which four standards are involved for which one determines six paired differences, one can devise $6!2^6 = 46080$ different designs (since one can imagine all permutations of the 6 difference determinations and each difference determination can be realized by one or the other of two opposed comparisons). The ensemble of the designs is composed of 23040 pairs such that, in each of them, subweighings of the same rank correspond to opposed comparisons.

We now note that, in the course of the classic subweighing involving the intercomparison of X and Y (which we will write as (X,Y)), the first and fifth operations involve X. If the following comparison is (X,Z), with of course Z≠Y, the first operation again involves X and there is no rotation of the weight table between these two comparisons. On the other hand, the two opposed comparisons (Y,X) and (Z,X) are separated by a rotation of the weight table. One might worry that the value of ϵ_i is a function of whether the subweighing of rank i is or is not preceded by a rotation of the weight table. It is therefore preferable that absences and presences of these rotations correspond in the opposed design, but we have just seen that an absence of rotation always corresponds to a rotation in the opposed designs. We must therefore exclude all designs for which a subweighing is not preceded by a rotation—either in the first design or in the opposed design. This can be simply expressed through the following conditions: for the two consecutive comparisons (X,Y) and (Z,T) it is necessary that Z≠X and T≠Y. With these restrictions, there still remain 1776 pairs of possible designs.[3]

One of the paired designs is, for example,

(A,B)	(B,A)
(B,C)	(C,B)
(C,D)	(D,C)
(D,A)	(A,D)
(A,C)	(C,A)
(B,D)	(D,B) .

The first column (i.e., the direct weighing) corresponds to the following conditional equations

$$m_A - m_B + \epsilon_1 = m_1$$

$$m_B - m_C + \epsilon_2 = m_2$$

$$m_C - m_D + \epsilon_3 = m_3$$

$$m_D - m_A + \epsilon_4 = m_4$$

$$m_A - m_C + \epsilon_5 = m_5$$

$$m_B - m_D + \epsilon_6 = m_6 \ .$$

This design, used by itself, would not have a marked advantage over the classic design. It can be shown that it would lead, nevertheless, to smaller errors but also to a significant overestimation of the variance of an observation. It is understood that if one chooses it from among designs which satisfy the conditions we have imposed (and which seem equivalent), one must also use the opposed design and take as conditional equations

$$m_A - m_B = (m_{1,1} - m_{1,2})/2$$

$$m_B - m_C = (m_{2,1} - m_{2,2})/2$$

$$m_C - m_D = (m_{3,1} - m_{3,2})/2$$

$$m_D - m_A = (m_{4,1} - m_{4,2})/2$$

$$m_A - m_C = (m_{5,1} - m_{5,2})/2$$

$$m_B - m_D = (m_{6,1} - m_{6,2})/2$$

in which the differences derived for a subweighing of rank i ($i = 1,2,\ldots 6$) are denoted by $m_{i,1}$ for the direct weighing and $m_{i,2}$ for the opposed weighing.

References

[1] Almer, H.E., J. Res. Nat. Bur. Stand. (U.S.) **76C** (1 and 2): 1-10 (January-June, 1972).

[2] Davis, R.S., J. Res. Nat. Bur. Stand. (U.S.) **90-4** 263-283 (July-August, 1985).

---

[3] P. Carré has rigorously derived this number. We have chosen, for the sake of brevity, to omit the derivation here.

Journal of Research of the National Bureau of Standards
Volume 90, Number 5, September–October 1985

Thermodynamics of Solution of $SO_2(g)$ in Water and of Aqueous Sulfur Dioxide Solutions

R. N. Goldberg and V. B. Parker
National Bureau of Standards, Gaithersburg, MD 20899

Accepted: June 19, 1985

A consistent set of thermochemical property values, $\Delta_f H^\circ$, $\Delta_f G^\circ$, S°, and C_p°, at 298.15 K is given for the known constituents of aqueous sulfur dioxide ($SO_2^0(aq)$, $HSO_3^-(aq)$, $SO_3^{2-}(aq)$, $H^+(aq)$, and $S_2O_5^{2-}(aq)$). Also tabulated are values of the mean ionic activity coefficients, osmotic coefficients, partial pressure of $SO_2(g)$, and the relative apparent molar enthalpy as a function of concentration of $SO_2(aq)$ at 298.15 K. The data analysis considered a wide variety of measurement techniques: calorimetric enthalpies of solution and reaction, heat capacities, equilibrium constants, solubilities, and vapor pressure measurements, both partial and total, over aqueous solutions of SO_2 for the temperature range 278 to 393 K. All auxiliary data have been taken from the most recent set of CODATA values which were converted to a standard state pressure of one bar (0.1 MPa). For the process $SO_2(g)=SO_2^0(aq)$, the selected "best" values are: $K=1.23\pm0.05$ mol kg^{-1} bar^{-1}, $\Delta G^\circ=-0.51\pm0.10$ kJ mol^{-1}, $\Delta H^\circ=-26.97\pm0.30$ kJ mol^{-1}, and $\Delta C_p^\circ=155\pm10$ J mol^{-1} K^{-1}. The standard state partial molar entropy of $SO_3^{2-}(aq)$, obtained by the analysis of data via two independent thermodynamic pathways is -15.40 ± 0.80 J mol^{-1} K^{-1} at 298.15 K. Parameters are given which extend the predictions to temperatures up to 373 K.

Key words: bisulfite; calorimetry; enthalpy; entropy; Gibbs energy; heat capacity; pyrosulfite; solubility; sulfite; sulfur dioxide; thermodynamic data; vapor pressure; water.

1. Introduction

The purpose of this paper is to present the results of an analysis of the thermodynamic properties of the SO_2+H_2O system. This study was performed as part of the analysis of the thermodynamics of sulfur and its compounds. The aqueous species of relevance to this study are neutral aqueous sulfur dioxide $SO_2^0(aq)$, bisulfite ion $HSO_3^-(aq)$, sulfite ion $SO_3^{2-}(aq)$, pyrosulfite ion $S_2O_5^{2-}(aq)$, and $H^+(aq)$. A consistent set of property values[1], $\Delta_f H^\circ$, $\Delta_f G^\circ$, S°, and C_p° at 298.15, is given for these species which results from the evaluation of the various processes involving them. Also included are values of the activity and osmotic coefficients, the relative apparent molar enthalpy (L_ϕ) as a function of concentration, and the partial pressure of $SO_2(g)$ over its aqueous solution.

About the Authors, Paper: R. N. Goldberg and V. B. Parker are chemists in the Chemical Thermodynamics Division of NBS' Center for Chemical Physics. The work they report on was sponsored by the U.S. Department of Energy, the NBS Office of Standard Reference Data, and the Design Institute for Physical Property Data of the American Institute of Chemical Engineers.

Throughout this paper we are adhering to the standard states and units used in the NBS Tables of Chemical Thermodynamic Properties [1][2]. All auxiliary thermochemical data have been taken from the most recent set of CODATA key values [2] which were converted,

---

[1] See Glossary of symbols at end of paper.

[2] Figures in brackets indicate literature references.

where appropriate, to a standard state pressure of one bar (0.1 MPa).

Thermodynamic data from a wide variety of measurement techniques are considered. They include calorimetric measurements of heat capacities and enthalpies of solution and reaction, measured equilibrium constants, entropies, solubilities, and vapor pressure measurements (both partial and total) over aqueous solutions. The approach adopted in this paper is to select a tentative set of values and accompanying uncertainties for the various processes involving sulfur (IV)-oxygen species and then to examine the various thermodynamic pathways for consistency. A final set of values for both the processes and thermodynamic properties is then selected which provides a "best" fit to all of the available data. The final uncertainties are adjusted to reflect the agreement or lack of it between the various data sets.

2. Key Processes Involved in the Evaluation

The species considered to be present in an aqueous solution of SO_2 (stoichiometric "$SO_2(aq)$") are neutral, unionized $SO_2^0(aq)$, $HSO_3^-(aq)$, $SO_3^{2-}(aq)$, $S_2O_5^{2-}(aq)$, and $H^+(aq)$. They are involved in the following equilibria:

$$SO_2^0(aq) + H_2O(l) = H^+(aq) + HSO_3^-(aq) \qquad (A)$$

$$HSO_3^-(aq) = H^+(aq) + SO_3^{2-}(aq) \qquad (B)$$

$$2\ HSO_3^-(aq) = S_2O_5^{2-}(aq) + H_2O(l) \qquad (C)$$

Selections are made for the various parameters that enter into the solution process of $SO_2(g)$ into water so that a set of "best" values of $\Delta G°$ and $\Delta H°$ can be obtained for the processes:

$$SO_2(g) = SO_2^0(aq) \qquad (D)$$

$$SO_2(g) + H_2O(l) = H^+(aq) + HSO_3^-(aq) \qquad (E)$$

The standard state partial molar entropy of $SO_3^{2-}(aq)$ can be obtained from the standard state Gibbs energy and enthalpy changes for the process:

$$SO_2(g) + H_2O(l) = 2\ H^+(aq) + SO_3^{2-}(aq) \qquad (F)$$

which is the summation of processes (A), (B), and (D). This partial molar entropy can also be obtained independently using the data for the process:

$$Na_2SO_3(cr) = 2\ Na^+(aq) + SO_3^{2-}(aq) \qquad (G)$$

together with the Third Law entropy for $Na_2SO_3(cr)$, the enthalpy of solution, and the value of $\Delta G°$ for the solution process (G).

3. Equilibrium Calculations on Aqueous Solutions of Sulfur Dioxide

The equilibrium constants for the equilibria in aqueous solutions of SO_2 as described in processes (A), (B), and (C) are:

$$K_A = [\hat{a}\{H^+(aq)\}\ \hat{a}\{HSO_3^-(aq)\}] \div [\hat{a}\{SO_2^0(aq)\}\ \hat{a}\{H_2O(l)\}] \qquad (1)$$

$$K_B = \hat{a}\{H^+(aq)\}\ \hat{a}\{SO_3^{2-}(aq)\}/\hat{a}\{HSO_3^-(aq)\} \qquad (2)$$

and

$$K_C = \hat{a}\{S_2O_5^{2-}(aq)\}\ \hat{a}\{H_2O(l)\} \div \hat{a}^2\{HSO_3^-(aq)\}. \qquad (3)$$

In the above equations, $\hat{a}_i$ is the activity of the ith species which is equal to the quantity $\hat{m}_i\hat{\gamma}_i$ where $\hat{m}_i$ and $\hat{\gamma}_i$ are, respectively, the molality and activity coefficient of that species. The symbol " ˆ " denotes a quantity which pertains to a species as distinct from a stoichiometric quantity [3]. The absence of the symbol " ˆ " over a quantity implies that the quantity is stoichiometric. The treatment of the experimental data often requires some assumptions about the activity coefficients of the species in solution and the mathematical solution of these three simultaneous, nonlinear equations for the molalities of all of the species considered to be present in solution. The first requirement will now be addressed.

Since we have no direct knowledge of the activity coefficients of these or any other individual aqueous charged species, it is necessary to make assumptions to obtain values for these activity coefficients. It is assumed that the activity coefficients can be calculated using the expression:

$$\ln\ \hat{\gamma}_i = -A_m \hat{I}^{1/2}\hat{z}_i^2/(1 + B\,\hat{I}^{1/2}) \qquad (4)$$

where A_m is the Debye-Hückel constant, B is an "ion-size" parameter, and $\hat{I}$ is the ionic strength which is calculated as:

$$\hat{I} = (1/2)\Sigma \hat{m}_i \hat{z}_i^2. \qquad (5)$$

The above summation extends over all of the species in the solution. The Debye-Hückel constants tabulated by Clarke and Glew [4] were used in our calculations. Staples and Beyer [5] have recently calculated values of the Debye-Hückel limiting slopes using the IAPS recommended values for the dielectric constants and density of water as a function of temperature [6,7]. These new limiting slopes differ by 0.24 and 0.61%, respectively, from the Gibbs energy and enthalpy limiting slopes given by Clarke and Glew [4]. The effect on our final recommended values due to uncertainties in the Debye-Hückel slopes is negligible in comparison with the experimental errors in measured properties which were extrapolated to zero ionic strength whenever possible.

The B parameter in eq [4] is a quantity which serves to account for the repulsive forces between the particles. If sufficiently accurate activity or osmotic coefficients were available for aqueous solutions of SO_2, it would be possible to regress a value for this parameter using a chemical equilibrium model [3]. Since this information is not available, it is necessary to work with estimated or inferred values of B. Evidence for a value of B equal to 1.5 was obtained when extrapolating measured values of K_A (sec. 4.1) to zero ionic strength; specifically, the extrapolation could frequently be made with a line of lesser slope when a value of 1.5 for B was used to calculate values of $\hat{\gamma}_i$. Also, a B value of 1.5 is typical of the values obtained by Hamer and Wu [8] from their correlation of the activity and osmotic coefficient data of aqueous uni-univalent electrolytes. Evidence for a value of B equal to 2.5 was found in the treatment of the heat of solution and heat of dilution data (sec. 4.2) to obtain values of ΔH_A°. There the best fit of the experimental data was frequently obtained using a value of B equal to 2.5. Less direct evidence for a value of 2.5 comes from a treatment of the osmotic coefficients of aqueous sulfuric acid using an equilibrium model [3] in which it was found that a value of B equal to 2.5 could represent the osmotic coefficients of sulfuric acid to within 1.2% up to a stoichiometric molality of 0.20 mol kg^{-1}. This stoichiometric molality corresponds to a calculated (see eq (5)) (species) ionic strength of 0.25 mol kg^{-1}. Since an aqueous solution of SO_2 with a stoichiometric molality of 2.0 mol kg^{-1} has a calculated ionic strength of only 0.21 mol kg^{-1}, the use of assumed values of B ranging from 1.5 to 2.5 will allow us to proceed with the necessary equilibrium calculations without serious error.

The solution of the simultaneous, nonlinear equations (1), (2), and (3) was accomplished using an iterative numerical procedure (CO5NCF) which is a part of a library of numerical analysis programs [9]. The calculation was made self-consistent both in regards to the activity of the water which is a participant in the equilibria (see reference [3]) and with eqs (4,5). This model allows us to calculate the amounts of each species in the solutions and values of the fractions (α_i) of the species, i.e., $\alpha_i = \hat{n}_i / m^{st}$ where m^{st} is the stoichiometric amount of SO_2 in solution calculated as $m^{st}(H_2SO_3)$ for convenience. It does not imply the physical existence of H_2SO_3. Results of these calculations at 298.15 K are given in table 1 using our final selected values of K_A, K_B, and K_C and a value of B equal to 2.0.

Throughout, where necessary and possible, reported equilibrium constants have been corrected to 298.15 K, for the activity, and to zero ionic strength.

4. The Thermodynamic Parameters for the Description of the Equilibria in Solution at 298.15 K and as a Function of Temperature

4.1 Process (A)

The reported equilibrium constants at 298.15 K for process (A), the first ionization of SO_2^0(aq), are summarized in table 2a. The most detailed investigations cited are those of Tartar and Garretson [15] and Huss and Eckert [24]. While previous critical evaluations [1,32] have relied upon the results of Tartar and Garretson [15], we have selected the data of Huss and Eckert [24], $K_A = 0.0139 \pm 0.004^a$ mol kg^{-1}, as being preferable since it is based upon two different measurement techniques (conductivity and ultraviolet spectrophotometric measurements) which avoid the possible systematic errors inherent in the use of electrochemical cells which were used by Tartar and Garretson [15].

Measurements of K_A as a function of temperature have been used to obtain values of ΔH_A° which are corrected to 298.15 K (see table 1) using the model of Clarke and Glew [33]. A value of $\Delta C_p^\circ = -272 \pm 10$ J mol^{-1} K^{-1} for process (A) was used in performing these calculations. This heat capacity change is based upon the calorimetrically determined standard state heat capacities of SO_2^0(aq) and HSO_3^-(aq) of, respectively, 195 ± 10 J mol^{-1} K^{-1} (Barbero et al. [34]) and -2 ± 10 J mol^{-1} K^{-1} (Allred et al. [35]). The uncertainties assigned to the calculated enthalpies in table 1b are purely statistical and refer to 95% confidence limits.

The model of Clarke and Glew [33] is based upon a Taylor series expansion of the heat capacity at a reference temperature (298.15 K). When $d\Delta C_p^\circ / dT$ is constant over the temperature range of interest, the

---

[a] The thermodynamic equilibrium constant is dimensionless. The units expression identifies the composition units used for components in K.

Table 1. Fractions (α) of species and the species ionic strength ($\hat{I}$) at 298.15 K in aqueous sulfur dioxide solutions as a function of the stoichiometric molality of SO_2 ($m^{st}(SO_2)$) and of (hypothetical) H_2SO_3($m^{st}(HSO_3)$). These values were calculated using the final selected values of K_A, K_B, and K_C and a value of B equal to 2.0.

$m^{st}(SO_2)$ mol kg^{-1}	$m^{st}(H_2SO_3)$ mol kg^{-1}	$\alpha(H^+)$	$\alpha(HSO_3^-)$	$\alpha(SO_3^{2-})$	$\alpha(SO_2^0)$	$2\alpha(S_2O_5^{2-})$	$\hat{I}$ mol kg^{-1}
0.001000	0.001000	0.9406	0.9404	0.7697×10^{-4}	0.0595	0.6058×10^{-4}	0.9407×10^{-3}
0.002000	0.002000	0.8949	0.8947	0.4036×10^{-4}	0.1051	0.1123×10^{-3}	0.1790×10^{-2}
0.003000	0.003000	0.8576	0.8574	0.2781×10^{-4}	0.1424	0.1573×10^{-3}	0.2573×10^{-2}
0.004000	0.004000	0.8261	0.8259	0.2141×10^{-4}	0.1739	0.1971×10^{-3}	0.3305×10^{-2}
0.005000	0.005000	0.7989	0.7986	0.1749×10^{-4}	0.2011	0.2329×10^{-3}	0.3995×10^{-2}
0.006000	0.006001	0.7750	0.7747	0.1485×10^{-4}	0.2250	0.2654×10^{-3}	0.4651×10^{-2}
0.007000	0.007001	0.7537	0.7534	0.1293×10^{-4}	0.2463	0.2952×10^{-3}	0.5277×10^{-2}
0.008000	0.008001	0.7345	0.7342	0.1148×10^{-4}	0.2655	0.3227×10^{-3}	0.5878×10^{-2}
0.009000	0.009001	0.7172	0.7168	0.1034×10^{-4}	0.2828	0.3483×10^{-3}	0.6457×10^{-2}
0.01000	0.01000	0.7013	0.7009	0.9412×10^{-5}	0.2987	0.3723×10^{-3}	0.7016×10^{-2}
0.02000	0.02000	0.5919	0.5913	0.5111×10^{-5}	0.4081	0.5524×10^{-3}	0.1185×10^{-1}
0.03000	0.03001	0.5271	0.5264	0.3589×10^{-5}	0.4729	0.6742×10^{-3}	0.1583×10^{-1}
0.04000	0.04002	0.4823	0.4816	0.2797×10^{-5}	0.5177	0.7670×10^{-3}	0.1932×10^{-1}
0.05000	0.05004	0.4487	0.4479	0.2307×10^{-5}	0.5513	0.8422×10^{-3}	0.2247×10^{-1}
0.06000	0.06006	0.4221	0.4212	0.1972×10^{-5}	0.5779	0.9056×10^{-3}	0.2537×10^{-1}
0.07000	0.07008	0.4003	0.3994	0.1728×10^{-5}	0.5997	0.9605×10^{-3}	0.2807×10^{-1}
0.08000	0.08011	0.3821	0.3810	0.1541×10^{-5}	0.6179	0.1009×10^{-2}	0.3062×10^{-1}
0.09000	0.09014	0.3664	0.3653	0.1393×10^{-5}	0.6336	0.1052×10^{-2}	0.3304×10^{-1}
0.1000	0.1001	0.3527	0.3516	0.1273×10^{-5}	0.6473	0.1092×10^{-2}	0.3535×10^{-1}
0.1500	0.1504	0.3035	0.3022	0.9015×10^{-6}	0.6965	0.1248×10^{-2}	0.4565×10^{-1}
0.2000	0.2007	0.2718	0.2705	0.7065×10^{-6}	0.7282	0.1363×10^{-2}	0.5456×10^{-1}
0.3000	0.3016	0.2318	0.2303	0.5015×10^{-6}	0.7682	0.1532×10^{-2}	0.6987×10^{-1}
0.4000	0.4029	0.2066	0.2049	0.3934×10^{-6}	0.7934	0.1655×10^{-2}	0.8308×10^{-1}
0.5000	0.5045	0.1887	0.1869	0.3260×10^{-6}	0.8113	0.1754×10^{-2}	0.9493×10^{-1}
0.6000	0.6065	0.1751	0.1732	0.2796×10^{-6}	0.8249	0.1836×10^{-2}	0.1058
0.7000	0.7089	0.1643	0.1624	0.2457×10^{-6}	0.8357	0.1906×10^{-2}	0.1159
0.8000	0.8116	0.1554	0.1534	0.2195×10^{-6}	0.8446	0.1968×10^{-2}	0.1254
0.9000	0.9148	0.1480	0.1459	0.1988×10^{-6}	0.8520	0.2023×10^{-2}	0.1344
1.0000	1.0183	0.1416	0.1395	0.1819×10^{-6}	0.8584	0.2072×10^{-2}	0.1430

temperature dependency of the equilibrium constant is given by:

$$
\begin{aligned}
R\ln K = &-\Delta G^{\circ}_{298.15}/298.15+\Delta H^{\circ}_{298.15}(1/298.15-1/T)\\
&+\Delta C^{\circ}_{p}[(298.15/T)-1+\ln(T/298.15)]\\
&+(298.15/2)\,(\mathrm{d}\Delta C^{\circ}_{p}/\mathrm{d}T)[(T/298.15)-\\
&(298.15/T)-2\,\ln(T/298.15)] \qquad (6)
\end{aligned}
$$

A convenient aspect of the above equation is that known values of ΔG°, ΔH°, ΔC°_{p}, and $(\mathrm{d}\Delta C^{\circ}_{p}/\mathrm{d}T)$ at a reference temperature provide all of the information needed to calculate the variation of the equilibrium constant with temperature.

Given in table 2b are enthalpy values obtained from the direct calorimetric measurements of Vanderzee [41], Dobrogowska and Hepler [39], and Zambonin and Jordan [38] and from the enthalpies of solution of $SO_2(g)$ as a function of concentration from the measurements of Johnson and Sunner [37] and Stiles and Felsing [36]. Values of ΔH°_{A} were also obtained from the heat of dilution measurements of Dobrogowska and Hepler [40] using a regression calculation in which ΔH°_{A} was varied in the equilibrium modeling calculation of L_{ϕ} (sec. 6) until agreement was obtained between measured and calculated values of the heat of dilution. A similar procedure was used for treating the aforementioned enthalpy of solution measurements of $SO_2(aq)$ where values of L_{ϕ} could be obtained from the concentration dependence of the heat of solution measurements. We adopt a value of $\Delta H^{\circ}_{A}=-17.80\pm0.40$ kJ mol^{-1} based largely on the direct calorimetric measurements of Dobrogowska and Hepler [39] and of Vanderzee [41] and the enthalpy of solution measurements of Johnson and Sunner [37]. Note that in table 1b the direct calorimetric measurements of Dobrogowska and Hepler [39] and of Vanderzee [41] are sensitive to the B parameter. This causes an uncertainty in the final value of ΔH°_{A} comparable to the errors in the measurements themselves.

4.2 Process (B)

The thermodynamic data for process (B), the second ionization for $SO_2^0(aq)$, are summarized in table 3.

Table 2. Thermodynamic parameters for process (A): $SO_2^0(aq)+H_2O(l)=H^+(aq)+HSO_3^-(aq)$.

a. Equilibrium constant at 298.15		b. Enthalpy change at 298.15 K		
Investigator(s)	K/mol kg^{-1}	Investigator(s)	ΔH°/kJ mol^{-1}	Method
Drucker [10]	$\sim 0.009^{a,b}$	Stiles and Felsing [36]	-17.2 ± 0.7^a	calorimetric-heats of solution
Sherill and Noyes [11]	0.012^c	Johnstone and Leppla [13]	-16.4 ± 0.9	K vs T
Britton and Robinson [12]	$\sim 0.008^a$	Johnson and Sunner [37]	-18.0 ± 0.6^a	calorimetric-heats of solution
Johnstone and Leppla [13]	0.0130^d	Deveze and Rumpf [19]	-16.7 ± 1.0	K vs T
Yui [14]	0.0127	Flis et al. [20]	$-16.7\pm10.$	K vs T
Tartar and Garretson [15]	0.0172	Zambonin and Jordan [38]	-14.3 ± 1.5	calorimetric-heats of solution
Frydman et al. [16]	0.02^a	Dobrogowska and Hepler [39]	-17.4 ± 0.3^b	calorimetric-neutralization
Ellis and Anderson [17]	0.014	Dobrogowska and Hepler [40]	-18.2 ± 1.0^a	calorimetric-heats of dilution
Ludemann and Franck [18]	0.0103^e	Vanderzee [41]	$-17.8_2\pm0.3^c$	calorimetric-neutralization
Deveze and Rumpf [19]	0.0145			
Flis et al. [20]	0.010			
Deveze [21]	0.013			
Sekine et al. [22]	0.014^a			
Beilke and Lamb [23]	0.0165^a			
Huss and Eckert [24]	0.0139			

[a] Corrections were applied.

[b] Drucker's [10] results are based on the measurements of McRae and Wilson [25] and of Walden and Centnerzwer [26].

[c] Based on measurements of Kerp and Bauer [27] and of Lindner [28].

[d] Based on measurements of Campbell and Maass [29] which are also reported in Beazley et al. [30] and Morgan and Maass [31].

[e] Pressure is 0.2 kbar (20 MPa).

[a] These values were obtained by regression calculations. Additional information obtained from these regressions are as follows: Stiles and Felsing [36], (21 points) B=2.5, $\Delta H_D^o=-27.14\pm0.10$ kJ mol^{-1}. Johnson and Sunner [37], (7 points) B=2.5, $\Delta H_D^o=-26.90\pm0.04$ kJ mol^{-1}. Dobrogowska and Hepler [40], (21 points) B=2.0 to 2.5. The uncertainties given here refer to two standard deviations as distinct from the overall assigned uncertainties given above.

[b] This value was obtained using the Guggenheim equation with $\beta=-0.2$. If an extended Debye-Hückel equation is used with B=2.5, a value of -17.17 kJ mol^{-1} is obtained; if B is set equal to 1.5 then $\Delta H^o=-17.61$ kJ mol^{-1}.

[c] This value was obtained using an extended Debye-Hückel equation with B=1.6. If a value of B=3.0 is used, a value of ΔH_A^o equal to -17.43 kJ mol^{-1} is obtained.

Table 3. Thermodynamic parameters for process (B): $HSO_3^-(aq)=H^+(aq)+SO_3^{2-}(aq)$.

a. Equilibrium constant at 298.15 K		b. Enthalpy change at 298.15 K		
Investigator(s)	$K\times10^8$/mol kg^{-1}	Investigator(s)	ΔH°/kJ mol^{-1}	Method
Jellinek [42]	500^a	Arkhipova et al. [45]	-12.1 ± 4.0	K vs T
Kolthoff [43]	$\sim 2^b$	Zambonin and Jordan [38]	-5.2 ± 0.8	calorimetric
Britton and Robinson [12]	$\sim 4^b$	Hayon et al. [46]	-11.6 ± 3.3	K vs T
Yui [14]	6.3	Teder [47]	-8.0 ± 5.0	K vs T
Tartar and Garretson [15]	6.24	Krunchak et al. [48]	-16.7 ± 6.7	K vs T
Frydman et al. [16]	$\sim 10^b$	Allred et al. [35]	-3.6 ± 0.2	calorimetric
Cuta et al. [44]	6.4^b	Vanderzee [41]	-3.67 ± 0.07	calorimetric
Arkhipova et al. [45]	5.05			
Hayon et al. [46]	6.3			
Teder [47]	7.9			
Krunchak et al. [48]	6.61			

[a] Uncorrected.

[b] Corrections were applied. The Cuta et al. [44] original extrapolation to zero ionic strength led to $K=(7.1\pm0.5)\times10^{-8}$ mol kg^{-1}.

From table 3a we select a value of K_B equal to $(6.5\pm0.5)\times10^{-8}$ mol kg^{-1}. As was done for process (A), values of ΔH_B^o were calculated from the variation of K_B with temperature using a constant value of $\Delta C_p^o=-262\pm14$ J mol^{-1} K^{-1} (C_p^o of $SO_3^{2-}(aq)=-264\pm10$ J mol^{-1} K^{-1} from Allred et al. [35]) and the Clarke and Glew model. The values of ΔH_B^o obtained in this way are in sharp disagreement with the recent calorimetric measurements of Allred et al. [35], Vanderzee [41], and Zambonin and Jordan [38]. The calorimetric values are preferred and a value of -3.65 ± 0.10 kJ mol^{-1} is adopted for ΔH_B^o based on the first two calorimetric values. Thus for process (B) at 298.15 K, the tentative selected values are: $K=(6.5\pm0.5)\times10^{-8}$ mol kg^{-1}. $\Delta G^o=41.02\pm0.20$ kJ mol^{-1}, $\Delta H^o=-3.65\pm0.10$ kJ mol^{-1}, and $\Delta C_p^o=-262\pm14$ J mol^{-1} K^{-1}. It will later be necessary to modify ΔG^o to 40.94 ± 0.20 kJ mol^{-1} to obtain thermodynamic consistency in a thermodynamic network, resulting in a predicted K_B of $(6.72\pm0.5)\times10^{-8}$ mol kg^{-1}.

4.3 Process (C)

The relatively few available equilibrium data for process (C), the formation of $S_2O_5^{2-}$(aq) from HSO_3^-(aq), are summarized in table 4a. The molar absorbance of $S_2O_5^{2-}$(aq) obtained by Connick et al. [52] was used to calculate values of K_C from the measurement data obtained by the earlier workers [49–51]. A value of $K_C=0.032\pm0.01$ mol^{-1} kg is adopted from the measurements of Connick et al. [52] who also obtained $\Delta H_C^\circ=4.6\pm4.0$ kJ mol^{-1} (table 4). A value of $\Delta C_p^\circ=-21\pm32$ J mol^{-1} K^{-1} is estimated for process (C) based upon an estimate of -100 ± 25 J mol^{-1} K^{-1} for the partial molar heat capacity of $S_2O_5^{2-}$(aq). Thus for process (C) at 298.15 K: $K=0.032\pm0.01$ mol^{-1} kg, $\Delta G^\circ=8.53\pm0.80$ kJ mol^{-1}, $\Delta H^\circ=-4.6\pm4.0$ kJ mol^{-1}, and $\Delta C_p^\circ=-21\pm32$ J mol^{-1} K^{-1}.

5. Vapor-liquid Equilibrium Data and the Henry's Law Constant for SO₂

There are two types of vapor-liquid-equilibrium measurements for the SO_2+H_2O system which were considered, namely, total pressure (SO_2+H_2O) measurements and partial pressure (SO_2) measurements. Knowing the total stoichiometric molality of SO_2 in a given solution, eqs (2,3) can be solved for $\hat{m}\{SO_2^0(aq)\}$. If the partial pressure of the SO_2(g) over that aqueous solution is known, the equilibrium constant for process (D), the Henry's Law constant, is calculated as

$$K_D=\hat{a}\{SO_2^0(aq)\}/f\{SO_2(g)\} \qquad (7)$$

In the above equation, the fugacity (f) of the SO_2(g) has been taken to be equal to its pressure and the activity of SO_2^0(aq) equal to its molality, i.e., $a\{SO_2^0(aq)\}=\hat{m}\{SO_2^0(aq)\}=\alpha\{SO_2^0(aq)\}m^{st}$. The latter assumption is consistent with eq (4). The former assumption is justified since (i) calculations showed that the effect of including the fugacities as calculated from available second virial coefficients [53] perturbed the calculated values of K_D by less than 1.3% up to a stoichiometric molality of 1.0 mol kg^{-1}, (ii) the values of K_D calculated from most of the data sets examined were extrapolated to zero mol kg^{-1}, and (iii) the scatter in the data sets does not justify this small correction.

Values of the Henry's Law constant (K_D) have also been calculated from measurements of total vapor pressures over aqueous SO_2 solutions. To do this calculation, the vapor pressure of the water over these solutions was calculated from the activity of the water using a chemical equilibrium model [3]. The partial pressure of SO_2(g) is obtained from

$$p\{SO_2(g)\}=p(\text{total})-p\{H_2O(g)\} \qquad (8)$$

The Henry's Law constant was then calculated using eq (7). Not considered here are the results of the few investigations summarized by references [54,55] which are of low precision. The Henry's Law constants obtained from the analysis of the data are given in table 5. Values of ΔG° and ΔH° for process (D) were calculated (see table 6) using the Clarke and Glew equation with ΔC_p° fixed at 155 J mol^{-1} K^{-1}.

We believe that the most reliable of the vapor-liquid-equilibrium investigations is that of Rabe and Harris [62]. The reasons are: (i) they measured the partial pressures of SO_2(g) rather than total pressures, (ii) the precision of their measurement is very good, (iii) the ΔH° calculated from their measurements is close to a calorimetrically determined value which will be discussed shortly, and (iv) they took care to minimize systematic errors due to analyses of the gas phase and aqueous solutions and also due to oxidation of the solutions and absorption of CO_2(g). The results of Vosolobe et al. [66], also based on partial pressure measurements, are close to those of Rabe and Harris [62]. We adopt a value of $\Delta G^\circ=-0.51\pm0.10$ kJ mol^{-1} for process (D). This corresponds to a Henry's Law constant of 1.23 ± 0.05 mol kg^{-1} bar^{-1} at 298.15 K. The assigned uncertainty is large enough to overlap with the mean value obtained from all the ΔG° values for process (D) summarized in table 6 which are based upon the partial and total pressure measurements.

Table 4. Thermodynamic parameters for process (C): $2\,HSO_3^-(aq)=S_2O_5^{2-}(aq)+H_2O(l)$.

a. Equilibrium constant at 298.15 K		***b. Enthalpy change at 298.15 K***		
Investigator(s)	K/mol^{-1} kg	Investigator(s)	ΔH°/kJ mol $S_2O_5^{2-}$(aq))$^{-1}$	Method
Golding [49]	~0.02[a]	Hayon et al. [46]	−3.3	K vs T
Arkhipova and Chistyakova [50]	~0.02[a]	Connick et al. [52]	−4.6	K vs T
Bourne et al. [51]	~0.02[a]			
Connick et al. [52]	0.032			

[a] Corrections were applied.

Table 5. Henry's Law constants (K) for process (D): $SO_2(g) = SO_2^0(aq)$.

T/K	K/mol kg^{-1} bar^{-1}
Data of Beuschlein and Simenson [56] at 0.51 g SO_2/100 g H_2O:	
308.75	0.80
314.15	0.71
320.15	0.59
325.15	0.53
335.75	0.38
344.15	0.31
351.75	0.26
358.15	0.23
365.15	0.20
372.15	0.18
378.35	0.16
384.95	0.14
386.15	0.13
Data of Beuschlein and Simenson [56] at 1.09 g SO_2/100 g H_2O:	
299.95	1.30
306.75	0.98
312.55	0.77
317.35	0.67
323.75	0.56
334.75	0.42
340.55	0.38
346.55	0.33
352.35	0.29
359.55	0.25
371.55	0.22
373.75	0.19
379.75	0.17
Data of Beuschlein and Simenson [56] at 4.36 g SO_2/100 g H_2O:	
296.35	1.28
300.35	1.15
303.15	1.03
306.15	0.93
309.15	0.85
311.35	0.79
315.35	0.71
318.35	0.65
321.35	0.60
325.15	0.55
328.45	0.51
331.55	0.47
334.35	0.44
336.15	0.42
340.35	0.38
343.95	0.36
345.55	0.35
Data of Beuschlein and Simenson [56] at 7.45 g SO_2/100 g H_2O:	
298.35	1.37
304.55	1.27
307.35	1.14
310.55	1.07
314.15	0.99
317.15	0.91
320.35	0.84
323.55	0.77
327.55	0.69
328.75	0.66
Data of Byerley [57]:	
298.15	1.26
323.15	0.78
Data of Campbell and Maass [29]:	
298.15	1.20
303.15	1.02
313.15	0.75
323.15	0.57
333.15	0.45
343.15	0.36
353.15	0.30
363.15	0.25
373.15	0.21
383.15	0.18
393.16	0.16
Data of Conrad and Beuschlein [58]:	
298.15	1.17
Data of Douabul and Reilly [59]:	
278.97	3.40
283.12	3.03
288.10	2.50
292.98	2.06
298.15	1.73
303.25	1.47
Data of Hudson [60]:	
283.15	2.35
288.15	1.85
293.15	1.54
303.15	1.07
313.15	0.75
Data of Hudson [60] cont'd:	
321.15	0.60
333.15	0.44
343.15	0.36
353.15	0.29
363.15	0.25
Data of Johnstone and Leppla [13]:	
298.15	1.20
308.15	0.85
323.15	0.57
Data of Maass and Maass [61]:	
283.15	2.50
289.65	2.00
295.15	1.63
298.15	1.50
300.15	1.38
Data of Morgan and Maass [31]:	
273.15	3.30
283.15	2.16
291.15	1.53
298.15	1.20
Data of Rabe and Harris [62]:	
303.15	1.02
313.15	0.76
323.15	0.58
333.15	0.45
343.15	0.36
353.15	0.30
Data of Tokunaga [63]:	
283.15	2.57
293.15	2.11
303.15	1.46
313.15	1.07
Data of Vosolobe et al. [64]:	
293.15	1.40
303.15	1.07
313.15	0.80
323.15	0.64
333.15	0.53

Table 6. Thermodynamic parameters for process (D), $SO_2(g)=SO_2^0(aq)$, at 298.15 K. ΔC_p^o was constrained to be equal to 155 J mol^{-1} K^{-1}.

Worker(s)	ΔG^o/kJ mol^{-1}	ΔH^o/kJ mol^{-1}
Hudson [60]	-0.581 ± 0.029	-27.31 ± 0.29
Maass and Maass [61]	-0.980 ± 0.053	-23.5 ± 1.9
Campbell and Maass [29]	-0.541 ± 0.097	-27.28 ± 0.65
Morgan and Maass [31]	-0.45 ± 0.12	-25.6 ± 2.2
Conrad and Beuschlein [58]	-0.389	
Johnstone and Leppla [13]	-0.435 ± 0.096	-25.7 ± 2.0
Beuschlein and Simenson [56]		
at 0.51 g SO_2/100 g H_2O:	-0.66 ± 0.15	-30.39 ± 0.98
at 1.09 g SO_2/100 g H_2O:	-0.74 ± 0.12	-28.99 ± 0.89
at 4.36 g SO_2/100 g H_2O:	-0.495 ± 0.022	-26.11 ± 0.27
at 7.45 g SO_2/100 g H_2O:	-0.97 ± 0.15	-22.6 ± 2.4
pooled:	-0.72 ± 0.24	-27.0 ± 3.4
Rabe and Harris [62]	-0.512 ± 0.068	-26.41 ± 0.67
Vosolobe et al. [64]	-0.475 ± 0.075	-22.2 ± 1.2
Tokunaga [63]	-1.26 ± 0.42	$-21.9\pm11.$
Douabul and Reilly [59]	-1.367 ± 0.079	-23.8 ± 2.0
Byerley [57]	-0.573	
Grand average[a]	-0.57 ± 0.10	-25.6 ± 1.2

[a] Discarding the results of Douabul and Reilly [59] and of Tokunaga [63].

6. The Calorimetric Enthalpy of Solution of $SO_2(g)$ in Water

Enthalpies of solution of $SO_2(g)$ in water have been measured several times [28,36,37,65–68]. Roth and Zeumer [68] have summarized the results of the earlier and not very precise investigations by Berthelot [66], Thomsen [65], and Lindner [28]. These early measurements will not be used. Instead, only the measured enthalpies of solution of $SO_2(g)$ in water from the more recent investigations by Johnson and Sunner [37], Stiles and Felsing [36], Ramsetter and Hantke [67], and Roth and Zeumer [68] will be considered. The measurements were treated in two different ways. The first used all of the parameters for processes (A), (B), and (C) to calculate [3] values of the excess Gibbs energy (G^{ex}) as a function of temperature from which the excess enthalpy (H^{ex}) was obtained using

$$H^{ex}=G^{ex}-T(\partial G^{ex}/\partial T)_p \tag{9}$$

The relative apparent molar enthalpy (L_ϕ) is equal to H^{ex}/m^{st}. It is a stoichiometric quantity.

In table 1b values of ΔH_A^o were obtained from the data of Johnson and Sunner [37], Stiles and Felsing [36], and Dobrogowska and Hepler [40]. These values were obtained using a minimization, or least-squares calculation. As a byproduct of that calculation, values of the B parameter were also inferred.

Using the previously adopted value of $\Delta H_A^o=-17.80$ kJ mol^{-1} and a value of B equal to 2.0 in the above procedure for calculating L_ϕ, the enthalpy of solution data were treated to obtain a value of $\Delta H_D^o=-27.00\pm0.30$ kJ mol^{-1}. Figure 1 shows the graphical treatment of the data; the calculations are shown in table 7. To obtain thermodynamic consistency with later calculations, this value is adjusted to -26.97 ± 0.30 kJ mol^{-1}. A value of $\Delta H_E^o=-44.77\pm0.50$ kJ mol^{-1} is obtained.

The above procedure is essentially equivalent to using the experimental heat of solution data at 298.15 K and correcting for the enthalpies of ionization of all of the species. Thus

$$\Delta H_{meas}=\Delta H_D^o+[1-\alpha\{SO_2^0(aq)\}][\Delta H_A^o + L_\phi(H^+\cdot HSO_3^-(aq))] \tag{10}$$

where $L_\phi(H^+\cdot HSO_3^-)$ is the relative apparent molar enthalpy for the (hypothetical) solution consisting only of the ions $H^+(aq)$ and $HSO_3^-(aq)$. This L_ϕ is different than the stoichiometric L_ϕ above. This type of procedure has been used previously by Wu and Young [69]. In eq (10) contributions to the enthalpy of $S_2O_5^{2-}(aq)$ and $SO_3^{2-}(aq)$ were neglected since they are negligible in comparison to that from $HSO_3^-(aq)$. In applying eq (10) below the electrostatic contribution $[L_\phi(H^+\cdot HSO_3^-)$ is negligible in comparison to the other terms and was neglected.

An alternative procedure is to use the above relationship (eq (10)) and to plot the experimental enthalpies of solution as a function of α. In such a plot, the slope yields a value of ΔH_A^o and the intercept at $\alpha=1$

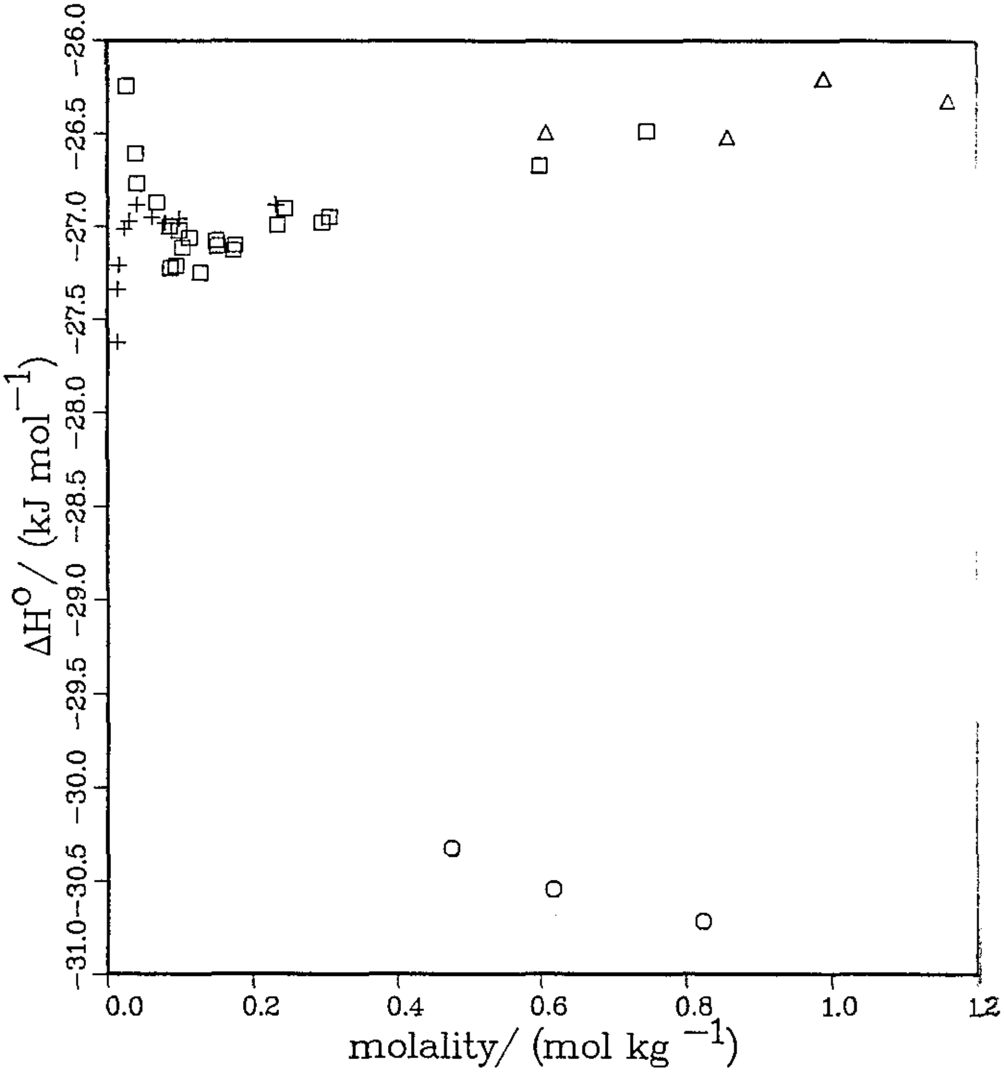

Figure 1-Values of ΔH° at 298.15 for process (D), $SO_2(g) = SO_2^0(aq)$, calculated from the heat of solution measurements of Johnson and Sunner [37] (+), Stiles and Felsing [36] [□], Ramstetter and Hantke [67] (○), and Roth and Zeumer [68](Δ). The final selected value of ΔH_D° is -26.97 ± 0.30 kJ mol^{-1}.

yields a value of ΔH_D°.The results of Stiles and Felsing [36] lead to $\Delta H_A^\circ = -17.1 \pm 1.2$ kJ mol^{-1} and $\Delta H_D^\circ = -27.1 \pm 0.8$ kJ mol^{-1}. The Johnson and Sunner [37] measurements, neglecting the three data points at the lowest concentrations, lead to $\Delta H_A^\circ = -18.00 \pm 0.34$ kJ mol^{-1} and $\Delta H_D^\circ = -26.76 \pm 0.40$ kJ mol^{-1}. This independent computational method serves as additional confirmation of the correctness of our selected values of ΔH_A° and ΔH_D°. The enthalpies of solution as a function of $\alpha(SO_2^0(aq))$ are shown in figure 2.

7. The Properties of $SO_3^{2-}(aq)$

The summation of processes (B) and (E) is:

$$SO_2(g) + H_2O(l) = 2H^+(aq) + SO_3^{2-}(aq) \qquad (F)$$

This process is useful for obtaining the thermodynamic properties of $SO_3^{2-}(aq)$.

7.1 The Enthalpy of Process (F)

The enthalpy changes already obtained for processes (B) and (E) lead to a value of $\Delta H_F^\circ = -48.42 \pm 0.50$ kJ mol^{-1}. Additional, direct experimental information leading to this enthalpy change is available from measurements on the enthalpy of solution of $SO_2(g)$ in either aqueous NaOH or KOH, represented as:

$$SO_2(g) + 2\ OH^-(aq) = SO_3^{2-}(aq) + H_2O(l) \qquad (F1)$$

Use of the relative apparent molar enthalpy data of Vanderzee and Noll [70] for aqueous Na_2SO_3 and the relative apparent molar enthalpy data tabulated by Parker [71] for aqueous KOH and NaOH leads to values of -164.4 ± 0.23, -162.3 ± 0.30, and -161.1 ± 2.5 kJ mol^{-1} for ΔH_{F1}° from the measurement data of Ramstetter and Hantke [67], Roth and Zeumer [68], and

Table 7. Enthalpy of solution data for $SO_2(g)$ in water at 298.15 K. The experimental results (ΔH_{meas}) are given in column six and the corrected results, ΔH° for process (D), are given in column seven. $\Delta H^\circ_{\mathrm{A}}$ has been constrained to equal -17.80 kJ mol^{-1}. The B parameter was set at 2.0.

$m^{\mathrm{st}}(H_2SO_3)$ mol kg^{-1}	$\alpha(HSO_3^-)$	$\alpha(SO_3^{2-})$	$\alpha(SO_2^0)$	$2\,\alpha(S_2O_5^{2-})$	ΔH_{meas} kJ mol^{-1}	ΔH°
Data of Stiles and Felsing [36]						
0.02594	0.5496	0.4074×10^{-5}	0.4497	0.6292×10^{-3}	-35.928	-26.243
0.03839	0.4880	0.2900×10^{-5}	0.5113	0.7533×10^{-3}	-35.196	-26.606
0.04074	0.4789	0.2755×10^{-5}	0.5204	0.7728×10^{-3}	-35.196	-26.767
0.06745	0.4047	0.1785×10^{-5}	0.5943	0.9468×10^{-3}	-33.982	-26.867
0.08470	0.3736	0.1469×10^{-5}	0.6254	0.1029×10^{-2}	-33.560	-26.997
0.08625	0.3711	0.1447×10^{-5}	0.6278	0.1036×10^{-2}	-33.744	-27.223
0.09383	0.3601	0.1346×10^{-5}	0.6389	0.1067×10^{-2}	-33.535	-27.210
0.09778	0.3547	0.1300×10^{-5}	0.6442	0.1083×10^{-2}	-33.250	-27.020
0.1020	0.3493	0.1254×10^{-5}	0.6496	0.1099×10^{-2}	-33.246	-27.112
0.1118	0.3377	0.1159×10^{-5}	0.6612	0.1134×10^{-2}	-32.987	-27.058
0.1273	0.3218	0.1038×10^{-5}	0.6771	0.1183×10^{-2}	-32.895	-27.247
0.1485	0.3036	0.9108×10^{-6}	0.6951	0.1243×10^{-2}	-32.401	-27.073
0.1504	0.3022	0.9015×10^{-6}	0.6965	0.1248×10^{-2}	-32.405	-27.102
0.1728	0.2866	0.8015×10^{-6}	0.7121	0.1303×10^{-2}	-32.154	-27.127
0.1754	0.2849	0.7914×10^{-6}	0.7137	0.1309×10^{-2}	-32.095	-27.097
0.2334	0.2549	0.6219×10^{-6}	0.7436	0.1425×10^{-2}	-31.459	-26.990
0.2435	0.2507	0.6002×10^{-6}	0.7478	0.1442×10^{-2}	-31.296	-26.901
0.2947	0.2324	0.5112×10^{-6}	0.7660	0.1522×10^{-2}	-31.049	-26.976
0.3050	0.2293	0.4967×10^{-6}	0.7692	0.1536×10^{-2}	-30.966	-26.948
0.5983	0.1742	0.2828×10^{-6}	0.8240	0.1829×10^{-2}	-29.719	-26.669
0.7461	0.1589	0.2354×10^{-6}	0.8391	0.1929×10^{-2}	-29.263	-26.482
Data of Ramstetter and Hantke [67]						
0.4744	0.1917	0.3432×10^{-6}	0.8066	0.1726×10^{-2}	-33.681	-30.324
0.6160	0.1721	0.2760×10^{-6}	0.8260	0.1843×10^{-2}	-33.556	-30.543
0.8223	0.1526	0.2171×10^{-6}	0.8454	0.1974×10^{-2}	-33.388	-30.718
Data of Roth and Zeumer [68]						
0.6066	0.1732	0.2796×10^{-6}	0.8249	0.1836×10^{-2}	-29.522	-26.490
0.8566	0.1500	0.2099×10^{-6}	0.8480	0.1992×10^{-2}	-29.142	-26.517
0.9876	0.1413	0.1866×10^{-6}	0.8566	0.2058×10^{-2}	-28.673	-26.201
1.1587	0.1321	0.1635×10^{-6}	0.8658	0.2133×10^{-2}	-28.543	-26.320
Data of Johnson and Sunner [37]						
0.01330	0.6566	0.7401×10^{-5}	0.3430	0.4417×10^{-3}	-39.212	-27.623
0.01416	0.6466	0.7003×10^{-5}	0.3529	0.4579×10^{-3}	-38.752	-27.340
0.01544	0.6328	0.6489×10^{-5}	0.3667	0.4807×10^{-3}	-38.376	-27.209
0.02272	0.5708	0.4626×10^{-5}	0.4286	0.5895×10^{-3}	-37.074	-27.013
0.02932	0.5301	0.3706×10^{-5}	0.4692	0.6670×10^{-3}	-36.309	-26.970
0.03964	0.4830	0.2854×10^{-5}	0.5162	0.7639×10^{-3}	-35.384	-26.882
0.06039	0.4204	0.1986×10^{-5}	0.5787	0.9076×10^{-3}	-34.342	-26.950
0.07829	0.3842	0.1590×10^{-5}	0.6148	0.1001×10^{-2}	-33.731	-26.981
0.09772	0.3548	0.1316×10^{-5}	0.6441	0.1083×10^{-2}	-33.187	-26.956
0.2312	0.2559	0.6344×10^{-6}	0.7427	0.1421×10^{-2}	-31.367	-26.881

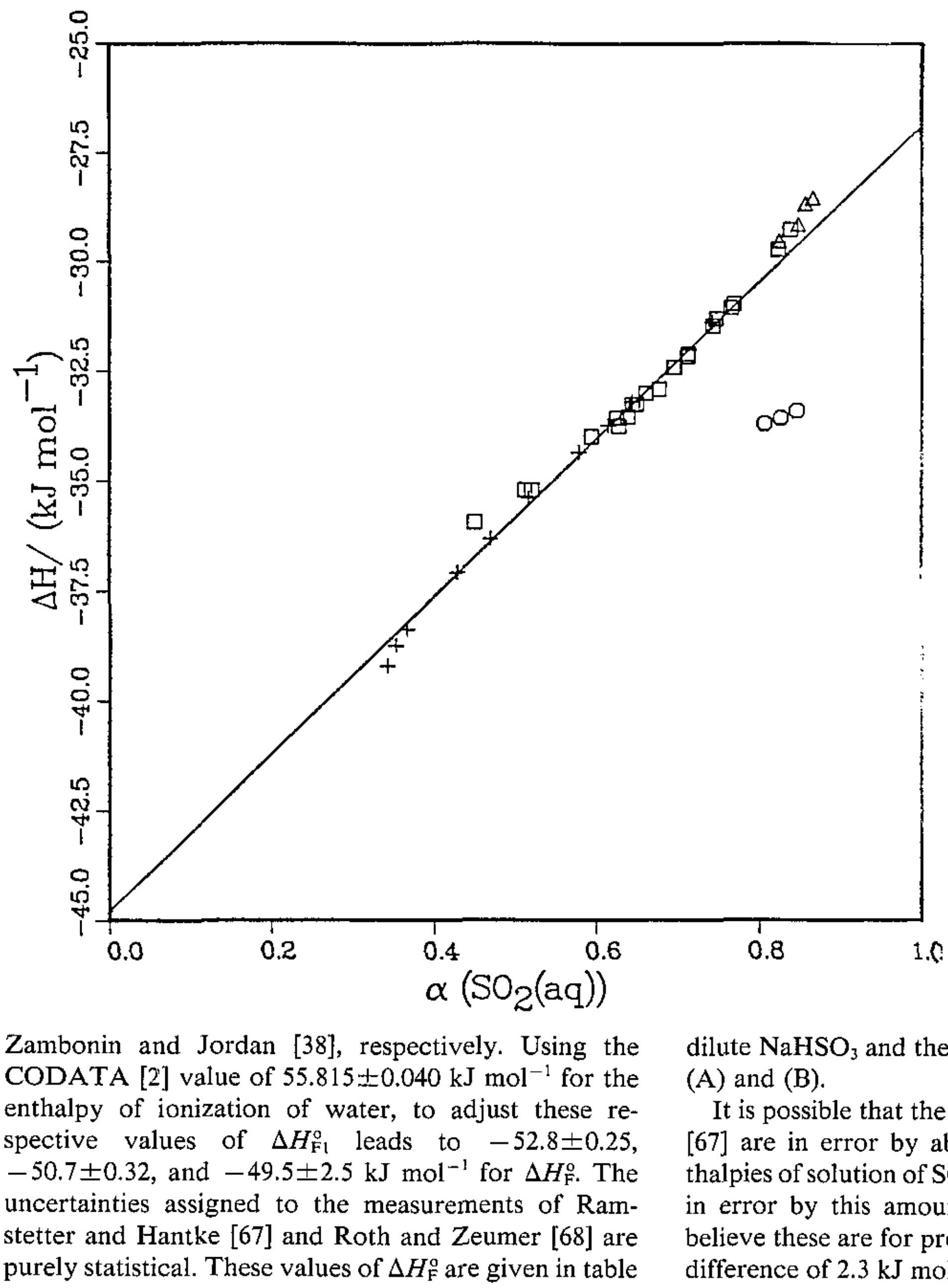

Figure 2-Experimental enthalpies of solution of $SO_2(g)$ in water at 298.15 K as a function of $\alpha(SO_2^0(aq))$. The measurements are those of Johnson and Sunner [37] (+), Stiles and Felsing [36] (□), Ramstetter and Hantke [67] (○), and Roth and Zeumer [68] (Δ). The straight line connects the selected values of $\Delta H_D^\circ = -26.97$ and $\Delta H_E^\circ = -44.77$ kJ mol^{-1}.

Zambonin and Jordan [38], respectively. Using the CODATA [2] value of 55.815 ± 0.040 kJ mol^{-1} for the enthalpy of ionization of water, to adjust these respective values of ΔH_{F1}° leads to -52.8 ± 0.25, -50.7 ± 0.32, and -49.5 ± 2.5 kJ mol^{-1} for ΔH_F°. The uncertainties assigned to the measurements of Ramstetter and Hantke [67] and Roth and Zeumer [68] are purely statistical. These values of ΔH_F° are given in table 8 together with the value obtained from Zambonin and Jordan's [38] measured enthalpy of solution of $SO_2(g)$ in dilute $NaHSO_3$ and their measured values for processes (A) and (B).

Table 8. ΔH° at 298.15 K for process (F): $SO_2(g) + H_2O(l) = SO_3^{2-}(aq) + 2\ H^+(aq)$.

Investigator(s)	ΔH°/kJ mol^{-1}
Ramstetter and Hantke [67]	-52.8 ± 0.23
Roth and Zeumer [68]	-50.7 ± 0.30
Zambonin and Jordan [38]	-49.5 ± 2.5
Zambonin and Jordan [38][a]	-47.36 ± 0.9
This evaluation	-48.42 ± 0.50

[a] Derived from the authors' measurements of the heat of solution of $SO_2(g)$ in $NaHSO_3(aq)$ and the enthalpy data for processes (A) and (B).

It is possible that the result of Ramstetter and Hantke [67] are in error by about 4 kJ mol^{-1} since their enthalpies of solution of $SO_2(g)$ in water (see figs. 1,2) were in error by this amount in the same direction as we believe these are for process (F). We cannot explain the difference of 2.3 kJ mol^{-1} between the ΔH° for process (F) obtained from Roth and Zeumer [68] and the tentative selection. As can be seen, the results of Zambonin and Jordan [38] bracket this selection. We believe that our selected value is best and needs no serious adjustment. This will become apparent in the next section. A modern, accurate measurement of ΔH° for process (F1) would be of value in confirming our selection.

7.2 The Standard State Entropy of $SO_3^{2-}(aq)$

7.2.1 From the Tentative Selections for Process (F)

Use of the values of ΔG° tentatively selected for processes (B) and (E) leads to a value of $\Delta G_F^\circ = 51.11 \pm 0.24$

kJ mol^{-1}. This value together with $\Delta H_F^\circ = -48.42\pm0.50$ kJ mol^{-1} leads to $\Delta S_F^\circ = -333.83\pm1.9$ J mol^{-1} K^{-1}. The CODATA [2] values for $S^\circ(H_2O,l)$ (69.950±0.030 J mol^{-1} K^{-1}) and $S^\circ(SO_2,g)$ (adjusted to 248.223±0.05 J mol^{-1} K^{-1} at a standard state pressure of one bar) are then used to calculate a value of -15.66 ± 2.0 J mol^{-1} K^{-1} for the standard state partial molar entropy of SO_3^{2-}(aq).

7.2.2 From Data on the Na_2SO_3 System

There is a direct path to the standard state partial molar entropy of SO_3^{2-}(aq) from the Third Law entropy of Na_2SO_3(cr) and the value of ΔG° and ΔH° for the process:

$$Na_2SO_3(cr) = 2\ Na^+(aq) + SO_3^{2-}(aq). \qquad (G)$$

The stable solid phase in equilibrium with saturated Na_2SO_3 solution at 298.15 K is the heptahydrate. The solubility measurements of Foerster et al. [72] and Kobe and Hellwig [73] lead to a solubility of 2.42±0.10 mol kg^{-1} at 298.15 K. The activity of water, 0.908±0.002, and the mean ionic activity coefficient, 0.190±0.005, at 2.42 mol^{-1} kg^{-1} are obtained by extrapolation of the results of the evaluation of Goldberg [74]. These result in $\Delta G^\circ = 4.016\pm0.20$ kJ mol^{-1} for the process:

$$Na_2SO_3{\cdot}7\ H_2O(cr) = 2\ Na^+(aq) + SO_3^{2-}(aq) + 7\ H_2O(l). \qquad (G1)$$

Arii [75] reports vapor pressure measurements over $Na_2SO_3{\cdot}7\ H_2O(c)$:

$$\log p(\text{mm of Hg}) = 10.656 - 2797.1/T. \qquad (11)$$

These data, after correction for the nonideality of the water vapor, and with the CODATA ΔG° value for the vaporization of $H_2O(l)$, leads to $\Delta G^\circ = 4.070\pm0.15$ kJ mol^{-1} for the process:

$$Na_2SO_3{\cdot}7\ H_2O(cr) = Na_2SO_3(cr) + 7\ H_2O(l). \qquad (G2)$$

The vapor pressure measurements of Tarassenkow [76] over $Na_2SO_3{\cdot}7\ H_2O(cr)$:

$$\log p(\text{mm of Hg}) = 9.949 - 2608/T \qquad (12)$$

were rejected since they lead to an unreasonable ΔS° for the decomposition process (G2). Combining ΔG_{G1}° and ΔG_{G2}° we obtain a value of $\Delta G_G^\circ = -0.054\pm0.25$ kJ mol^{-1} at 298.15 K.

Above 308 K the stable solid phase in equilibrium with an aqueous solution of Na_2SO_3 is the anhydrous salt. Extrapolation to 298.15 K of the solubility measurements of Foerster et al. [72] and of Kobe and Hellwig [73] leads to a solubility of 3.26±0.05 mol kg^{-1}. With an exptrapolated [74] mean ionic activity coefficient of 0.19±0.03, $\Delta G_G^\circ = 0.13\pm0.6$ kJ mol^{-1} is obtained, confirming the result obtained from the pathway involving the heptahydrate.

The recent enthalpy of solution measurements of Vanderzee [41] yield $\Delta H_G^\circ = -13.26\pm0.04$ kJ mol^{-1} and are preferred to the earlier measurements of de Forcrand [77] and Kennedy and Lister [78] which yield $\Delta H_G^\circ = -14.9\pm2.0$ kJ mol^{-1} and -18.62 ± 0.4 kJ mol^{-1}, respectively.

With $\Delta H_G^\circ = -13.26\pm0.04$ kJ mol^{-1} and $\Delta G_G^\circ = -0.054\pm0.25$ kJ mol^{-1} a value of $\Delta S_G^\circ = -44.29\pm0.85$ kJ mol^{-1} is obtained. Use of $S^\circ\{Na_2SO_3(cr)\} = 145.94\pm1.2$ J mol^{-1} K^{-1} at 298.15 K [79] and the CODATA entropy for Na^+(aq) of 58.45±0.15 J mol^{-1} K^{-1}, results in a partial molar entropy of -15.25 ± 1.5 J mol^{-1} K^{-1} for SO_3^{2-}(aq). This value is in very good agreement with the value of -15.66 ± 2.0 J mol^{-1} K^{-1} obtained from the SO_2(g) cycle.

8. The Oxidation of SO_2(aq) to H_2SO_4(aq)

There are two calorimetric determinations involving the oxidation of SO_2(aq) that were used by the CODATA "Key Values" task group [2] as input for their evaluation of $\Delta_f H^\circ$ of SO_4^{2-}(aq). Reversing their procedure, we use the final CODATA value of $\Delta_f H^\circ(SO_4^{2-}(aq)) = -909.34\pm0.40$ kJ mol^{-1} and these reactions to obtain $\Delta_f H^\circ(H_2SO_3(\text{in } 2500\ H_2O))$. The reactions investigated by Johnson and Sunner [37] and by Johnson and Ambrose [80] are, respectively:

$$Br_2(l) + 2\ H_2O(l) + SO_2(\text{in } 2500\ H_2O) = H_2SO_4(\text{in } 2500\ H_2O) + 2\ HBr(\text{in } 1250\ H_2O);\quad \Delta H = -232.09\pm0.42 \text{ kJ mol}^{-1}$$

and

$$Cl_2(g) + 2\ H_2O(l) + SO_2(\text{in } 2500\ H_2O) = H_2SO_4(\text{in } 2500\ H_2O) + 2\ HCl(\text{in } 1250\ H_2O);\quad \Delta H = -323.34\pm0.60 \text{ kJ mol}^{-1}$$

Using $\Delta_f H^\circ$ (HBr(in 1250 H_2O)) $= -121.06\pm0.15$ kJ mol^{-1} and $\Delta_f H^\circ$ (HCl(in 1250 H_2O)) $= -166.695\pm0.10$ kJ mol^{-1} from CODATA [2] and Parker [66] and $\Delta_f H^\circ$ (H_2SO_4(in 2500 H_2O)) $= -895.43\pm0.40$ kJ mol^{-1} from

CODATA and Wu and Young [69] we obtain $\Delta_f H°(SO_2(\text{in } 2500\ H_2O)) = -333.80 \pm 0.60$ kJ mol^{-1} and -333.82 ± 0.65 kJ mol^{-1}, respectively.

Independently, from the present evaluation we obtain $\Delta_{sol}H = -37.14 \pm 0.20$ kJ mol^{-1} for $SO_2(g) \rightarrow SO_2(\text{in } 2500\ H_2O)$ which leads to $\Delta_f H°(SO_2(\text{in } 2500\ H_2O)) = -333.95 \pm 0.28$ kJ mol^{-1}. This agreement, well within the assigned uncertainties, substantiates the selections made here. More importantly, however, the enthalpy relationships between aqueous solutions of SO_2 and $H_2SO_4(aq)$ over a range of concentrations are well defined.

9. Final Selected Values for the Processes and Properties

The very near agreement in the partial molar entropy of $SO_3^{2-}(aq)$ obtained via two independent thermochemical pathways may be fortuitous and the result of a cancellation of errors. Nevertheless, it serves to confirm the selections made for the processes used in obtaining the partial molar entropy of $SO_3^{2-}(aq)$. Tables 9 and 10 contain, respectively, our recommended values for the processes and for the thermodynamic properties of the species. These values are consistent with the forthcoming CODATA "Key Values" [2] adjusted to a standard state pressure of one bar. Note that we have adopted a final "best" value of -15.40 ± 0.8 J mol^{-1} K^{-1} for $S°(SO_3^{2-},aq)$ and adjusted the value of K_B to be $(6.7_2 \pm 0.5) \times 10^{-8}$ mol kg^{-1} rather than the value of $(6.5 \pm 0.5) \times 10^{-8}$ mol kg^{-1} used in section 4.2. In tables 9 and 10 the uncertainties have also been adjusted in light of the agreement of the data. The thermodynamic parameters for process (C) are not as well known as those for the other processes. Consequently, the property values for $S_2O_5^{2-}(aq)$ have much larger uncertainties than have been assigned to the other species.

Table 9. Recommended values for the processes involving sulfur(IV)-oxygen species at 298.15 K.

The value of the equilibrium constant for a given process can be calculated using the equation: $R\ ln\ K = -\Delta G°_{298.15}/298.15 + \Delta H°_{298.15}[1/298.15 - 1/T] + \Delta C°_p[(298.15/T) - 1 + ln(T/298.15)] + (298.15/2)(d\Delta C°_p/dT)[(T/298.15) - (298.15/T) - 2\ ln(T/298.15)]$.

Process		$\Delta G°$	$\Delta H°$	$\Delta C°_p$	$d\Delta C°_p/dT$[a]
		kJ mol^{-1}		J mol^{-1} K^{-1}	J mol^{-1} K^{-2}
$SO_2^0(aq) + H_2O(l) = H^+(aq) \pm HSO_3^-(aq)$	(A)	10.60 ± 0.10	-17.80 ± 0.40	$-272. \pm 10$	1.7[a]
$HSO_3^-(aq) = H^+(aq) + SO_3^{2-}(aq)$	(B)	40.94 ± 0.20	-3.65 ± 0.10	$-262. \pm 14$	−2.7[a]
$2\ HSO_3^-(aq) = S_2O_5^{2-}(aq) + H_2O(l)$	(C)[b]	8.53 ± 0.80	-4.6 ± 4.0	$-21. \pm 25$[a]	−1.9[a]
$SO_2(g) = SO_2^0(aq)$	(D)	-0.51 ± 0.10	-26.97 ± 0.30	$155. \pm 10$	−0.035[a]
$SO_2(g) + H_2O(l) = H^+(aq) + HSO_3^-(aq)$	(E)	10.09 ± 0.14	-44.77 ± 0.40	$-117. \pm 14$	+1.7[a]

[a] Estimated.

[b] For process (C), all values refer to one mole of $S_2O_5^{2-}(aq)$.

Table 10. Recommended thermodynamic property values at 298.15 K in SI units and at a standard state pressure of one bar (0.1 MPa).

Species	$\Delta_f H°$	$\Delta_f G°$	$S°$	$C°_p$
	kJ mol^{-1}		J mol^{-1} K^{-1}	
S(cr, rhombic)	0	0	32.054 ± 0.050[a]	22.686 ± 0.050
$SO_2(g)$	-296.81 ± 0.20[a]	-300.09 ± 0.21[a]	248.223 ± 0.05[a]	39.842 ± 0.020
$SO_2^0(aq)$	-323.78 ± 0.32	-300.60 ± 0.23	159.48 ± 0.75	$195. \pm 10$
$S_2O_5^{2-}(aq)$	-973.6 ± 4.1	-808.61 ± 0.94	$154. \pm 13.$	-100.0 ± 25[b]
$HSO_3^-(aq)$	-627.41 ± 0.32	-527.14 ± 0.25	134.17 ± 0.65	$-2. \pm 10$
$H_2SO_3^0(aq)$[c]	-609.61 ± 0.32	-537.74 ± 0.23	229.43 ± 0.75	$270. \pm 10$
$SO_3^{2-}(aq)$	-631.06 ± 0.40	-486.20 ± 0.33	-15.40 ± 0.80	$-264. \pm 10$
$Na_2SO_3(cr)$	-1098.48 ± 0.42	-1010.08 ± 0.46	145.94 ± 0.80	120.25
		Auxiliary Values[a]		
$O_2(g)$	0	0	205.152 ± 0.005	29.378 ± 0.003
$H_2(g)$	0	0	130.680 ± 0.005	28.836 ± 0.002
$H_2O(l)$	-285.830 ± 0.04	-237.141 ± 0.04	69.950 ± 0.030	75.300
Na(cr)	0	0	51.30 ± 0.20	28.23 ± 0.20
$Na^+(aq)$	-240.34 ± 0.06	-261.95 ± 0.10	58.45 ± 0.15	—

[a] CODATA [2] selections.

[b] Estimated.

[c] Convention, property values are set equal to the sum of those of $SO_2^0(aq)$ and $H_2O(l)$.

The property values given in table 10 for HSO_3^-(aq), SO_3^{2-}(aq), and $S_2O_5^{2-}$(aq) are based upon a more current data base than are the property values given in earlier evaluations [1,32,81,82]; for this reason they are to be preferred.

Table 11 gives properties ($\gamma_\pm$, ϕ, L_ϕ) and $p(SO_2)$, the partial pressure of SO_2(g) over aqueous solutions of SO_2 as a function of the stoichiometric molality. The values of L_ϕ are obtained as a composite of experimental heat of dilution data [40], heat of solution data [36,37], and the use of the equilibrium model (sec. 6) to calculate values of L_ϕ using the selected property values for processes (A) (B), and (C). The properties $\gamma_\pm$, ϕ, and L_ϕ all have contributions from the various species in the solution. L_ϕ is a bulk, stoichiometric property and is the difference in enthalpy between the real solution and the hypothetical standard state solution consisting of H^+(aq) and HSO_3^-(aq). The activity and osmotic coefficients were calculated using a speciation model of the solution and are expressed in table 11 for the final treatment of aqueous SO_2 solutions as a uni-univalent electrolyte. The excess Gibbs energy is given by $G^{ex}=\nu[m^{st}(H_2SO_3)]RT(ln\ \gamma_\pm-\phi+1)$, where $\nu=2$.

Values of $\gamma_\pm$ can also be calculated from the spectroscopic data of Huss and Eckert [24] if one assumes a value of K_A and a value of the B parameter in eq (5) and applies the equation:

$$K_A=[\hat{\gamma}(H^+)\hat{\gamma}(HSO_3^-)]\hat{m}(H^+)\hat{m}(HSO_3^-)] \div[\hat{m}(SO_2)a_w] \qquad (13)$$

to calculate the product $[\hat{\gamma}(H^+)\hat{\gamma}(HSO_3^-)]$. Since γ_i^{st} is equal to the quantity $\gamma_i\hat{m}_i/m_i^{st}$, the (stoichiometric) mean *ionic activity coefficient* ($\gamma_\pm$) *can be calculated using:*

$$\gamma_\pm=[\gamma(H^+)\hat{\gamma}(HSO_3^-)\hat{m}(H^+)\hat{m}(HSO_3^-)] \div[m^{st}(H^+)m^{st}(HSO_3^-)] \qquad (14)$$

The average deviation between the values of $\gamma_\pm$ calculated in this way using the spectroscopic data of Huss and Eckert [24] and the values of $\gamma_\pm$ calculated from the equilibrium model is 0.0026.

Table 11. Stoichiometric thermodynamic properties of aqueous sulfur dioxide solutions at 298.15 K. The values of $\gamma_\pm$, ϕ, and L_ϕ are relative to (H^+) (HSO_3^-).

$m^{st}(SO_2)$ mol kg^{-1}	$m^{st}(H_2SO_3)$	$\gamma_\pm$	ϕ	L_ϕ kJ mol^{-1}	$p(SO_2)$ bar (0.1 MPa)
0.001000	0.001000	0.909	0.960	1.09	0.0000483
0.002000	0.002000	0.855	0.934	1.91	0.000171
0.003000	0.003000	0.812	0.914	2.58	0.000347
0.004000	0.004000	0.777	0.897	3.14	0.000566
0.005000	0.005000	0.748	0.883	3.61	0.000818
0.006000	0.006001	0.722	0.870	4.05	0.00110
0.007000	0.007001	0.699	0.859	4.43	0.00140
0.008000	0.008001	0.679	0.849	4.78	0.00173
0.009000	0.009001	0.661	0.840	5.08	0.00207
0.01000	0.01000	0.644	0.832	5.36	0.00243
0.02000	0.02000	0.533	0.777	7.36	0.00664
0.03000	0.03001	0.468	0.745	8.54	0.0115
0.04000	0.04002	0.424	0.723	9.36	0.0168
0.05000	0.05004	0.391	0.706	9.99	0.0224
0.06000	0.06006	0.366	0.693	10.46	0.0282
0.07000	0.07008	0.345	0.682	10.85	0.0341
0.08000	0.08011	0.327	0.674	11.17	0.0402
0.09000	0.09014	0.312	0.666	11.44	0.0464
0.1000	0.1001	0.300	0.659	11.66	0.0527
0.1500	0.1504	0.253	0.635	12.56	0.0850
0.2000	0.2007	0.224	0.620	13.14	0.119
0.3000	0.3016	0.188	0.600	13.89	0.188
0.4000	0.4029	0.165	0.588	14.33	0.258
0.5000	0.5045	0.149	0.579	14.64	0.330
0.6000	0.6065	0.137	0.572	14.89	0.403
0.7000	0.7089	0.127	0.566	15.08	0.477
0.8000	0.8116	0.120	0.561	15.24	0.551
0.9000	0.9148	0.113	0.556	15.37	0.625
1.0000	1.0183	0.107	0.553	15.48	0.700

10. The Effects of Perturbations of Parameters on Calculated Quantities in the Equilibrium Modelling of SO_2+H_2O

The model used to describe the SO_2+H_2O system used several thermodynamic parameters: ΔG°, ΔH°, and ΔC°_p at 298.15 K for processes (A), (B), and (C), and a B parameter (see eq (4)). It is of interest to examine the effects of perturbations in these parameters on the quantities which can be calculated from the model, namely $\alpha(SO_2^0)$, $\gamma_\pm$, ϕ, and L_ϕ. The examination of the effects of these perturbations will be confined to the reference temperature of 298.15 K.

The results of these calculations are summarized in table 12. Each parameter has been perturbed by the limit of error assigned to that parameter in table 9; the B parameter has been perturbed by 0.5. As a result of these calculations, the values of K_A and B are found to be most significant in calculating value of $\alpha(SO_2^0)$, $\gamma_\pm$, and ϕ. In the calculation of L_ϕ, the most important quantity is ΔH°_A followed by K_A and B.

The results of these calculations are useful in estimating uncertainties in the thermodynamic properties that were calculated from the equilibrium model. Thus the calculated values of K_D are uncertain by at least 1 to 3% because of uncertainties in the calculation of $\alpha(SO_2^0)$. The uncertainties in $\gamma_\pm$, ϕ, and L_ϕ are given in the last column of table 12; uncertainties in $\gamma_\pm$ range from 0.2 to 1.4%, in ϕ from 0.1 to 0.3%, and in L_ϕ from 2 to 3%. This sensitivity analysis does not consider possible coupling effects between the parameters varied.

11. Extension of the Temperature Range Over Which Properties and Equilibrium can be Calculated

In the model used, it was assumed that the heat capacities of the species were independent of temperature. This was necessary since we have no direct knowledge of the temperature dependency of the heat capacities of

Table 12. Percentage effects on calculated quantities ($\alpha(SO_2^0)$, $\gamma_\pm$, ϕ, and L_ϕ) due to perturbations in the parameters of the model used to describe the thermodynamics of aqueous SO_2 solutions. The parameters which were perturbed were B in eq (4) (1.5 instead of 2.0), K_A (0.0143 instead of 0.0139 mol kg^{-1}), K_B (7.2×10^{-8} mol kg^{-1} instead of 6.7×10^{-8} mol kg^{-1}), K_C (0.042 instead of 0.032), ΔH°_A (−18.20 instead of −17.80 kJ mol^{-1}), ΔH°_B (−3.75 instead of −3.65 kJ mol^{-1}), and ΔH°_C (−8.60 instead of −4.60 kJ mol^{-1}). The temperature is 298.15 K in all cases. A discussion of the effects of the B parameter on the value of ΔH°_A is given in section 4.1.

m	Parameters Modified							
mol kg^{-1}	B	K_A	K_B	K_C	ΔH°_A	ΔH°_B	ΔH°_C	total in quadrature
				Effects on $\alpha(SO_2^0)$				
0.001	−0.08	−2.5	-3.3×10^{-5}	−0.0018	—	—	—	2.5
0.01	−0.34	−1.5	-8.3×10^{-6}	−0.0093	—	—	—	1.5
0.10	−0.50	−0.63	-1.3×10^{-6}	−0.022	—	—	—	0.80
1.00	−0.47	−0.23	-2.0×10^{-7}	−0.035	—	—	—	0.52
				Effects on $\gamma_\pm$				
0.001	−0.044	+0.16	-1.7×10^{-5}	−0.00091	—	—	—	0.16
0.01	−0.17	+0.64	-3.8×10^{-6}	−0.0047	—	—	—	0.66
0.10	−0.25	+1.1	-5.8×10^{-7}	−0.011	—	—	—	1.1
1.00	−0.23	+1.3	-8.5×10^{-8}	−0.017	—	—	—	1.3
				Effects on ϕ				
0.001	−0.021	+0.075	+0.028	−0.00045	—	—	—	0.08
0.01	−0.07	+0.26	$+3.5\times10^{-5}$	−0.00045	—	—	—	0.27
0.10	−0.05	+0.28	$+6.0\times10^{-6}$	−0.0040	—	—	—	0.28
1.00	−0.021	−0.15	$+1.0\times10^{-6}$	−0.0050	—	—	—	0.15
				Effects on L_ϕ				
0.001	−0.028	−2.4	-1.8×10^{-3}	−0.0056	+2.1	-6.6×10^{-4}	−0.011	3.2
0.01	−0.27	−1.5	-4.8×10^{-5}	−0.014	+2.3	-1.6×10^{-5}	−0.014	2.8
0.10	−0.44	−0.61	-3.5×10^{-6}	−0.027	+2.3	-1.1×10^{-6}	−0.018	2.4
1.00	−0.42	−0.22	-5.0×10^{-7}	−0.041	+2.3	-1.3×10^{-7}	−0.025	2.4

these species. However, using the "correspondence principle" of Criss and Cobble [83] and the parameters given by them in their table 1, the quantitiy (dC_p°/dT) can be estimated for the aqueous species HSO_3^-, SO_3^{2-}, and $S_2O_5^{2-}$. Doing this, we obtain $(dC_p^\circ/dT) = +1.7$, -1.0, and $+1.5$ J mol^{-1} K^{-2} for HSO_3^-, SO_3^{2-}, and $S_2O_5^{2-}$, respectively. From data on the heat capacity of water and of $SO_2(g)$ [7,82], $(dC_p^\circ/dT) = +0.0066$ J mol^{-1} K^{-2} for $H_2O(l)$ and $(dC_p^\circ/dT) = +0.035$ J mol^{-1} K^{-2} for $SO_2(g)$; we estimate $(dC_p^\circ/dT) = 0$ for $SO_2^0(aq)$. These estimates are combined to yield values of $+1.7$, -2.7, -1.9, -0.035, and $+1.7$ J mol^{-1} K^{-2} for $(d\Delta C_p^\circ/dT)$ for processes (A), (B), (C), (D), and (E), respectively. These values are also given in table 9.

If eq (6) is used without including a $(d\Delta C_p^\circ/dT)$ term, values of $K_A = 0.00153$ mol kg^{-1} and $K_D = 0.213_3$ mol kg^{-1} bar^{-1} at 373.15 K are calculated. Inclusion of the $(d\Delta C_p^\circ/dT)$ terms and the estimates of them given above leads to $K_A = 0.00171$ mol kg^{-1} and $K_D = 0.212_8$ mol kg^{-1} bar^{-1} at 373.15 K. Thus, over the temperature range 273 to 373 K the thermodynamics of the $SO_2 + H_2O$ system are reasonably well described by the parameters given in tables 9 and 10. A plot of the measured Henry's Law constants as a function of temperature and the curve calculated using the final selected values given in table 9 are shown in figure 3. A particularly useful series of experiments would be the measurement of the heat capacities of aqueous sulfur dioxide solutions from 273 K to temperatures greater than 373 K.

We thank Drs. Celina Dobrogowska, Loren Hepler, Peter Tremaine, and Cecil Vanderzee for sharing the results of their research with us prior to publication and for their helpful discussions.

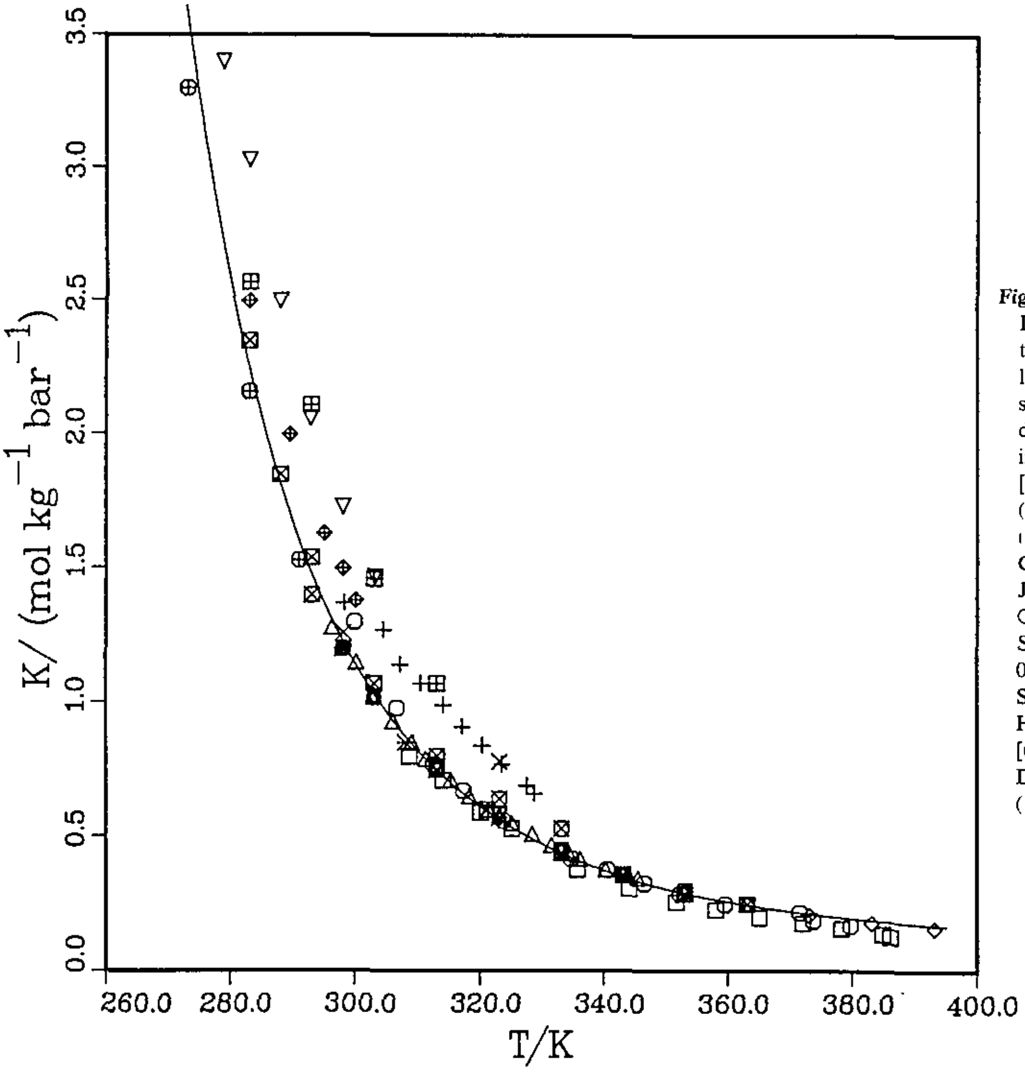

Figure 3–The measured Henry's Law constants as a function of temperature and the curve (solid line) calculated using the final selected values in table 9. The data sets and their corresponding symbols are: (⊠) Hudson [60]; (◈) Maass and Maass [61]; (◇) Campbell and Maass [29]; (⊕) Morgan and Maass [31]; (⧄) Conrad and Beuschlein [58]; (✳) Johnstone and Leppla [13]; (□, O, Δ, and +) Beuschlein and Simenson [56] at, respectively, 0.51, 1.09, 4.36, and 7.45 g SO_2/100 g H_2O; (⊠) Rabe and Harris [62]; (⊗) Vosolobe et al. [64]; (⊞) Tokunaga [63]; (▽) Douabul and Reilly [59]; and (X) Byerley [57].

Glossary

Roman

a	activity
f	fugacity
m	molality/mol kg^{-1}
n	amount or number of moles of substance
p	pressure
z	charge
A_m	Debye-Hückel constant; $A_m = 1.17642$ kg$^{1/2}$ mol$^{-1/2}$ at 298.15 K
B	parameter in Debye-Hückel equation
C_p	heat capacity at constant pressure
G	Gibbs energy
H	enthalpy
I	ionic strength
K	equilibrium constant
L_ϕ	relative apparent molar enthalpy, equal to H^{ex}/m
R	gas constant; $R = 8.31448$ J mol^{-1} K^{-1}
S	entropy
T	temperature

Greek

α	fraction
γ	activity coefficient
ϕ	osmotic coefficient
Δ	change

Superscripts

ex	excess
st	stoichiometric
$\circ$	value of a quantity which applies at the thermodynamic standard state
0	designates a neutral species, i.e., SO_2^0
$\hat{}$	species quantity

Subscripts

f	formation property
ϕ	apparent molar quantity
$\pm$	mean ionic property
w	designates water

References

[1] Wagman, D. D.; W. H. Evans, V. B. Parker, I. Halow, S. M. Bailey, R. H. Schumm, K. L. Churney, and R. L. Nuttall. "The NBS Tables of Chemical Thermodynamic Properties," J. Phy. Chem. Ref. Data, Vol. 11, Supplement No. 2 (1982).

[2] CODATA Task Group on Key Values for Thermodynamics, draft final report, 1984.

[3] Goldberg, R. N., J. Res. Natl. Bur. Stand. **89:** 251 (1984).

[4] Clarke, E. C. W., and D. N. Glew, Trans. Faraday Soc. **76:** 1911 (1980).

[5] Beyer, R. P., and B. R. Staples, private communication.

[6] Uematsu, M., and E. U. Franck, J. Phys. Chem. Ref. Data **9:** 1291 (1980).

[7] Haar, L.; J. S. Gallagher, and G. S. Kell, NBS/NRC Steam Tables; Washington, DC: Hemisphere Publishing Corp. (1984).

[8] Hamer, W., and Y. C. Wu, J. Phys. Chem. Ref. Data **1:** 1047 (1972).

[9] NAG FORTRAN Library Manual, Mark 9 edition; Oxford: NAG Central Office (1982).

[10] Drucker, K., Z. Phys. Chem. (Leipzig) **49:** 563 (1904).

[11] Sherrill, M. S., and A. A. Noyes, J. Am. Chem. Soc. **48:** 1861 (1926).

[12] Britton, H. T. S., and R. A. Robinson, Trans. Faraday Soc. **28:** 531 (1932).

[13] Johnstone, H. F., and P. W. Leppla. J. Am. Chem. Soc. **56:** 2233 (1934).

[14] Yui, N., Bull. Inst. Phys. Chem. Res. **19:** 1229 (1940).

[15] Tartar, H. V., and H. H. Garretson, J. Am. Chem. Soc. **63:** 808 (1941).

[16] Frydman, M.; G. Nilsson, T. Rengemo, and L. G. Sillen, Acta Chem. Scand. **12:** 878 (1958).

[17] Ellis, A. J., and D. W. Anderson, J. Chem. Soc.: 1765 (1965).

[18] Ludemann, H. D., and E. U. Franck, Ber. Bunsenges. Phys. Chem. **72:** 523 (1968).

[19] Deveze, D., and P. Rumpf, C. R. Hebd. Seances Acad. Sci. **258C:** 6135 (1964).

[20] Flis, I. E.; G. P. Arkhipova and K. P. Mischenko, J. Appl. Chem. USSR (Eng. Trans.) **38:** 1466 (1965); Zh. Prikl. Chim. (Lengingrad) **38:** 1494 (1965).

[21] Deveze, D. C., R. Hebd. Seances Acad. Sci. **263C:** 392 (1966).

[22] Sekine, T.; H. Iwaki, M. Sakairi, F. Shimada, and M. Inarida, Bull. Chem. Soc. Japan **41:** 1 (1968).

[23] Beilke, S., and D. Lamb, Am. Inst. Chem. Eng. J. **21:** 402 (1975).

[24] Huss, A., Jr., and C. A. Eckert, J. Phys. Chem. **81:** 2268 (1977).

[25] McCrae, J., and W. E. Wilson, Z. Anorg. Allg. Chem. **35:** 11 (1903).

[26] Walden, P., and M. Centnerszwer, Z. Phys. Chem. (Leipzig) **42:** 432 (1903).

[27] Kerp, W., and E. Bauer, Arb. Gesund. **26:** 297 (1907).

[28] Lindner, J., Montash. Chem. **33:** 613 (1912).

[29] Campbell, W. B., and O. Maass, Can. J. Res. **2:** 42 (1930).

[30] Beazley, W. B.; W. B. Campbell, and O. Maass, The Physical Properties of Sulphite Liquors, Dominion Forest Service—Bulletin 93; Ottawa, Canada; J. O. Patenaude (1938).

[31] Morgan, M., and O. Maass, Can. J. Res. **5:** 162 (1931).

[32] Rossini, F. D.; D. D. Wagman, W. H., Evans, S. Levine, and I. Jaffe, Selected Values of Chemical Thermodynamic Properties—National Bureau of Standards Circular 500;

Washington, DC; U. S. Government Printing Office (1952).

[33] Clarke, E. C. W., and D. N. Glew, Trans. Faraday Soc. **62:** 539 (1966).

[34] Barbero, J.; L. G. Hepler, K. G. McCurdy, and P. R. Tremaine, Can. J. Chem. **61:** 2509 (1983).

[35] Allred, G. C.; J. W. Larson, and L. G. Hepler, Can. J. Chem. **59:** 1068 (1981).

[36] Stiles, A. G., and W. A. Felsing, J. Am. Chem. Soc. **48:** 1543 (1926).

[37] Johnson, W. H., and S. Sunner, Acta Chem. Scand. **17:** 1917 (1963).

[38] Zambonin, G., and J. Jordan, Anal. Chem. **41:** 437 (1969).

[39] Dobrogowska, C., and L. G. Hepler, J. Solution Chem., **12:** 153 (1983).

[40] Dobrogowska, C., and L. G. Hepler, personal communication.

[41] Vanderzee, C. E., personal communication.

[42] Jellinek, K., Z. Phys. Chem. (Leipzig) **76:** 257 (1911).

[43] Kolthoff, J. M., Z. Anorg. Allg. Chem. **109:** 69 (1920).

[44] Cuta, F.; E. Beranek, and J. Pisecky, Coll. Czech. Chem. Commun. **23:** 1496 (1958).

[45] Arkhipova, G. P.; I. E. Flis, and K. P. Mischenko, J. Appl. Chem. USSR (Eng. Trans.) **37:** 2275 (1964); Zh. Prikl. Khim. (Leningrad) **37:** 2306 (1964).

[46] Hayon, E.; A. Treinin, and J. Wolf, J. Am. Chem. Soc. **94:** 47 (1972).

[47] Teder, A., Sven. Papperstidn. **75:** 704 (1972).

[48] Krunchak, E. G.; A. G. Rodichev, Ya. S. Khvorostin, B. S. Krumagal'z, V. G. Krunchak, and Yu. I. Yusova, Zh. Neorg. Khim. **18:** 2859 (1973); Russ. J. Inorg. Chem. **18:** 1519 (1973).

[49] Golding, R. M.; J. Chem. Soc.: 3711 (1960).

[50] Arkhipova, G. P., and I. I. Chistyalsova, J. Appl. Chem. USSR (Eng. Trans.) **44:** 2242 (1971); Zh. Prikl. Khim (Leningrad) **44:** 193 (1971).

[51] Bourne, D. W. A.; T. Higuchi, and I. H. Pitman, J. Pharm. Sci. **63:** 865 (1974).

[52] Connick, R. E.; T. M. Tam., and E. von Deuster, Inorg. Chem. **21:** 103 (1982).

[53] Dymond, J. H., and E. B. Smith, The Virial Coefficients of Pure Gases and Mixtures. A Critical Compilation; Oxford: Clarendon Press (1980).

[54] Sherwood, T. K., Ind. Eng. Chem. **17:** 745 (1925).

[55] Van Arsdel, W. B., Partial Vapor Pressures of Aqueous Solutions of Sulfur Dioxide, in "The International Critical Tables of Numerical Data, Physics, Chemistry, and Technology; Volume III: 302; E. W. Washburn, editor, New York: McGraw Hill Book Company (1928).

[56] Beuschlein, W. L., and L. O. Simenson, J. Am. Chem. Soc. **62:** 610 (1940).

[57] Byerley, J. J.; G. L. Rempel, and V. T. Le, J. Chem. Eng. Data, **25:** 55 (1980).

[58] Conrad, F. H., and W. L. Beuschlein, J. Am. Chem. Soc. **56:** 2254 (1934).

[59] Douabul, A., and J. Riley, J. Chem. Eng. Data **24:** 274 (1979).

[60] Hudson, J.C., J.Chem. Soc.: 1332 (1925).

[61] Maass, C. E., and O. Maass, J. Am. Soc. **50:** 1352 (1928).

[62] Rabe, A. E., and J. F. Harris, J. Chem. Eng. Data **8:** 333 (1963).

[63] Tokunaga, J., J. Chem. Eng. Data **19:** 162 (1974).

[64] Vosolobe, J.; A. Simecek, J. Michalek, and B. Kadlec, Chem. Promsyl. **15:** 401 (1965).

[65] Thomsen, J., Thermochemische Untersuchungen, Vol. 2; Leipzig: Johann Ambrosium Barth (1882).

[66] Berthelot, M.; Ann. Chim. Phys. **1:** 73 (1884).

[67] Ramstetter, H., and G. Hantke, Z. Phys. Chem. Bodenstein Festband 662 (1931).

[68] Roth, W.A., H. Zeumer, Z. Elektrochem. **38:** 164 (1932).

[69] Wu, Y. C., T. F. Young, J. Res. Natl. Bur. Stand. **85:** 11 (1980).

[70] Vanderzee, C. E., and L. A. Noll, J. Chem. Thermodyn. **15:** 1015 (1983).

[71] Parker, V. B.; Thermal Properties of Aqueous Uni-univalent Electrolytes, NSRDS-NBS 2; Washington, DC: U.S. Government Printing Office (1965).

[72] Foerster, F.; A. Brosche, C. Norberg-Schulz, Z. Phys. Chem. (Leipzig) **110:** 435 (1924).

[73] Kobe, K. A., and K. C. Hellwig, Ind. Eng. Chem. **47:** 1116 (1955).

[74] Goldberg, R. N., J. Phys. Chem. Ref. Data **10:** 671 (1981).

[75] Aari, K., Rikagaku Kenkyusho Iho **7:** 891 (1928).

[76] Tarassenkow, D. N., Z. Anorg. Allg. Chem. **169:** 407 (1928).

[77] de Forcrand, M., Ann. Chim. Phys. [6] **3:** 242 (1884).

[78] Kennedy, M. B., and M. N. Lister, Thermochim. Acta **7:** 249 (1973).

[79] Kelly, K. K., and G. E. Moore, J. Am. Chem. Soc. **66:** 293 (1944).

[80] Johnson, W. H., and J. R. Ambrose, J. Res. Natl. Bur. Stand. **67A:** 427 (1963).

[81] Brewer, L., Thermodynamic Values for Desulfurization Processes, in ACS Symposium Series No. 188—Flue Gas Desulfurization; J. L. Hudson and G. T. Rochelle, editors; Washington, DC: American Chemical Society (1982).

[82] Gurvich, L. V.; I. V. Veits, V. A. Medvedev, G. A. Khachkuruzov, V. S. Yungman, G. A. Bergman, et al., Thermodynamic Properties of Individual Substances; V. P. Glushko (general editor); Moscow: Izdatel'stvo "Nauka" (1979).

[83] Criss, C. M., and J. W. Cobble, J. Am. Chem. Soc. **86:** 5385 (1964).

Journal of Research of the National Bureau of Standards
Volume 90, Number 5, September-October 1985

SRM 1970: Succinonitrile Triple-Point Standard—A Temperature Reference Standard Near 58.08 °C

B. W. Mangum

National Bureau of Standards, Gaithersburg, MD 20899

and

Samir El-Sabban

National Institute of Standards, Cairo, Egypt

Accepted: August 8, 1985

Triple-point-of-succinonitrile cells have been tested and established as Standard Reference Material (SRM) 1970. Of the 115 cells tested, 109 were accepted as SRM 1970. Five of the 115 cells had triple-point temperatures lower than 58.0785 °C (the low-temperature limit established for SRM 1970) and, consequently, were rejected. One of the 115 cells broke during tests on it. The mean value of the triple-point temperatures (obtained by freezing) of the 109 cells is 58.0796±0.0015 °C, where the uncertainty is the total estimated uncertainty relative to the International Practical Temperature Scale of 1968, Amended Edition of 1975. The standard deviation of the triple-point temperatures is 0.48 mK. The purity of the succinonitrile of the SRM 1970 cells is estimated to range from 99.999,97% to 99.999,84%. The preparation of the cells, the various tests performed on them, and the procedure recommended for their use are described.

Key words: SRM 1970; standard reference materials; succinonitrile; thermometry; temperature fixed point.

1. Introduction

Accurate temperature measurements and control are necessary for many tests conducted in clinical and biomedical laboratories in order to obtain accurate and meaningful results which form the basis for the diagnosis and treatment of diseases. Temperature fixed points at temperature values near those at which specific tests are conducted can be used as *the* reference temperatures for those tests. This simplifies temperature measurements and significantly improves the precision and accuracy of clinical measurements and results.

The calibration of thermometers is performed, either directly or indirectly, through the use of temperature fixed points. Because temperature fixed points provide reproducible environments, they have been the basis for practical temperature scales, such as the International Practical Temperature Scale of 1968, Amended Edition of 1975 (IPTS-68(75) [1][1], and they are used in the calibration of thermometers on those scales. In addition to the defining fixed points of the scales, there are other equilibrium states which are internationally recognized as secondary reference points [2]. Those points may be used for calibration of thermometers which do not operate over the entire range of the defining fixed points

About the Authors: B. W. Mangum is with the Temperature and Pressure Division in NBS' Center for Basic Standards where Samir El-Sabban served as a guest scientist.

---

[1] Numbers in brackets indicate literature references.

of the IPTS-68(75) and when a comparison calibration against a primary standard is not practicable. The feasibility of the use of the triple-point temperature of succinonitrile (SCN) as a temperature reference point near 58.08 °C has been investigated [3,4,5] and the quality of this fixed point has been found to be very high, the point containing most of the features desired in a fixed point.

Based on the previous work [4,5], we have developed an SCN triple-point Standard Reference Material (SRM). It is designated SRM 1970—the Succinonitrile Triple-Point Standard, and it is available from the Office of Standard Reference Materials (OSRM) of the National Bureau of Standards (NBS). These SRM devices are easy to use and with them one can confidently expect to achieve a calibration point with an uncertainty no greater than ±0.0015 K.

The remainder of this publication describes SRM 1970, the tests performed on the cells of SCN, the conditions under which the cells were tested, the results obtained from testing 109 SRM 1970 cells, and the procedure recommended for using the cells for calibration of thermometers.

2. Experimental Details

2.1 Preparation of SCN Triple-Point Cells

a) Cell Design

Previous experience [4,5] with SCN led to the design shown in figure 1 for the triple-point cells constituting SRM 1970. These cells, sealed under vacuum, contain approximately 60 grams of high-purity SCN and are designed to be used with small thermometers, those not exceeding 4.5 mm (0.180 inches) in diameter. The thermometer well of a cell will accommodate most precision thermistor thermometers, many platinum resistance thermometers, and most diode thermometers.

b) Purification of SCN and Filling of the Triple-Point Cells

The SCN used for SRM 1970 was purified by the zone-refining technique by Prof. M. E. Glicksman and his students at Rensselaer Polytechnic Institute. The apparatus used for the zone-refining was a modified version of that described previously by M. E. Glicksman et al. [3].

Three pyrex cells were attached to the zone-refining tube of the second stage of refining so that three cells could be prepared from the purest half of the purified material of that second stage. Each of the two stages of purification consisted of at least 20 passes through the 6 zones of the zone-refining furnace. This amount of purification had been determined previously by one of us (BWM) [6] to be the minimum required if three cells were to be routinely obtained from each zone-refined lot. During the early part of the second stage of zone-refining, the three pyrex cells were washed several times with the purified SCN.

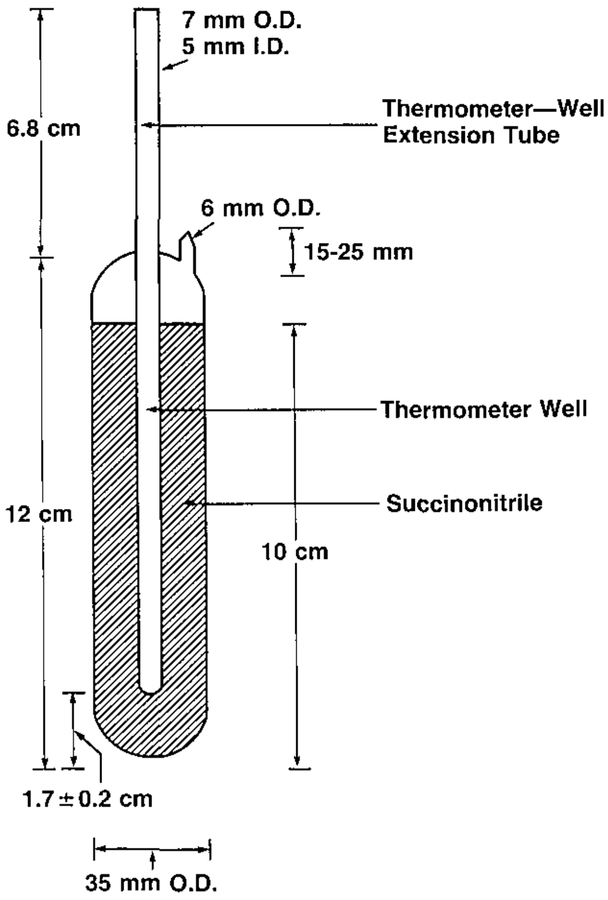

Figure 1–Cross-sectional drawing of the succinonitrile triple-point cell.

2.2 Testing of SCN Triple-Point Cells for SRM 1970

Using the inner-sheath technique, the freezing behaviors of 115 cells were studied in a well-stirred constant-tempeature oil bath maintained at 56.880 °C. Some preliminary tests were made prior to testing the cells to determine the optimum bath temperature to be used for the freezing experiments, consistent with a reasonable time for the freeze. The temperature selected was 56.880 °C.

Prior to a freezing experiment, the SCN was melted by immersing the cell in a water bath at about 80 °C. When the material had completely melted, the cell was removed from the bath and its exterior dried. The cell was inverted several times to obtain a uniform distribution of any impurities present in the sample. Then a cold copper rod was inserted into the thermometer well, which contained a small amount of light mineral oil, to initiate the growth of an SCN sheath around the well. When the sheath appeared to be 2 to 3 mm thick, the rod was removed. This sheath provided an inner liquid-solid interface which the thermometer sensed. Once the sheath had formed, the cell was placed in the constant-temperature oil bath and a thermistor thermometer was inserted into the thermometer well. A second liquid-solid interface then formed along the outer wall of the cell. After the cell was placed in the oil bath, temperature measurements were begun and they were made at regular intervals until complete solidification of the SCN in the cell. These tests were made with groups of three cells at a time.

2.3 Apparatus Used in the Realization of the Triple-Points of the SCN Cells

a) Constant Temperature Bath

A constant temperature bath was used to study the freezing behavior of all cells tested for suitability as SRM 1970. The bath has a 10.6-liter capacity; such baths are commercially available. A light mineral oil (a dimethylpolysiloxane) with a 17-centistokes viscosity at 40 °C was used as a bath medium. The bath has both a cooling and a heating system and its temperature was controlled by their combined effect. The cooling was obtained by passing a stream of pressurized air at room-temperature through the cooling coil. The heater, and consequently the whole bath, was controlled by a commercially-available proportional controller with a thermistor as a sensing element. By this means, it was possible to maintain a uniform bath temperature, constant to ±0.5 mK, for a time much longer than that needed for a freezing or a melting experiment. This high level of temperature control, however, is not necessary in the normal use of SRM 1970.

b) Thermometers

Bead-in-glass probe-type thermistor thermometers were used in this investigation. They were calibrated over the range 0 °C to 70 °C against a standard platinum resistance thermometer (SPRT), all located in a copper block immersed in a constant-temperature oil bath. The copper block, a high-capacity, good thermal conductor, ensured thermal equilibrium among the thermistors and with the SPRT, and damped any sudden temperature fluctuations. The measurements of the SPRT resistances, from which the temperatures of calibration were obtained, were made with a Cutkosky 400-Hz bridge [7] which, with a strip-chart recorder, has an equivalent temperature resolution of about 1.5 μK and an inaccuracy of about 0.01 mK. The SPRT itself had been calibrated previously on the IPTS-68(75) in the NBS Platinum Resistance Thermometer Calibration Laboratory using the same bridge. A constant-current source and high-quality $6\frac{1}{2}$-digit digital voltmeter (DVM) were used for potentiometric measurements of the thermistors. The uncertainty of the resistance measurements of the thermistor thermometers corresponded to about ±0.25 mK. By fitting the equation

$$1/T = A + B \log R + C(\log R)^2 + D(\log R)^3 \qquad (1)$$

to the data, where T is the temperature in kelvins and R is the thermistor resistance in ohms at temperature T, the constants A, B, C and D were determined. The temperature value then derived from eq (1) for a measured thermistor-thermometer resistance agreed with that measured with the SPRT to within ±1 mK.

c) Temperature Measurement System

The temperature measuring system used in testing the cells for suitability as SRM 1970 was the same automated one used in calibrating the thermistor thermometers and it consisted of a microcomputer, a $6\frac{1}{2}$-digit DVM, a constant-current source, a standard resistor, and the calibrated thermistor thermometers [5].

3. Results and Discussion

3.1 Freezing Experiments

During previous work on SCN [4,5], it was found that the fastest and easiest way of producing a temperature plateau for calibration of thermometers was by the freezing technique, using the inner sheath method, rather than by the melting technique. Adopting this technique for testing the cells for suitability as SRM 1970, we conducted freezing experiments on a cell in the oil bath maintained at 57.480 °C, 57.180 °C and 56.880 °C in order to ascertain a reasonable bath temperature at which to test all of the remaining cells. The results for three freezes of cell A-1/3 in a bath at 57.480 °C are shown in figure 2; those for the cell in a bath at 57.180 °C are shown in figure 3; and those for the cell in a bath at 56.880 °C are shown in figure 4. We depict the combined results of these freezes of A-1/3 in figure 5, which shows that although the times required for the freezes vary with the temperature of the bath, as expected, the temperatures of the plateaus are unaffected by the bath temperature. Based on these results, it was decided to test all cells in a bath maintained at 56.880 °C, a temperature that does not require an unreasonable amount of time for measurements.

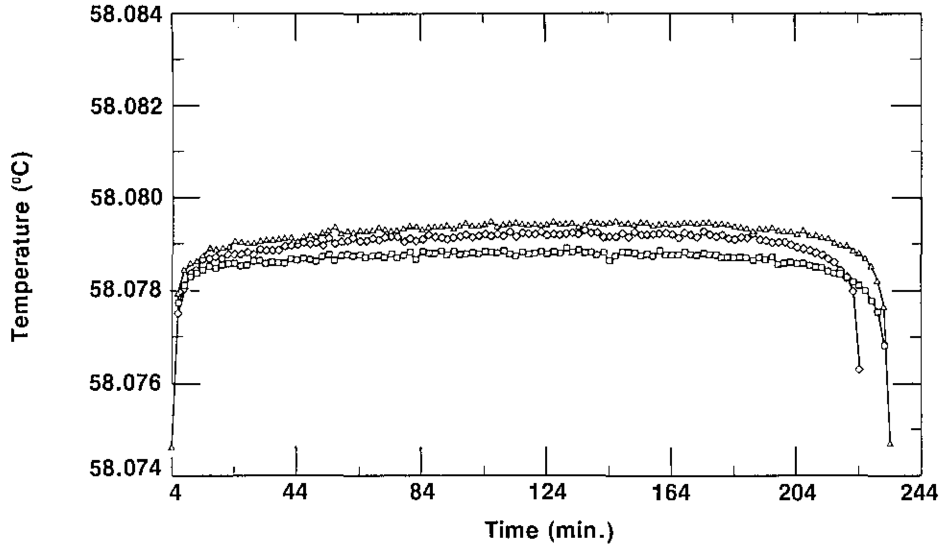

Figure 2–Freezing curves of sample A-1/3 obtained on different days, each time in an oil bath maintained at 57.480 °C.

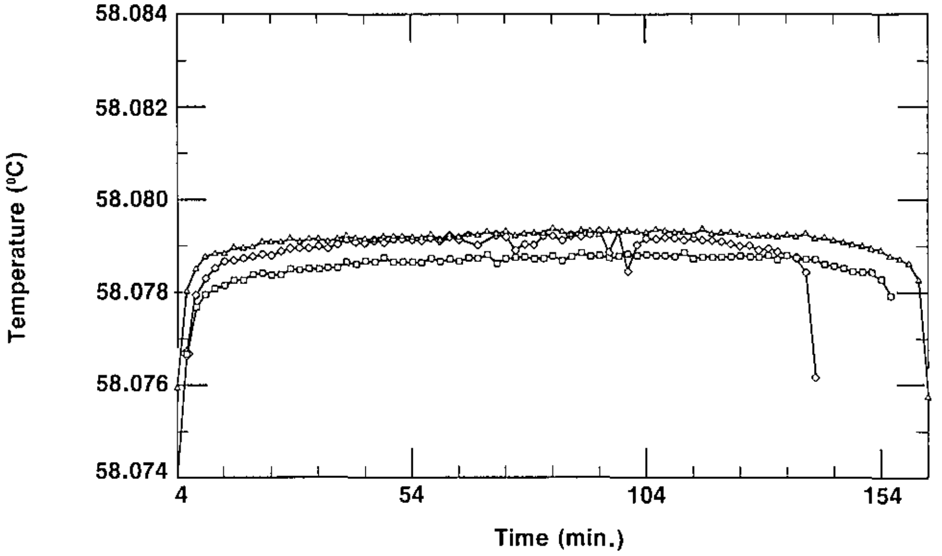

Figure 3–Freezing curves of sample A-1/3 obtained on different days, each time in an oil bath maintained at 57.180 °C.

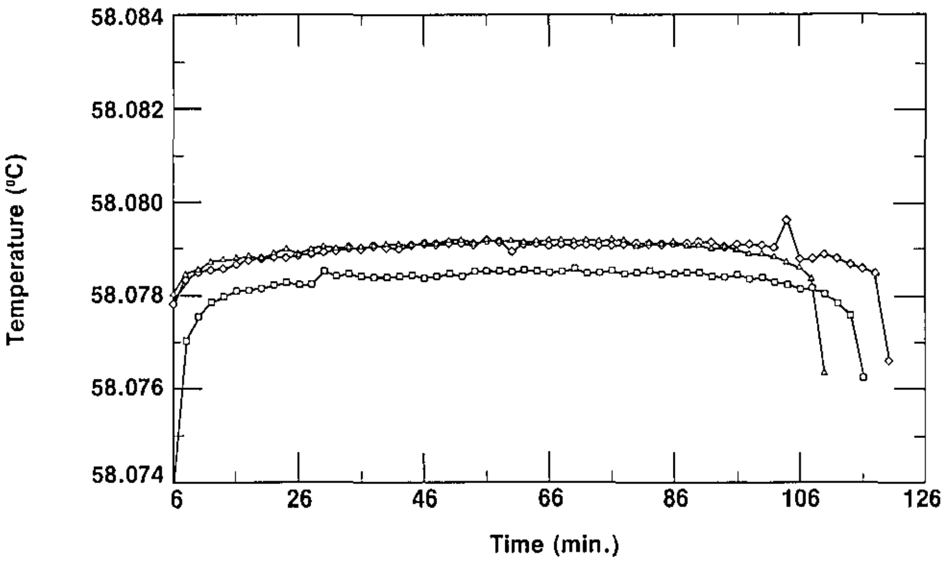

Figure 4–Freezing curves of sample A-1/3 obtained on different days, each time in an oil bath maintained at 56.880 °C.

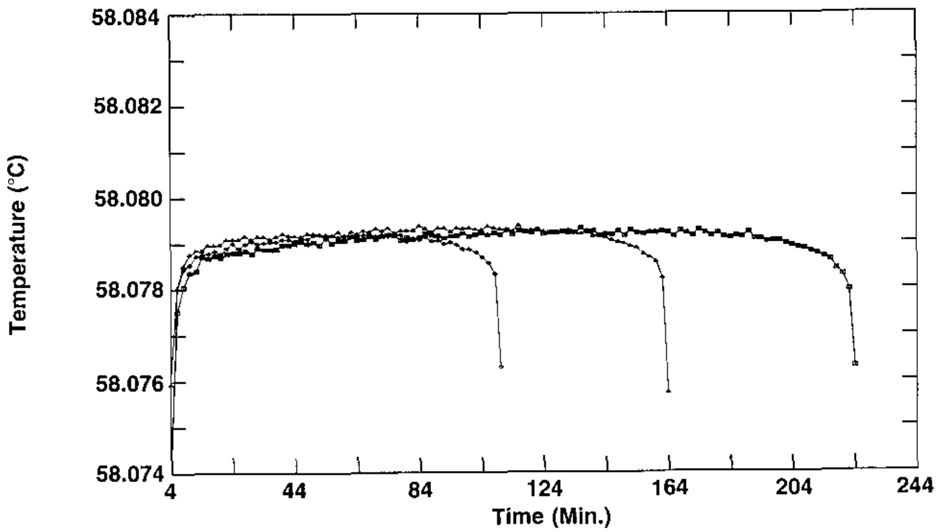

Figure 5–Freezing curves of SRM 1970 cell A-1/3 in constant temperature baths at three different temperatures. The data represented by ■ were obtained with the cell in a bath at 57.480 °C, those data represented by ▲ were obtained with the cell in a bath at 57.180 °C, and those data represented by ♦ were obtained with the cell in a bath at 56.880 °C.

Figures 2, 3, and 4 also show the irreproducibility of the freezing curves of A-1/3. The spread among the three curves of figure 2 is 0.7 mK, that for the curves of figure 3 is 0.6 mK, and that for the curves of figure 4 is 0.8 mK. We see, however, that the plateaus of the individual curves in figures 2, 3, and 4 are fairly flat and that the freezing range of any given curve is about 0.5 mK. The initial rapid rise in temperature indicated by the thermistor thermometer is due primarily to the temperature of the oil and thermistor in the thermometer well of the cell rising to the temperature of the SCN liquid-solid interface. The curvature at the ends of the freezes is due mainly to the low thermal conductivity of the SCN [8]. Also, the small but gradual increase in temperature during about the first two-thirds of any given freeze is attributed to the low thermal conductivity of SCN. The small but gradual decrease in temperature during the last one-third of a freeze is attributed to the low thermal conductivity of SCN, to increased concentration of impurities in the liquid SCN due to rejection of those impurities at the solid-liquid interface as the SCN solidifies, and to decreased effective immersion of the thermistor thermometer because of the way the outer solid-liquid interface progresses inward.

It can be seen from figure 5 that even with the temperature of the bath set 1.2 °C below the triple-point temperature of SCN, there is still a plateau lasting about 2 hours, a sufficient time for most calibrations. If longer times are required for calibrations, the bath temperature must be set closer to the triple-point value of the SCN.

As indicated in section 2, the column of the second stage of the zone-refining apparatus normally contained enough SCN to fill three cells from the top half (the purest part and that which constituted a "lot") of the column. The cells were given a designation ending with -1/X, -2/X or -3/3. The letter X represents either the number 2, indicating that only two cells (-1/2 and -2/2) were filled from that lot, or the number 3, indicating that three cells (-1/3, -2/3 and -3/3) are in the lot. The first cell to be filled would normally be considered to be the purest of a lot, the second cell to be filled the next purest, and the third cell to be filled the least pure. In some cases, however, one could envision cells designated -2/X being purer than cells designated -1/X due to the washing of the empty (contaminated?) cells and the impurity not being removed from the first portion of the SCN in the subsequent zone refining. The spread of the triple-point temperatures of the cells of SCN of a given lot is an indication of the relative purity of those cells. Figures 6, 7, and 8 show freezing curves of three groups of cells, each group constituting a lot. All of the freezing curves were obtained at the same bath temperature using the same technique of freezing. The behavior portrayed in these figures is typical of all the lots of SRM 1970. Figure 6 displays freezing curves of cells L28-1/3, L28-2/3, and L28-3/3; these cells belong to lot L28. The temperatures of the plateaus of these cells have a spread of 1.3 mK. Figures 7 and 8 display the curves obtained during freezing of cells L25-1/3, L25-2/3 and L25-3/3, which belong to lot L25, and cells L34-1/3, L34-2/3 and L34-3/3, which belong to lot L34, respectively. The curves obtained during freezing of lot L25 show a spread of about 0.1 mK while those of lot L34 have a spread of about 0.8 mK. We see that all of the curves of figures 6, 7, and 8 are rather flat, with freezing ranges of about 0.5 mK, ranges which are due primarily to curvature near the ends of the freezes. These small freezing ranges indicate high purity.

We see from figures 6, 7, and 8 that the differences in temperature values of the plateaus of the different cells of a given lot are small but the temperatures are, nevertheless, not exactly the same. This is primarily because

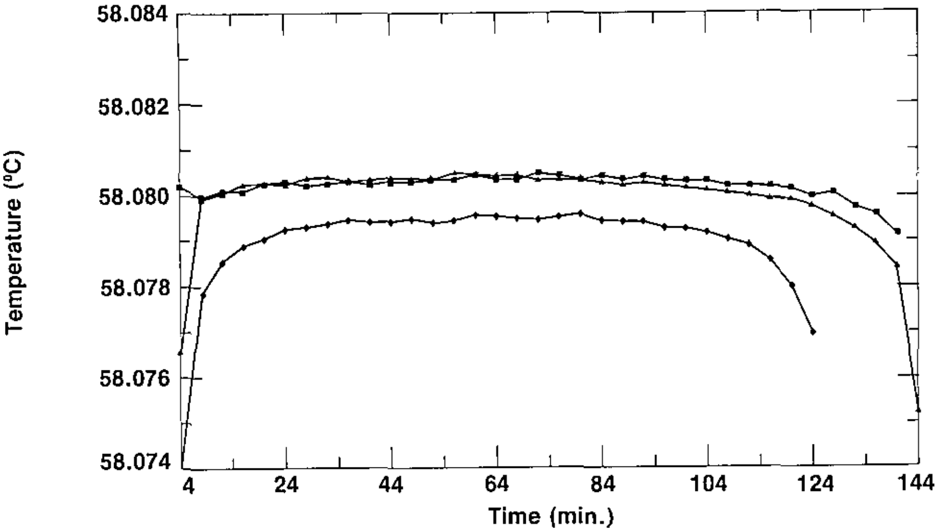

Figure 6–Freezing curves of the three cells of lot L28 obtained in a bath at 56.880 °C. ■ represent data for L28-1/3, ▲ represent *data for* L28-2/3, *and* ♦ *represent* data for L28-3/3.

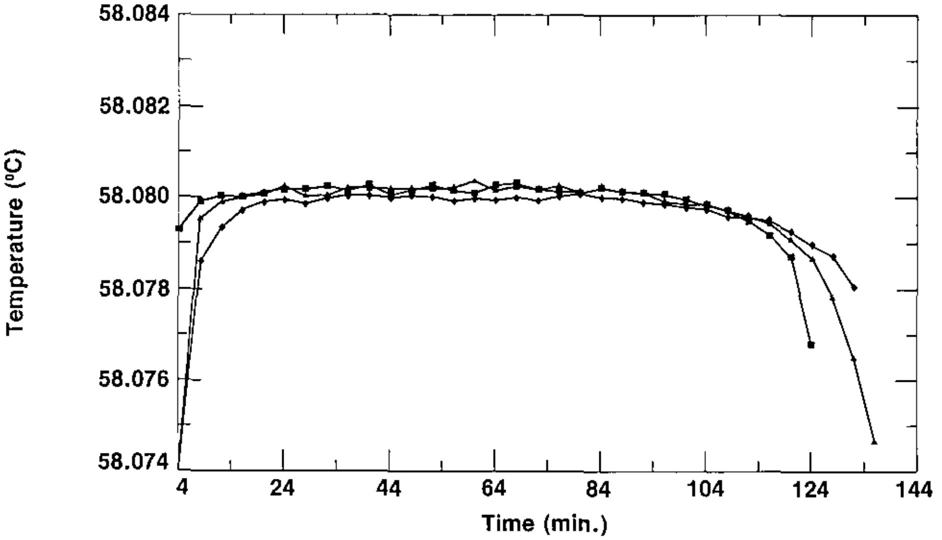

Figure 7–Freezing curves of the three cells of lot L25 obtained in *a bath at 56.880 °C.* ■ represent data for L25-1/3, ▲ represent data for L25-2/3, and ♦ represent data for L25-3/3.

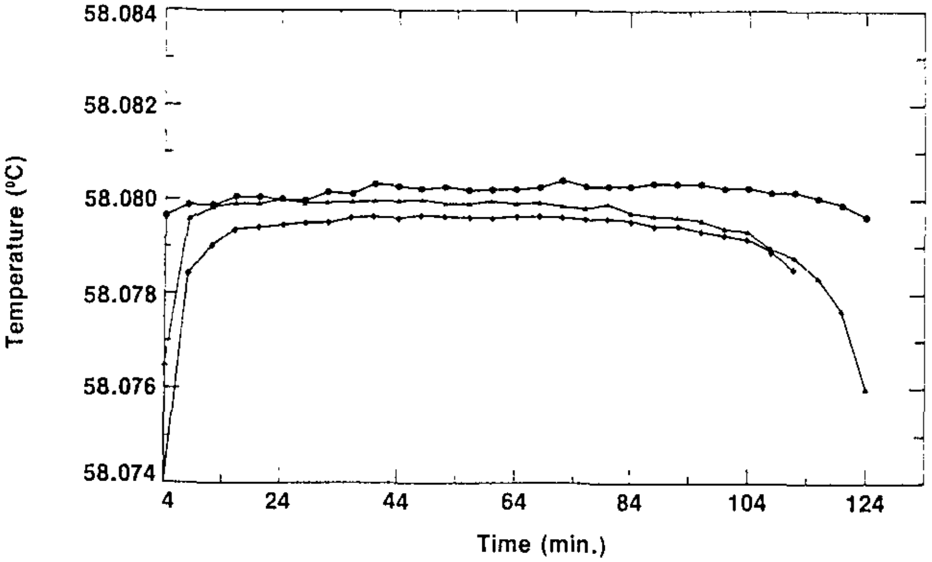

Figure 8–Freezing curves of the three cells of lot L34 obtained in a bath at 56.880 °C. ● represent data for L34-1/3, ▲ represent data for L34-2/3, and ♦ represent data for L34-3/3.

the cells were filled with SCN from different sections on the same zone-refining column and the purity varies along the column. The results for the cells of lot L34, shown in figure 8, as well as the results for cells of lots L28 and L25, shown in figures 6 and 7, respectively, demonstrate that the purities of the cells belonging to a given lot have different values and that generally they are in the sequence discussed above. The freezing curves of the three samples of lot L34, taken from the same zone-refining tube, clearly demonstrate how impurities affect the triple-point values of the SCN. It is seen in figure 8 that the first sample has a plateau during freezing which has the highest temperature of any sample of that lot. The third cell, L34-3/3, was the last sample in that lot to be extracted from the column and it has a plateau with the lowest temperature, while sample L34-2/3 is purer since it was the second sample to be taken from the column. This distribution of temperatures of the plateaus is as expected from purification of a material by zone refining.

The spread of the temperature values of the plateaus of the curves obtained during freezing of the cells of each of the 39 lots of SRM 1970 cells are shown in figure 9. This shows that the spread ranges from 0.1 mK to 1.3 mK. Only four lots, however, have a spread in temperature exceeding 0.9 mK, and those have a spread of 1.3 mK. These results indicate that the cells of SRM 1970 are of high purity.

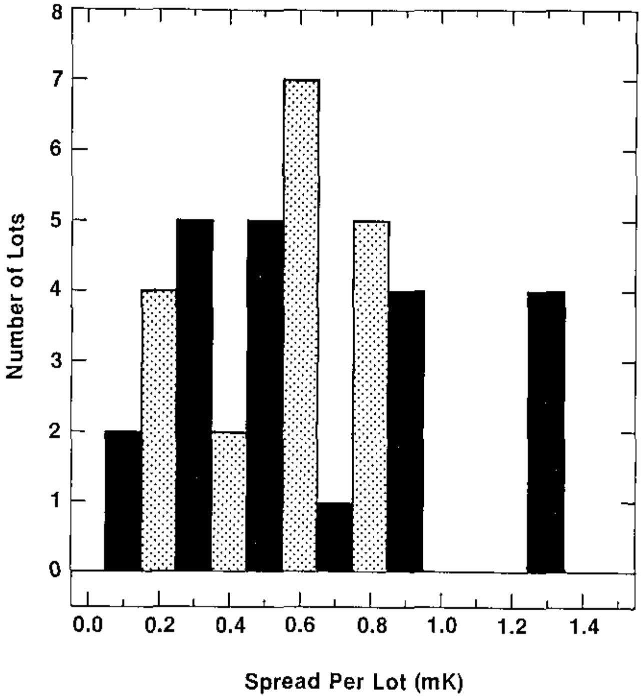

Figure 9–Histogram of the spread of the temperatures of the plateaus of the freezing curves of the cells of a given lot.

Our results indicate that in general, then, the purity of the cells of a given lot is such that the first extracted sample (comprising the -1/X cell) is the purest, the second cell filled is the next purest, and the last cell (the third cell) will be the most impure. Histograms of the triple-point temperatures for the cells labelled -1/X, -2/X, and -3/3 are presented in figure 10. The highest plateau temperature measured for any cell was 58.0806 °C, as indicated in the histograms for cells labelled -1/X and -2/X. Although the lowest temperature indicated in the histograms is 58.0785 °C, five cells with triple-point temperatures below that value were tested but were rejected as being unsuitable to qualify as SRM 1970. Consequently, they are not included in the histograms. The bottom section of figure 10 shows that the values of the 39 cells with designations -1/X have the

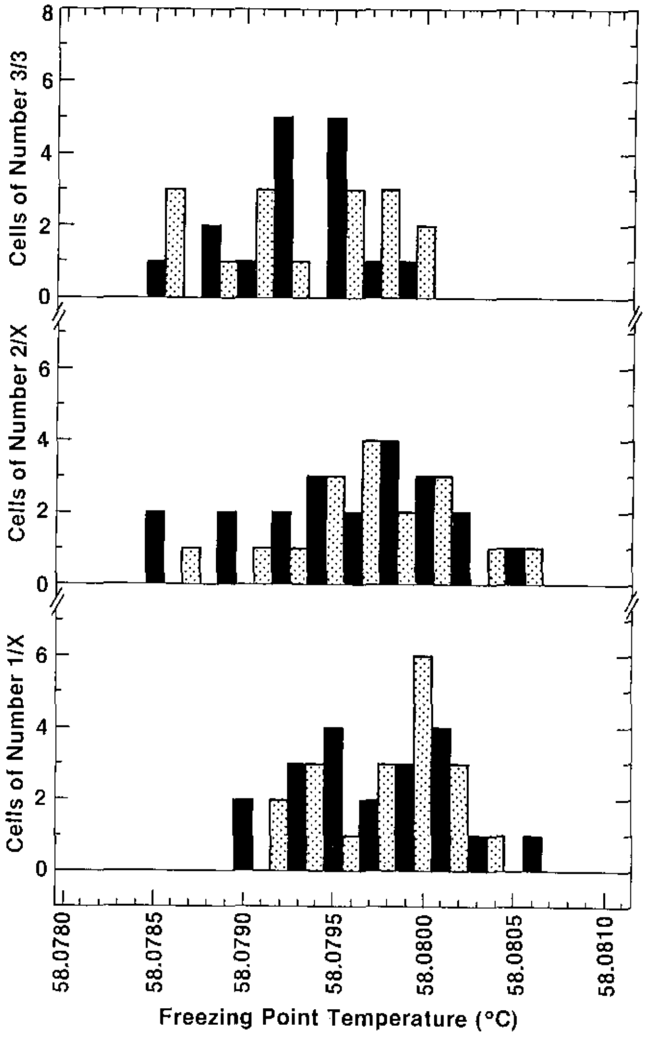

Figure 10–Histograms of the triple-point temperatures obtained by freezing of all of the SRM 1970 cells. The histogram at the bottom of the figure is for cells designated -1/X; that in the middle of the figure is for cells designated -2/X; and that at the top of the figure is for cells designated -3/3.

smallest spread of triple-point temperatures and a mean value (58.0797 °C) which is slightly higher than that of the other groups of cells (those designated -2/X and -3/3). *Also, no cell of the bottom portion of figure 10 has* a triple-point temperature below 58.0790 °C. The middle section of figure 10 shows that the 38 cells with designations -2/X are spread over the entire range from 58.0785 °C to 58.0806 °C, with a mean value of 58.0796 °C, a value only slightly less than that for the group of cells designated -1/X. The top section of figure 10 shows the distribution of the 32 cells designated -3/3. These cells have triple-point temperatures ranging from 58.0785 °C to 58.0800 °C, with a mean value of 58.0793 °C.

In order to demonstrate the reproducibility of the triple-point temperatures observed for the SRM 1970 cells, we present four typical freezing curves in figures 11 and 12. In figure 11, we display two freezing curves of cell D-3/3, obtained on different days in a bath at the same temperature and using the same thermistor thermometer. This shows the excellent reproducibility that is obtainable with most of the cells comprising SRM 1970. In figure 12, the freezing curves characteristic of cell K-2/3 are presented. As shown there, the curves did not reproduce exactly, but the spread was less than 1 mK. The results in figure 12 are typical of those for cells with the greatest irreproducibility. All 109 SCN cells accepted as SRMs were tested for reproducibility in the manner just described. In figure 13, we present a histogram which indicates that each of the cells had freezing curves which repeated to within a value lying between 0.0 mK and 0.8 mK. Figure 13 shows that 77% of the cells, i.e., 84 cells, have a spread of ⩽0.3 mK and only 25 cells had freezing curves with plateaus that differed by amounts ranging from 0.4 to 0.8 mK.

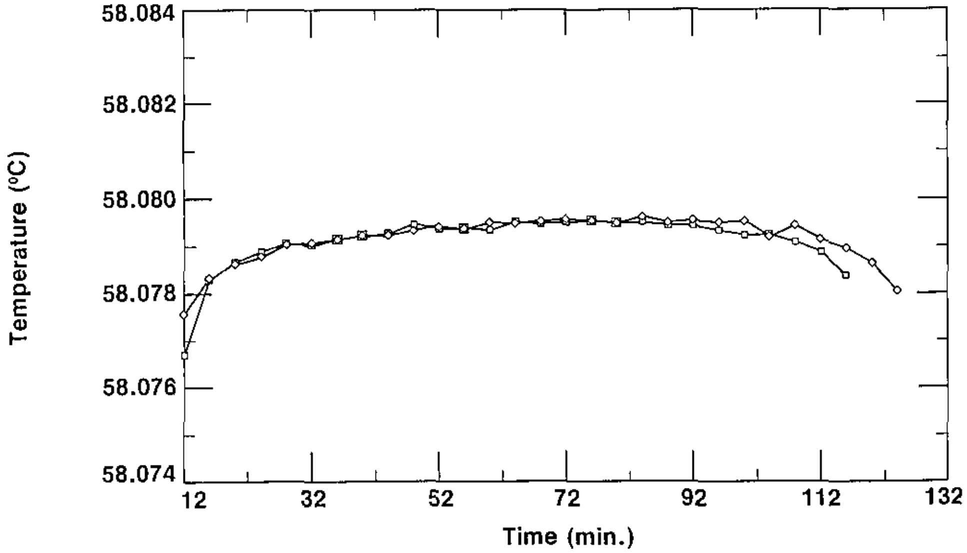

Figure 11–Freezing curves of cell D-3/3 obtained on different days in a bath at 56.880 °C, showing reproducibility.

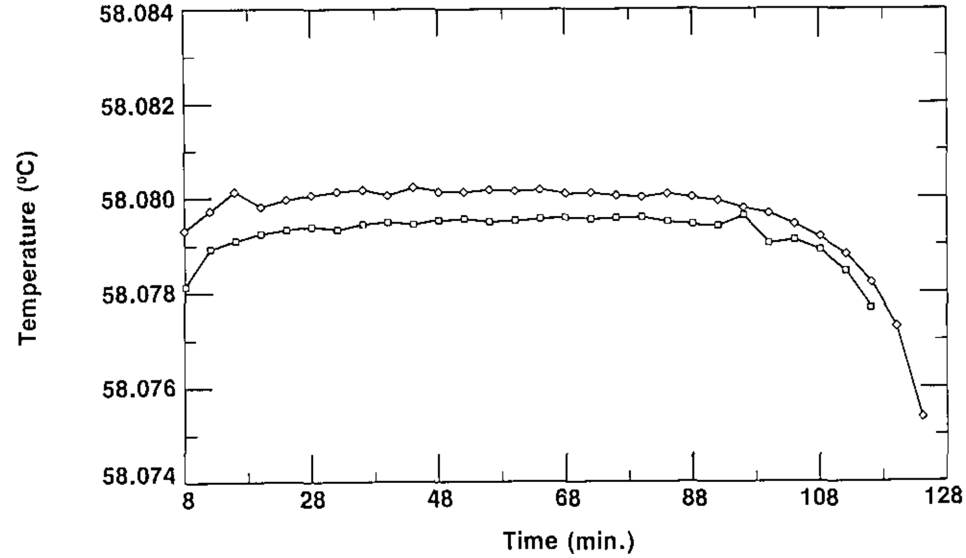

Figure 12–Freezing curves of cell K-2/3 obtained on different days in a bath at 56.880 °C, showing reproducibility.

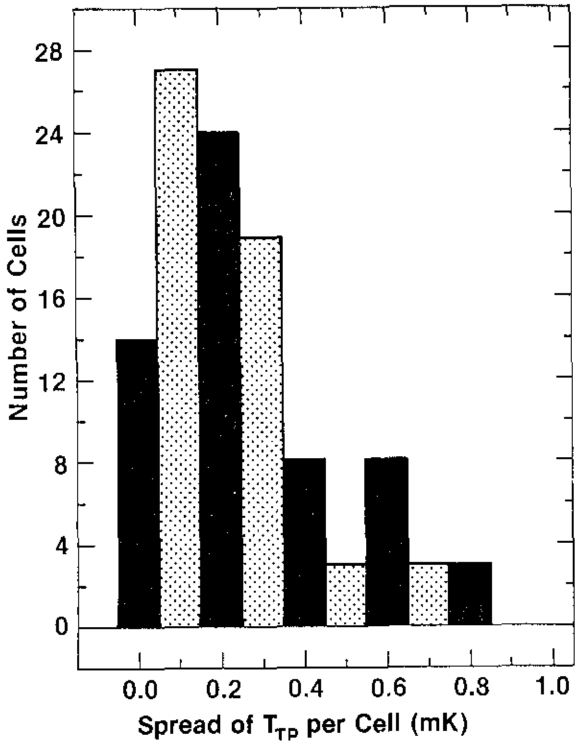

Figure 13–Histogram of the spread of the triple-point temperatures of each of the cells of SRM 1970, as determined from freezing curves.

As indicated previously, we obtained 115 cells of SCN to investigate for suitability as SRM 1970. Of these, 109 were accepted as the standard reference material; five were rejected because the temperatures of the plateaus of their freezing curves had values below 58.0785 °C, and one cell broke. The histogram shown in figure 14 represents the distribution of the triple-point temperatures of 111 of the 115 cells considered in our study. It is shown that the triple-point values of the 111 cells are distributed over a temperature range of 2.6 mK, i.e., from 58.0780 °C to 58.0806 °C, with the two cells having triple-point temperatures below 58.0785 °C being rejected as unsuitable for SRM 1970. All of the five cells rejected for SRM 1970 belonged to the group designated -3/3. The mean triple-point temperature of the 109 cells accepted as SRM 1970 is 58.0796 °C, with a standard deviation of 0.48 mK and a standard error of 0.04 mK.

3.2 Determination of Purity

The reproducibility of the plateau of the freezing curve of a sample and the depression of the triple-point/freezing-point temperature is a function of the purity of that sample [9–11]. Similarly, the purity determines, in

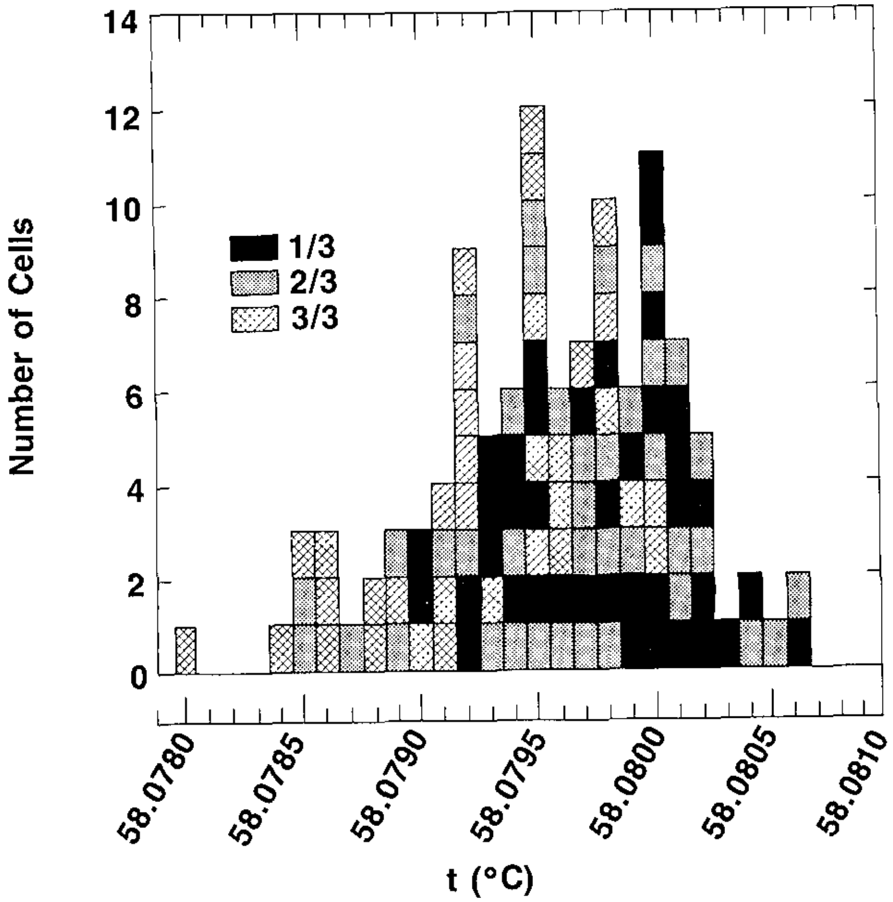

Figure 14–Histogram of the triple-point temperatures of the 109 cells of SCN accepted as SRM 1970, plus those of two cells with temperatures below 58.0785 °C, which, consequently, were rejected as being unsuitable for SRM 1970.

a major way, the melting behavior of a sample [9–11]. Both of these methods, of course, are nondestructive and can be used for any SRM 1970 cell. In using melting curves to estimate the purity of the SCN used as SRM 1970, three cells were selected with triple-point temperatures distributed over the range of temperatures observed for the plateaus of all the cells. One sample chosen was L27-2/3, with a triple-point value of 58.0804 °C; another cell chosen was L20-2/3, with a triple-point temperature of 58.0787 °C; the third cell selected was L26-2/3, which has a triple-point temperature of 58.0795 °C. Prior to the melting experiments from which the purities were to be determined, the cells were prepared as follows. The SCN was completely melted by placing the cells in a water bath at about 80 °C. Then the cells were removed, wrapped with paper towels, placed in dewars, and a slow stream of air directed into their thermometer wells to cause the SCN to freeze slowly from inside outwards. This technique of freezing causes the impurities that have distribution coefficients with values less than 1.0 (impurities rejected on freezing) to be more concentrated on the outermost parts of the cells. After the SCN had solidified and cooled to room temperature, the cells were immersed in a constant temperature bath, maintained at a temperature of 58.38 °C, and their temperatures monitored as the SCN melted. Melting curves of the three cells are presented in figure 15.

Assuming that the law of dilute solutions [12,13] is valid for our samples, the purity can be estimated by using the equation [9]

$$X=\frac{A(T_{f1}-T_{f2})}{[(1/f1)-(1/f2)]} \tag{2}$$

where X is the total mole fraction of impurities in the sample, T_{f1} and T_{f2} are the temperatures when the fractions melted are $f1$ and $f2$, respectively, and A is the cryoscopic constant ($=0.004318$ K^{-1} for SCN). The constant A can be calculated from the equation

$$A=\frac{\Delta H}{RT_0^2} \tag{3}$$

where R is the gas constant (8.314 J mol^{-1} K^{-1}), ΔH is the latent heat of fusion (3703.45 J mol^{-1} for SCN), and T_0 is the melting-point temperature in kelvins. Although the plateaus are not perfectly flat (i.e., there is a finite melting range), the temperature difference between the bath and the SCN is essentially constant, and thus the absorption of heat per unit time by the sample during melting is nearly constant. Consequently, the fraction of SCN melted can be considered to be directly proportional to the time elapsed during the melting and eq (2) can be rewritten as

$$X_2=\frac{A(T_{t1}-T_{t2})}{[(1/t1)-(1/t2)]} \tag{4}$$

where $f1$ and $f2$ have been replaced by $t1$ and $t2$, respectively.

Since the three melting curves shown in figure 15 do not have perfectly flat plateaus and since that for L20-2/3 has an odd shape, we selected various parts of each curve for possible use in estimating the impurity content. The results of the calculations are given in table 1. Values of $f2$ (or $t2$) greater than 0.8 were not used

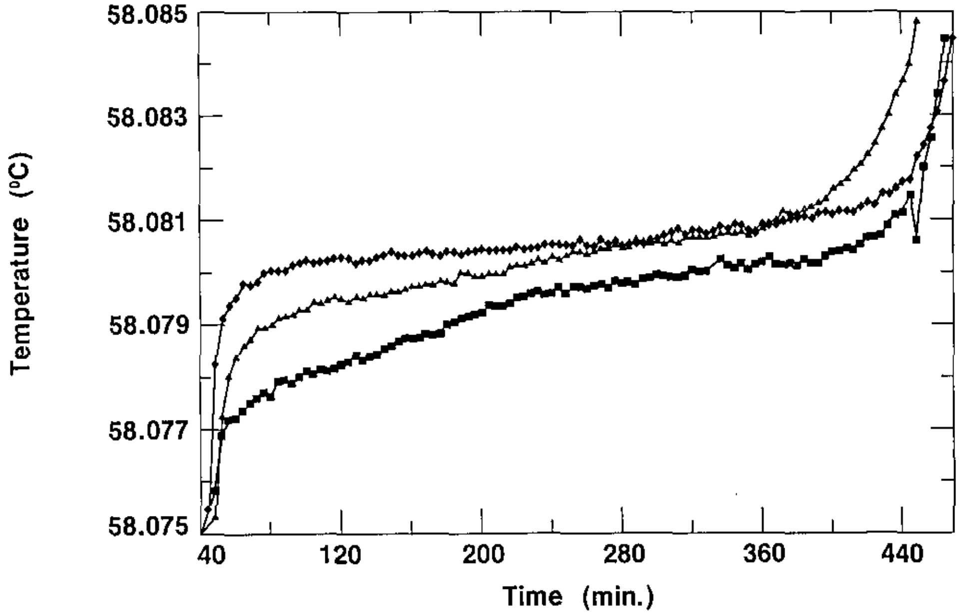

Figure 15–Melting curves of cells L20-2/3, L26-2/3 and L27-2/3 obtained in a bath maintained at 58.380 °C. ■ represent data obtained for cell L20-2/3, ▲ represent data obtained for cell L26-2/3, and ◆ represent data obtained for cell L27-2/3.

Table 1. Mole fraction of impurities for three cells (representative of all 109 cells of SRM 1970 certified to date), determined by using different parts of the melting curves.

$(t1, t2)$	X		
	L27-2/3	**L26-2/3**	**L20-2/3**
(0.1,0.5)	0.000,000,162	0.000,000,593	0.000,001,510
(0.1,0.6)	0.000,000,207	0.000,000,725	0.000,001,553
(0.1,0.7)	0.000,000,302	0.000,000,755	0.000,001,611
(0.1,0.8)	0.000,000,394	0.000,000,888	0.000,001,627
(0.2,0.5)	0.000,000,288	0.000,001,151	0.000,001,870
(0.2,0.6)	0.000,000,388	0.000,001,424	0.000,001,941
(0.2,0.7)	0.000,000,604	0.000,001,450	0.000,002.054
(0.2,0.8)	0.000,000,805	0.000,001,726	0.000,002,071
(0.3,0.7)	0.000,000,906	0.000,002,491	0.000,002,718
(0.3,0.8)	0.000,001,242	0.000,002,900	0.000,002,692
Best estimates:			
of X	0.000,000,**3**	0.000,000,**7**	0.000,00**1,6**
of Purity	99.999,**97**%	99.999,**93**%	99.999,**84**%

because at values greater than that the temperatures of the thermometers were already being influenced by the environment external to the cells. As indicated in table 1, 10 different sections of each curve were used to estimate the amount of impurity. In these determinations of impurities, eq (4) is most sensitive over about the first 50% of the melt. Consequently, we gave that section of the curves the greatest weight in estimating the amount of impurities present. The measurements on the three samples, representative of all the SRM 1970 cells, then, lead to estimates of the purities ranging from 99.999,97% to 99.999,84%, as listed in table 1.

4. Recommended Procedure for Use of SRM 1970 in Calibration of Thermometers

To obtain the best results in calibrating thermometers using SRM 1970, the SCN should be totally melted, a thin sheath of solid SCN prepared around the thermometer well, the cell placed in an appropriate constant-temperature environment, and the SCN slowly frozen while calibrating the thermometers.

The constant-temperature environment should be a well-stirred fluid bath (preferably containing a light purified mineral oil) maintained at a temperature between 56.88±0.05 °C and 58.02±0.05 °C. Since the SRM 1970 cells are approximately 18 cm long, the bath should be at least 20 cm deep from the top level of the oil. The cell or cells should be mounted vertically in a suitable holder in such a way that the cells are completely immersed below the level of the oil.

A temperature controller capable of controlling the bath to ±0.02 °C is preferred, but one could use a bath controlled to only ±0.05 °C, or even ±0.1 °C.

The detailed procedure preferred for calibrating a thermometer through the use of an SRM 1970 SCN triple-point cell is as follows: First, get a well-stirred oil bath to a temperature between 56.88±0.05 °C and 58.02±0.05 °C and maintain it at that point. Melt the sample by immersing the cell in a water bath at about 80 °C. After the SCN has completely melted, remove the cell from the bath and invert it several times to ensure a thorough mixing of the liquid material (SCN and all impurities) inside the cell. Dry the cell, put a small amount of oil in the thermometer well, and insert a cold copper rod into the well. After the formation of what appears to be a 2-to-3 mm sheath of the SCN around the thermometer well, put the cell into the temperature-controlled bath, using a holder to support the cell. Insert the thermometer to be calibrated in the thermometer well of the cell, wait 10 to 15 minutes, then read the indication of the thermometer. The *correct* temperature of the thermometer at that time is the value indicated on the SRM certificate.

If more than one thermometer is to be calibrated, the second and succeeding thermometers should be allowed 5 to 10 minutes to reach thermal equilibrium with the freezing SCN in the cell before obtaining measurements on them. The smaller the mass of the thermometer and the closer the temperature of the thermometer is to that of the fixed point, the shorter the length of time required for equilibrium.

If the minimum recommended temperature of the environment of the cell (56.88 °C) is used, there should be a minimum time of about 90 minutes available for calibration from the time the SCN sheath is prepared. If more than about 75% of the SCN has solidified and if more calibrations are needed, remove the SCN cell, repeat the previous procedure to prepare a new inner sheath, and put the cell back into the oil bath (or other controlled environment) for further calibration work. These consecutive freezing and melting processes may be repeated as often as required.

5. Summary and Conclusions

In certifying cells of SCN for SRM 1970, 115 cells of purified SCN were tested, 5 of them were rejected because their freezing values were below 58.0785 °C, and one was broken. The remaining 109 samples were accepted as suitable for the temperature standard reference material. The triple-point temperatures of all the cells accepted for SRM 1970 ranged between 58.0806 °C and 58.0785 °C, with a mean value of 58.0796 °C, a standard deviation of 0.48 mK and a standard error of 0.04 mK.

In realizing the fixed points of these cells, the freezing technique was found to be more convenient than that of melting because of the greater flatness of the freezing curves. Each of the cells showed a high degree of reproducibility, with the observed temperature values of the plateaus differing by no more than 0.8 mK. These properties make the cells very attractive as a temperature reference standard for the calibration of thermometers near 58.08 °C.

The SCN used was of high purity; that is clear from the fact that the spread of freezing-point temperatures values per lot did not exceed 1.3 mK. The amount of impurities estimated to be present in the SCN of the 109 cells of SRM 1970 has values ranging from 3×10^{-7} to 1.6×10^{-6}, equivalent to a purity of the SCN ranging from 99.999,97% to 99.999,84%. This is a very high purity for an organic material.

Thermometers with diameters no greater than 4.5 mm may be calibrated in these SRM 1970 cells.

6. References

[1] The International Practical Temperature Scale of 1968, Amended Edition of 1975, Metrologia **12**, 7–17 (1976).

[2] Bureau International des Poids et Mesures, Monograph, Supplementary Information for the IPTS-68 and the EPT-76, Pavillon de Breteuil, F-92310 Sevres (1983).

[3] Glicksman, M. E.; P. W. Voorhees and R. Setzko, The triple-point equilibria of succinonitrile—its assessment as a temperature standard, in Temperature: Its Measurement and Control in Science and Industry **5**, American Institute of Physics: New York, NY, 321–326 (1982).

[4] Mangum, B. W., The succinonitrile triple-point standard: a fixed point to improve the accuracy of temperature measurements in the clinical laboratory, Clin. Chem. **29**, 1380–1384 (1983).

[5] Mangum, B. W., Triple point of succinonitrile and its use in the calibration of thermistor thermometers, Rev. Sci. Instrum. **54**, 1687–1692 (1983).

[6] Mangum, B. W., private communication.

[7] Cutkosky, R. D., An ac resistance thermometer bridge, J. Res. Natl. Bur. Stand. (U.S.) **74C**, 15–18 (1970).

[8] Smit, W. M., Errors occurring in the determination of temperature-heat content curves, Anal. Chim. Acta **17**, 23–35 (1957).

[9] Smit, W. M., Melting curves (temperature-heat content curves) as criteria for purity, Recuil Trav. Chim. des Pays-Bas. **75**, 1309–1320 (1956).

[10] Furukawa, G. T.; J. L. Riddle, W. R. Bigge, and E. R. Pfeiffer, Application of some metal SRM's as thermometric fixed points, *Natl. Bur. Stand. (U.S.) Spec. Publ. 260-77, Supt. of* Documents, U.S. Gov't. Printing Office: Washington, DC 20402 (1982).

[11] Riddle, J. L.; G. T. Furukawa and H. H. Plumb, Platinum resistance thermometry, NBS Monograph 126, Supt. of Documents, U.S. Gov't Printing Office: Washington, DC 20402 (1973).

[12] Lewis, G. N., and M. Randall, Thermodynamics and Free Energy of Chemical Substances, McGraw-Hill Book Co.: New York, NY, p. 238 (1923).

[13] White, W. P., Estimating impurities by means of the melting point curve, J. Phys. Chem. **24**, 393–416 (1920).

Journal of Research of the National Bureau of Standards
Volume 90, Number 5, September-October 1985

Performance Assessment of Automatic Speech Recognizers

David S. Pallett
National Bureau of Standards, Gaithersburg, MD 20899

Accepted: September 3, 1985

This paper discusses the factors known to influence the performance of automatic speech recognizers and describes test procedures for characterizing their performance. It is directed toward *all* the stakeholders in the speech community (researchers, vendors and users) consequently, the discussion of test procedures is not directed toward the needs of specific users to demonstrate the performance characteristics of any *one* specific algorithmic approach or particular product. It relies significantly on contributions from an emerging consensus standards activity, especially material developed within the IEEE Working Group on Speech I/O Performance Assessment.

Key words: acoustics; automation; computer technology; signal processing; speech recognition; speech understanding.

1. Summary

This paper identifies and documents factors influencing automated speech recognition performance. Procedures are outlined that are important in designing and implementing performance tests. Documentation is outlined which should clearly define test conditions. Definitions of terms are contained in the Appendix.

Definitive tests to *fully* characterize automatic speech recognizer or system performance cannot be specified at present. However, it is possible to design and conduct performance assessment tests that make use of widely available speech data bases, use test procedures similar to those used by others, and that are well documented. These tests provide valuable benchmark data and informative, though limited, predictive power. By contrast, tests that make use of speech data bases that are not made available to others and for which the test procedures and results are poorly documented provide little objective information on system performance. Such tests might be termed "incomparable" in that the data obtained cannot be meaningfully compared with data for other tests or for other systems.

Speech recognizers are the central element in speech recognition systems, and primary attention in this paper is directed to tests of recognizers as system components. Testing overall systems performance and the human-machine interface involves the more difficult task of developing measures of speech understanding. The factors described in this paper are necessary, but not sufficient concerns in tests of integrated human-machine and speech understanding systems.

A number of recommended testing procedures are described in general terms. These procedures are deliberately not specified in detail because it is recognized that no one detailed procedure could meet the widely-

About the Author: David S. Pallett is a physical scientist in the Systems Components Division of NBS' Institute for Computer Sciences and Technology.

varying needs for test data among researchers, vendors, and users. At present, decisions concerning the best way to implement specific tests and their applicability for research purposes, for commercial products and for proposed applications are best left to the judgment of the researcher, vendor, and user. It is, however, their responsibility to discuss and document the specific test procedures and data supporting their claims for algorithm or product performance. These discussions should be based on the considerations outlined herein.

Since automatic speech recognition is still an emerging technology, a standard terminology has not yet been established. Current activities with the IEEE Acoustics, Speech and Signal Processing Society include a Working Group on Speech I/O Systems Performance Assessment. This Working Group has contributed to the suggested definitions of terminology in this paper.

2. Introduction

Researchers and systems designers in a number of agencies of the Federal Government have worked closely to identify the capabilities of automatic speech recognition technology and to exchange research findings. Applications studied to date include data entry, package sorting, and command/control in aircraft cockpits. These studies have demonstrated the need for careful planning of trial applications and the value of thorough analysis of performance data. Industry is showing ever increasing interest in commercial applications of the technology.

However, until automatic speech recognition technology is a well established element in the human-machine interface, continuing efforts must be made to identify the relative importance of factors influencing performance and to develop and specify definitive test procedures. Tests conducted using these procedures will then serve to clearly demonstrate appropriate uses of the technology and to document the associated productivity benefits.

The paper is the first to report on the development of detailed and specific test procedures for performance assessment in the Institute for Computer Sciences and Technology at the National Bureau of Standards. The overall focus is assessing the performance of speech recognizers as system components, with emphasis on laboratory benchmark tests. The discussion in this first paper is introductory in nature. Continuing attention to these issues, along with the contributions of consensus standards groups, will result in the development of detailed procedures for both benchmark tests of speech recognizers and for measuring human performance.

3. Factors Which Influence Speech Recognizer Performance

Successful implementation of automatic speech recognition technology presents numerous challenges. In many cases these challenges are met through the selection and imposition of constraints on the many factors known to influence performance. Corresponding constraints must be imposed on the structure of performance tests if meaningful performance data are to be obtained. The need for these constraints arises, in large part, from the high inherent variability of unconstrained speech.

The inherent variability of speech arises from the nature of speech and the articulatory process. "Speech is based on a sequence of discrete sound segments that are linked in time. These segments, called phonemes, are assumed to have unique articulatory and acoustic characteristics. When speech sounds are connected to form larger linguistic units, the acoustic characteristics of a given phoneme will change as a function of its immediate phonetic environment because of the interaction among various anatomical structures (such as the tongue, lips, and vocal chords) and their different degrees of sluggishness [1][1]." This variability in the articulatory gestures involved in the production of speech and the interactions that arise from adjoining segments are important factors contributing to the difficulty in successfully implementing automatic recognition of continuous speech.

Humans have well developed abilities to adapt to and accommodate this variability, but at present it is a critical barrier to the automatic recognition of unconstrained speech. Automatic speech recognition systems have difficulty discriminating between linguistically meaningful and insignificant variations. This variability significantly complicates the process of testing.

There are numerous other factors that further complicate the task of successfully implementing and testing automatic speech recognition technology. These factors make it desirable to clearly anticipate the effects they may cause when designing, implementing, and documenting performance assessment tests. Appropriate recognition of these factors will increase the value of the test results as benchmarks for comparative purposes and enhance the predictive power of the tests. The following factors describe the main sources of variability that should be considered when testing automatic speech recognition technology.

---

[1] Figures in brackets indicate literature references.

Speech Related Factors

The form of the speech has a great effect on the difficulty of recognition. Isolated words or discrete utterances are easiest to recognize. Connected words, even if spoken carefully, are more difficult to recognize because the beginnings and ends of each word are affected by the adjacent words. Fluent or continuous speech is much more difficult to recognize becasue the sound segments (particularly those at the beginnings and ends of the words) tend to merge, and stress patterns affect the loudness and distinctiveness of vowels.

Speaker Related Factors

There are important differences in the way different individuals speak. Factors contributing to these differences which may affect performance and which can be readily documented include:

Age: Voice quality in adolescence and old age often differs from that in mid-life.

Sex: Certain speech characteristics, such as pitch and vocal tract length, tend to be gender-specific for adults. Some speech recognition systems employ features that may have been optimized for user groups such as adult males, so that it is important to document age and sex data for the test speaker population.

Dialect History: Some speakers are dialect chameleons, and adapt quickly and convincingly to the dialect characteristics of a new region. Others retain some qualities of previous dialects while changing other qualities. Because pronunciation of many words depends strongly on dialect, documentation of dialect data may be particularly important in tests of speaker-independent recognizers.

Speech Idiosyncrasies: Speech-associated anomalies can be expected to affect recognition performance adversely. Stuttered, lisped or slurred speech patterns and unusual characteristics should be identified and noted. For some individuals, speaker-generated noises such as lip smacks, tongue clicks, "um, ers, and ahs," etc., will degrade performance. Speech levels vary with changes in vocal effort and in the distance between the speaker's mouth and microphone.

Changes in rate of speech introduce additional complications. In words spoken rapidly, some of the sound segments may be shortened or deleted or altered in quality. A slow rate of speech may cause vowels to have drifting frequency spectra and extremely long silence gaps in plosive consonants (e.g., "p," "t," "d," etc.).

When enrolling, testing, and using speech recognizers, the use of chewing gum and smoking should be noted and/or controlled. Care should be taken when selecting a test speaker population so that these characteristics are appropriately represented.

Speech Variability: Individual speakers vary in the degree of consistency with which they repeat words. Some speakers produce nearly identical repetitions of individuals words or utterances, even under stressful conditions. Others produce highly varied repetitions (e.g., words such as "eight" with the final consonant occasionally deleted, or with highly variable pitch). The former category will make a recognizer perform best, and is sometimes referred to as "sheep." The latter is sometimes referred to as "goats."

Motivation and/or Fatigue: Degradation of performance for speaker-dependent systems can be expected as motivation degrades or fatigue increases. It is useful to obtain samples of speech under these conditions in order to estimate the degree of performance degradation.

Task Related Factors

The design of vocabularies for successful application of this technology is an important consideration. Limited size vocabularies require careful planning. Vocabularies should be natural to the task and sufficiently distinct to ensure recognition with few substitution errors.

Performance is greatly improved by the imposition of syntactical constraints. In many task dialogs there are only a few possible choices at each point in the task, thus making the recognition task much simpler, faster, and more reliable.

Physical exertion, fatigue, and other stressing factors must be considered and documented in designing experiments and assessing performance. The voice pitch and loudness or vocal effort of the speaker change due to stress, as do the spectral components.

Environmental Factors

The input speech signal to a speech recognizer is affected by background noise, reverberation, and transmission channel phenomena (e.g., the use of telephone lines or wireless microphones). These environmental factors may lead to spurious responses by the recognizer. The performance of speech recognizers will generally be lower when telephone lines are used for input than with direct microphone input because the frequency response is limited and noise artifacts make correct recognition more difficult. The use of wireless microphones may lead to recognition errors due to transmission channel cross-talk, RF interference, signal fading, dropouts, etc.

Other Factors

Human recognition of speech involves an imperfectly known set of decision criteria. Automatic speech recognition devices apply specific, but (to some degree) arbitrarily chosen, decision criteria in order to effect recognition. Optimum settings of these decision criteria, including the associated reject thresholds, are extremely important. However, the optimum settings of these decision criteria are controlled by the vocabulary, the design of applications software (i.e., the implementation of syntactic constraints, error-correction protocols, etc.), and the characteristics of the individual user's speech and personal preference. Experimentation is required in order to determine the optimum setting of the reject thresholds. As an alternative to the selection and use of optimum settings of the reject thresholds, the reject capability may be disabled, to simulate a forced choice response. This procedure is frequently chosen for benchmark tests.

In some cases, the system may also have the ability to return ordered word lists. Typically, these word lists are ordered according to the distance measure between the input word and the reference templates or word models or in order of descending probability. The application of higher level constraints such as syntax then may lead to correct identification of the utterance. While this process may emulate human decision criteria and typical decision trees, it can complicate assessment.

4. Considerations in Developing Test Procedures

The design and implementation of tests to define the performance of automatic speech recognizers requires that attention be paid to many of the previously described factors influencing performance. A systematic process of experimental design and testing is indicated in this section to account for these factors. This process includes:

- Selecting an experimental design that either
  - (a) models an application, or
  - (b) provides benchmark data.
- Selecting speakers to represent the user population or some relevant subset.
- Selecting a test vocabulary that either
  - (a) exemplifies that used in an application, or
  - (b) has been used by others for benchmark test purposes.
- Training the system, or constructing the reference patterns to be used by speaker-dependent recognizers.
- Characterizing the test environment in order to document complicating factors such as factory noise, communications channel limitations, or task-related factors.
- Recording the test material to permit verification of the validity of the test results and reuse of the test material.
- Scoring the test results. Procedures are outlined for both isolated and connected word data.
- Pragmatic considerations to ensure that equipment is properly operating, that tests are conducted in a manner that is consistent with manufacturer's recommendations, and other related factors.
- Statistical considerations to indicate the statistical validity of performance data.
- Documentation of test conditions and performance data to allow evaluation of published data.

Tests designed and carried out accounting for these factors will be valuable in identifying the strengths and weaknesses of automatic speech recognition systems. The importance of performance assessment procedures has been emphasized in a recent study by the Committee on Computerized Speech Recognition Technologies of the National Research Council. Their report [2] recommends that: "... performance should be measured within a realistic task scenario, both within the laboratory and in actual operational settings, including worst case conditions. Laboratory benchmark tests using standard vocabularies, experienced users, and controlled environments are useful for comparing recognizers, but they are not efficient for predicting actual performance in operational systems. Adequate methods are needed for measuring both human and recognizer performance under realistic conditions. The importance of performance measurement techniques cannot be over emphasized since they provide the data for decisions about system design and effectiveness ...."

Experimental Design

There are two complementary approaches to designing performance assessment tests. These approaches are summarized in table 1.

In one approach, a set of benchmark test conditions is defined (e.g., use of a "standard" speech vocabulary and data base, and no use of syntax to actively control the recognition vocabulary). Little or no effort is taken to model an application. This approach provides valuable comparative performance information. It does not directly predict performance in real applications.

Table 1. Alternative approaches to test design.

Test Conditions	*Benchmark Tests*	*Applications Tests*
Vocabulary	Benchmark or Reference Vocabulary	Applications Specific (Task) Vocabulary
Data Base	Widely Available Recorded Data Base	(Variable)
Use of Syntax	Little or No Use of Syntax	Syntactically Constrained Word Sequence or Imposed Task Grammar
User Interaction	None	(Variable)
Predictive Power	Very Limited	Less Limited
Data Analysis	Detailed	(Variable)
Documentation	Thorough	(Variable)

A second approach consists of carefully selecting test conditions in order to simulate a field application. The use of syntactically constrained word sequences may dramatically enhance performance and is acceptable for user applications. The design of a test vocabulary should include specifying the structure of the grammar and the frequency of occurrence of each item. This approach may have greater predictive power in inferring performance in specific applications, but it complicates comparisons between differing applications. Because of the diverse applications proposed for recognizers, simulation of many different applications and the needs of differing users becomes very difficult and/or costly.

In both approaches to testing, simple averages such as error rates, recognition accuracy, etc., are often inadequate to indicate performance. It is important to determine and document the most frequently occurring confusion pairs (e.g., "five-nine" confusions, where a spoken five is recognized incorrectly as "nine"). Presentation of this data in the form of a confusion matrix is very informative. For an N-word vocabulary, the confusion matrix is an N-by-N-matrix of input versus output, a form of stimulus-response matrix representation. Correct recognition responses fall along the diagonal of this matrix, and substitution responses comprise the off-diagonal elements.

Selecting the Test Speaker Population

In both benchmark tests and in applications tests, care should be taken to select speakers for the tests that are in some sense representative of the ultimate users of the technology. For example, in applications tests of industrial quality control data entry systems, the most valuable test speakers will ordinarily be quality control personnel. In research or benchmark tests, the test speakers are ordinary adult males and/or females with "neutral" dialects. In extraordinary circumstances, efforts are taken to obtain representative speakers with regional dialects. However, representative sampling of all potential users is not always possible or necessary. The characteristics of the test speakers and both their user training and system enrollment procedures should be documented. While all of the documented factors may not significantly affect performance, the documentation will indicate to others whether the test group is of particular interest or relevance.

Some recognizers impose limitations on the duration of words, or of silence gaps within words considered as single words or strings. Other constraints may apply to the number of words which may constitute a connected string. These constraints may have important consequences in some applications and for some individual speakers (e.g., if the durations of the speaker's stop gaps are longer than a limit set by the manufacturer, the word or phrase may be segmented into two utterances, and will not be correctly recognized).

Because speech recognizers use enrollment data to build reference template sets, prototypes, or other internal representations of the words to be recognized, it is important that the enrollment data for speaker-dependent systems provide representative samples of the user's speech. The enrollment and test data should include speech that is characteristic of the application, possibly including fatigued or stressed speech. These requirements may complicate enrollment and test procedures and, when slighted, generally result in lower performance in an application. Other important factors include the degree of cooperation of the users and their familiarity with the equipment.

Automatic speech recognition algorithms and commercial systems perform best on systems trained for the intended user's voice. Such speaker dependent recognizers provide some degree of language independence, depending on the type of acoustic-phonetic representation or pattern matching algorithm used by the device. They may perform equally well when used with several languages. However, speaker independent recognizers are expected to be language and dialect dependent to the extent that they rely on phonological rules and specific data bases for the development of internal representations. The issue of language or dialect independence may be very important for some applications.

Speaker independent systems do not rely on the data obtained from the individual user's voice. Rather, they are designed using training or enrollment data from many speakers and incorporate representations (e.g., template sets derived by studying clusters of individual speakers' templates or word models derived from statistical analysis of many individual speakers' word models) based on features which are presumed not to vary from individual to individual. This is a crucial assumption (that the system relies on features that are relatively consistent) and its successful implementation is the key to success in speaker independent automatic speech recognition technology. It is essential to ensure that the most important variabilities and dialect related factors have been accounted for when designing and testing such systems. These requirements become increasingly challenging if large vocabularies are required and response must be available ina time period comparable with the duration of an utterance (i.e., real-time recognition).

When selecting test speakers for "speaker independent" systems there are a number of special concerns. Perhaps most importantly, a representative sampling of the intended user population should be obtained in order to appropriately represent regional dialect and/or transmission channel effects for the intended user and applications population. A statement describing the efforts taken to represent the user population should be included as part of the documentation. When conducting tests of these systems, it is important to exclude data from the test material that might have been used in constructing internal representations used by the recognizer. Casual recognition experiments using template sets generated from one person or a small number of people typically demonstrate highly variable performance. Sometimes recognition performance may be quite good or quite poor for some individuals, and frequently there will be good performance on some words and poor performance on others. For these reasons, casual experimentation to demonstrate "speaker independence" for systems designed to be speaker dependent is not recommended.

Selecting the Test Vocabulary

The actual performance of any given speech recognition system in both benchmark tests and applications is critically dependent upon the vocabulary items that must be distinguished at any given time. Both the number of items to be distinguished and the acoustic similarity or complexity of these items are critical factors.

Brief monosyllabic words (e.g., yes, no, go, the natural alphabet except for "w" etc.) are more difficult to recognize than longer polysyllabic words or brief phrases spoken and intended to be recognized as single items (e.g., Massachusetts, California, "start printing," "left bracket"). These more complex utterances contain much more acoustic information and redundancy than monosyllables. In actual applications, this fact is used to construct vocabularies that retain many of the qualities of a natural interaction while selecting somewhat more complex acoustical characteristics to maximize system performance.

For these reasons, it is necessary to explicitly state the test vocabulary. It is, of course, desirable to use a test vocabulary that is identical to the intended vocabulary for the application.

One parameter often used to characterize recognition system performance is that of the vocabulary size. Vocabulary sizes, ranging from approximately 40 to several hundred words, are not unusal at present. However, in order to enhance performance, it is often appropriate to use syntactic constraints. This is implemented through the imposition of an artificial language grammar to constrain the vocabulary choices at each stage of a task in a given application. In many cases, this is not only appropriate but will lead to significantly enhanced productivity by imposing a desired order for completing the intended task.

Restricted vocabularies and formatted messages are widely used for speech communications in situations such as air traffic control and military tasks in which high speech comprehension is required. Acceptance of these constraints in isolated and connected word speech recognition applications will result in higher performance, but must be explicitly stated when documenting system performance.

It is important to distinguish between the total vocabulary capacity (typically a function of total memory available to the system) and other measures of the effective *vocabulary size (typically functions of the structure* of the imposed artificial language grammar). For artificially constrained tasks, the average number of alternative words that the system has to choose from at any time is given by the "perplexity," or dynamic branching factor for the imposed artificial language.

A 10-word recognizer, requiring discrimination between the digits (0–9) with all transitions equally likely is typically more difficult than a system with a several hundred word total vocabulary and branching factor of only 5. Even if the larger vocabulary has a branching factor of 10, the larger vocabulary may be easier than the 10-word digit vocabulary if the vocabulary words tend to be longer and more discriminable than the digits (eight of which are monosyllables).

The benefits achieved through the use of syntactic constraints may be better addressed in separately documented tests. In syntactically constrained tasks, performance results ought to be reported with the following information to describe the characteristics of the imposed grammar.

(a) Complete description of the task grammar including full specification of the vocabulary at each task state or menu choice.
(b) Frequencies of transitions from each task state to successive states.
(c) Dynamic branching factor or perplexity.
(d) Frequency of occurrence in the test material of each vocabulary item.

Training

Two meanings of the word "training" are sometimes found in the literature of current speech recognition technology. A clear distinction must be made between them.

In one meaning, the user's speech is used to "train" the recognizer for the specific test or applications vocabulary. During this process, reference patterns ("template sets," "voice patterns," "voice prints," etc.) or more complex word models are developed and become the stored internal representations used for comparison with subsequently input speech in the recognition process. This process is referred to as "enrollment" without ambiguity.

A second meaning of the term "training" refers to that process in which the user of a recognizer becomes familiar with the device or system. During this "user training," many factors may combine to influence the user's speech. Generally, familiarization with the devices leads to improved performance, and the user learns to adapt to explicit, as well as implicit constraints on the form of the input speech.

One factor in user training that tends to improve performance uses feedback provided to the user. To date, most recorded speech data base material has not been obtained under circumstances allowing user feedback. The recorded speech data base material has been obtained in response to prompts or in list-reading tasks. The nature of the feedback provided to the test speaker should be documented along with a description of any prompts provided to the user or the tasks conducted by the user while providing test material.

In tests conducted on integrated systems (as opposed to tests on system components), time must be allowed for familiarization with the system and to observe the nature of performance improvement or degradation. In most cases, after a period of initial user training, performance can be improved significantly by simply re-enrolling the user. The new internal representations should then be more representative of the experienced user's typical speech, and poor initial performance due to the lack of user familiarity will be improved. Documented performance ought to represent the data obtained with experienced or fully trained users.

Characterizing the Environment

Both the operational environment and the speech signal transmission system providing input to speech recognition systems are important environmental factors influencing performance. For example, in an industrial quality control voice data entry application, the talker's environment might be a noisy factory floor, while the speech signal transmission environment may be a wireless microphone. When modeling an application, the acoustic environment and signal transmission channel should closely simulate the intended operational environment.

When access to the actual intended operational environment is limited or costly (e.g., in tests of systems for use in operational aircraft), using accurate simulations can provide a cost-effective test environment. By accurately modelling the environment, the value of such tests is enhanced by increasing the correlation between the test data obtained in the simulation and the actual operational environment.

Because laboratory test data are often not applicable to the user's operational environment, the responsibility for tests in operational environments becomes a critical element in dialogues between vendors and users.

When an actual operational environment is used, as for laboratory tests, care must be taken to control and document all potentially relevant characteristics of the test environment. Environmental noise tends to interfere with communication between humans. It also tends to degrade speech recognition system performance and it is best to separately conduct certain benchmark tests in which all background and transmission noise is minimized. These tests tend to provide information on optimum system performance because acoustic-phonetic information is not obscured by the noise. Comparison of benchmark test data with operational data can indicate the existence of noise-related limitations on performance for which noise control measures or improved transmission channels can lead to improved performance.

There are at least three types of noise that can affect performance:

-Ambient or background noise. This noise originates with the operation of nearby machinery such as office equipment, with ventilation systems, with people conversing in the vicinity of the user, or within the applications environment such as the crewspace of an aircraft. When microphones are located some distance from the talker's mouth, reflections from nearby surfaces such as desk-tops and room walls constitute a form of multipath interference that can be comparable to increased ambient noise in degrading system performance.

- Transmission or channel noise. Such noise is inherent in using wireless microphones or telephone lines, and (for long distance lines, in particular), may be due to signal processing devices such as echo suppression, multiplex, or satellite transmission systems.
- Inadvertent test speaker noises. These may originate in coughs, stammers, "ers," "ums," excessive breath noise, and speech extraneous to the selected recognition vocabulary.

Characterization of noise is important in interpreting operational test results. Attention should be directed to performance limitations that may be due to the following factors:

1) *The noise experienced by the speaker.* Speakers may modify their speech significantly in the presence of high noise. Typical modifications include speaking more loudly or slowly and taking care to articulate more carefully than otherwise. Minimal characterization provides the A-weighted sound level (dBA) experienced by the user. Because masks, helmets, and some headsets affect the perceived noise level, their use by the test speaker should be noted. More detailed characterization should include spectral content (e.g., third-octave band analyses) and the temporal nature (e.g., steady-state, intermittent or impulsive). Impulsive noise can lead to substantial degradations in performance, but full characterization of this noise is difficult to achieve without sophisticated instrumentation.
2) *The speech signal-to-noise input to the recognizer (prior to any recognition system signal processing).* The use of a noise-cancelling microphone can effectively eliminate much of the noise environment of the test speaker, even in a high noise environment. However, the signal-to-noise properties of the signal input to the recognizer may be a critical factor in limiting performance in noisy environments. The relative importance of the differing characteristics of speech emitted in a noisy environment vs. the degraded signal-to-noise properties is not yet well understood. Different algorithms and/or devices are probably affected to differing degrees.
3) *Type and characteristics of the microphone.* Useful characteristics to note include close-talking or noise-cancelling, directionality, whether push-to talk or otherwise manually switched, distance from the speaker's mouth, etc. The effectiveness and frequency response of noise cancelling microphones are influenced by the distance to the sound sources. Thus, placement of the microphones should be documented.
4) *Verbal characterization and description of the origin of the noise.* Typical characterizations use terms such as "buzzy," "hum," "static," etc., and descriptions of the origin are "office environment," "package sorting machinery," "receiving platform," etc.

If a speech signal transmission system other than direct microphone input is used, attention should be directed to these additional factors:

5) *Limitations on the transmission channel bandwidth and frequency response.* Cite the upper and lower cut-off frequencies and any significant deviations from flat frequency response over the cited bandwidth.
6) *Other limitations on the transmission channel.* Significant performance limitations may be due to other effects such as automatic volume control attack and release characteristics, signal compression and/or limiting, phase distortion, additive noise in transmission, etc. In general, these effects are difficult to characterize.

While it is possible and, in many cases, desirable to record speakers in sound isolated (low ambient noise) and anechoic (dead) environments in order to access subtle details of the speech signal, care must be taken in generalizing these observations to infer the nature of speech in the presence of noise. A first-order procedure involves the addition of white, pink, or carefully shaped noise spectra to the speech data after they are collected. Factors associated with the acoustic environment that influence the speakers include both ambient noise level and the degree of reverberation.

Speakers may compensate for these factors by speaking more loudly or enunciating more clearly or slowly. It should be specified what ambient noise was audible to the speaker at the time of collection.

In addition to acoustic environmental influences on the speaker, the task environment modifies the speaker's performance. Different types of tasks will affect the speaker's speech to varying degrees. Routine speech tasks such as list-reading or responding to visual prompts displayed on terminals produce less word-to-word variations than speech produced when there are concurrent physical tasks or shifts in attentional focus usually associated with other cognitive activities such as inspection, measurement, etc. For these reasons, higher recognition system performance can be expected from speech obtained from list reading than when there is concurrent tasking, and the talker's task environment should be fully described in the results.

Recording the Test Material

Many performance assessment tests make use of recorded speech data bases. Others are conducted "live." The general practice of recording the test material, even for "live" testing, is recommended. The recorded material provides a means of replicating the results obtained and verifying that the test material was properly input to the system. It also provides material to be used in additional measurements on similar systems as well as for analysis of the input audio signal. Using recorded test material offer the advantage of providing samples of speech obtained on different occasions, separated by days or weeks. This can account for some of the day-to-day variation and may more accurately model potential applications.

Recorded speech data bases exist in both digitally recorded and stored formats and analog recorded formats. The digital formats have greater signal-to-noise ratio than the analog format. Reference data bases that are widely used in research and testing have been recorded with 16-bit samples at sample rates from 10.0 to 20.0 Khz. Signal-to-noise ratios in excess of 90 dB are feasible using this technology. A widely used format for analog recordings is the use of quarter-inch magnetic tape at 7.5 inches per second, providing maximum signal-to-noise ratios of the order of 60 dB. The use of cassette tape recorders is not generally recommended for benchmark test purposes for a number of reasons including increased print-through.

Another advantage of using digital storage for data base material is that each speech token may readily be assigned an accompanying "header" to indicate the origin of that particular token. Comparable systems are feasible using analog storage, storing the header information in an encoded analog signal on the second channel of a two-channel tape recorder, but these systems may require specialized interfaces.

Newly developed recording technology includes use of 14 or 16 bit digital sampling and pulse code modulation systems to encode the signals for storage on Beta or VHS format video recorders, referred to as PCM/VCR recording technology. This technology offers many of the advantages of digital sampling and storage at lower costs for the storage medium than more traditional digital storage media, and offers the capability of copying the data with less degradation than for analog recordings.

A number of speech data bases have been widely used in testing and serve to provide test material for benchmarks. They are available in several recorded formats [3].

Scoring Isolated Word Data

Scoring isolated word recognition systems performance presents fewer challenges than for connected word systems. The relative ease in scoring isolated word data arises from the fact that most errors tend to be substitutions: deletions or insertions are easily identified when they do occur.

It is ordinarily presumed that the individual speech tokens (e.g., words or short phrases with minimal intra-word pauses) are separated in time by pauses that are long enough to permit the recognition system to respond. Indications that this may not be the case will be found if there is a high incidence of deletions or substitutions, and it should be noted that the origin of these errors may be due to a problem with the system's response time for the data base used for these tests.

Prior to detailed data analysis, it is instructive to critically listen to the recorded test material, particularly for those portions of the test material where unusual numbers of errors may have occurred. These errors may be due to noise artifacts or departures from proper script-reading or responses to prompts. If this is the case, the recorded tokens or artifacts must be editorially deleted from the test material prior to testing. Objective analysis of the data must include full documentation of these decisions regarding certification of the test material. It is preferable not to delete any data if the process of obtaining and using the test material was carefully structured and monitored.

Preliminary analysis of the performance data should identify and tabulate words which were correctly recognized, words provided as input for which substitution errors occurred, words provided as input for which there was no response (deletion errors or rejections), and instances in which a response occurred without a corresponding appropriate input (insertion errors). The raw data should be summarized by determining the corresponding correct recognition percent as well as the substitution, deletion, and insertion error percent.

In comparative testing of differing systems, it is generally preferable to disable the reject capability, so that each system returns a forced choice response. In this case, words provided as input for which there is no response are unambiguously classified as leading to deletion errors.

In performing benchmark tests to compare different recognizers, the removal of all syntax constraints may be preferable. These constraints, like the setting of reject threshold, may affect systems differently. If, following data analysis, recognition errors are concentrated on several specific words or utterances, then re-enrolling the speaker on these words or substituting acoustically distinctive synonymous words may substantially improve performance.

In other tests, particularly those at the integrated system level or in modelling an application, the use of the reject capability is an important feature that should be included in the test program. Tests in this case should document the settings of the reject threshold and/or other decision criteria, identify and tabulate words input to the system for which the reject response occurred, and determine the rejection percent. It is also valuable to determine and document the ratio of total errors to rejections, because this information may be useful in the design of applications software.

More detailed analysis of systems performance can be documented and easily reviewed by constructing a confusion matrix. Analysis of these data will provide valuable insights into systems performance and the design of successful vocabularies.

Another useful measure may be appropriate for those systems that present several ranked words for approval. In this case, recognition accuracy as a function of the word rank is a useful parameter, since it provides a measure of the probability that the second, third, ... Nth candidate is correct if the higher ranked candidate is incorrect. If it is known *a priori* that a recognizer will be implemented in an application that will impose higher level constraints, such as a syntactically controlled (sub-) vocabulary, then it is appropriate to determine and report the probabilities that the correct word is to be found among the top N candidates on the ordered list. This practice is inappropriate if the imposition of higher level constraints is impractical in a typical application.

It is sometimes desirable to have measures of a system's capacity to reject words that are not in its recognition vocabulary. This is particularly appropriate for those applications involving inexperienced users or those unaccustomed to using artificial grammars or syntactic constraints. For experienced users, it may be safe to assume that the input is limited to words in the recognition vocabulary, in which case out-of-vocabulary rejection capabilities are less critical.

In order to test a system's capability to reject out-of-vocabulary utterances a secondary test can be performed using the same recognition data base used for other tests. In this case, however, a subset of the recognition vocabulary is selected and the system is re-enrolled using only this subset of the entire test vocabulary. The entire data base is then used for test purposes, with responses that occur for words that are not part of the active vocabulary (the selected subset) being classified as "false acceptances." Documentation in such a test must include the total test vocabulary and the specified active vocabulary.

Scoring Connected Word Data

There are several ways to score recognition performance on connected word strings. The most stringent method is to record the percentage of strings completely recognized, i.e., the number of correct strings divided by the total number of strings tested. Another method is to calculate the percentage of individual words correctly recognized. The number of substitutions, deletions and insertions is calculated for each string, and the total error count is divided by the total number of words in the strings tested. The string scoring may be done by procedures involving strict left-to-right alignment or by a best-case pattern match.

Left-to-right alignment procedures involve matching each word in the input string with a corresponding word in the response string starting with the first (leftmost) member of each string. Obviously, the occurrence of an insertion or deletion will shift the position of words in the response string so that succeeding responses will be compared with inappropriate members of the string. For example, if the input string is 12345, and a deletion error results in a response string 1345, a left-to-right alignment procedure correctly scores the first digit as a correct recognition, but would cite the three following responses as substitution errors (e.g., "3" for "2," "4" for "3," "5" for "4") and would detect the presence of a deletion error only at the last digit (e.g., no response for "5"). In this case, one correct recognition, three substitutions, and one deletion would be indicated where in fact there were four correct recognitions, and one deletion (and the deletion error is, in fact, mis-identified). This left-to-right pattern match procedure, though well defined and easy to implement, is in many cases a very poor or worst-case pattern match.

When using best-case pattern match procedures, individual words are matched so as to minimize the number of errors within the string. That is, if the input string is 12345, and the output string 1345, it is inferred that there were four correct recognitions and one deletion.

Selection of the most appropriate scoring method involves consideration of the relevant application, and particularly the manner of verification and correction by the speaker. Where the manner of correction involves repetition of an entire string, the string error rate may be most appropriate. The aligned word recognition scores would be appropriate measures for those cases in which correction may be possible by backing up one word at a time for the end of the string.

Using best-case pattern match procedures tends to avoid some of these complications, but there are no generally agreed-upon procedures to uniquely define the best-case match criteria. Specific details of these procedures are beyond the scope of this paper. Purchasers of systems for which these considerations are significant should discuss the scoring procedures used by vendors.

Pragmatic Considerations

Prior to conducting tests, care should be taken to make sure that the equipment is functioning properly. Because recognition systems are designed to perform with distorted and/or variable input, determining proper functioning is not a simple task. Malfunctioning components can masquerade as an imperceptible input distortion, and the system will appear to work but not as well as it should. Check procedures should include tests to confirm consistency of recognition results with results obtained previously using recorded speech, checks of input amplitude settings, and the use of any available software diagnostics. These tests should be routinely conducted at the time of testing.

In tests of commercially available systems, the manufacturer's recommendations should be followed in order to obtain optimum performance. If the manufacturer's recommendations are not followed, some degradation in performance may be expected.

Manufacturers may suggest procedures to use regarding:

-Recommended number of enrollment tokens.
-Presentation order of enrollment tokens.
-Required minimal interval pause duration.
-Amplitude and gain control settings (to accurately simulate live input if recorded input is used). Amplitude settings should not be readjusted, once set.
-Microphone position.

Specify if a "press to talk" switch is used. The use of "press to talk" microphones provides an input signal to the recognizer that has very little or no signal amplitude between words. Depending on the particular recognizer's procedure for accommodating input signals during inter-word pauses, this may lead to either improved or degraded performance, relative to the use of conventional unswitched microphones. Choice of "open" or "press-to-talk" microphones should be determined by operational considerations (e.g., if it is an accepted practice in a proposed application or if it is required to activate a remote system) as well as whether the use of one or the other may lead to optimum performance with a given recognizer.

Proper connection of peripheral devices and electronic components should be checked before testing to eliminate ground loops and extraneous noise. The speech (audio) signal input to the recognizer should be monitored by the experimenter to verify that no extraneous noise is being introduced.

Statistical Considerations

Performance assessment tests of any automatic speech recognition system require that a large number of speech tokens be input to the system by many users and that detailed analysis of the test data be conducted. Thorough testing requires the use of large test speech data bases, substantial data storage, and time. Disregard for these facts inevitably leads to misleading conclusions regarding system performance.

Researchers, vendors, users and developers of automatic speech recognition technology each have different needs for performance data. The significance and interpretation of the data vary because of the different goals each group seeks to achieve. Consequently the degree of concern for the statistical validity of the performance data is variable.

Factors that need to be considered in structuring statistically valid performance tests include the size of the test speaker (user) group, the number of test tokens, and the amount of enrollment material provided to the system.

For benchmark testing, a concise statement of the number of test speakers, number of test utterances, and number of errors of each type should be given. It is recommended that the performance documentation should include a statement of the total error rate and the confidence level implied by each statistic. Statistical tables should be consulted to interpret the results, and the assumptions made in computing the statistics should be stated explicitly [4].

For tests that model an application, statistically based considerations include sampling the intended user population and range of tasks to be implemented using speech recognition and defining the variability in the noise environment.

For speaker independent recognition technology, particular attention need be paid to sampling the intended user population and communications channels. Dialect-related effects and variations in the quality of telephone connections make it difficult to obtain consistent performance from current low-cost remote access speaker independent recognition technology. Testing of this technology must be based on large speech data bases.

To obtain optimal performance from each of the systems to be compared in a benchmark test the appropriate vendor-recommended training procedure must be followed. Because some recognizers make use of single-token enrollment, while others build increasingly more reliable statistically based word models in the process of enrollment (and, possibly, in operating in a speaker-adaptive mode), appropriate enrollment procedures often vary significantly from one system to another.

No generally accepted rules have yet been developed for statistically reliable speech recognition system test procedures. In view of the many factors influencing performance, most researchers and vendors attempt to carefully control known sources of variable performance. Those researchers and vendors whose products build increasing accuracy with statistically large enrollment data are particularly conscious of the need for statistically large enrollment and test data bases. Though no generally accepted rules for adequate statistical sampling presently exist, data analysis should seek to define the distribution of performance data, as well as mean values. Decisions regarding apparent superiorities of algorithms or products cannot be reliably made if the differences in mean values are smaller than the associated variances. Reference to handbooks of experimental statistics can be valuable in avoiding misinterpretation of the test data.

In principle, extensive and statistically valid testing involves the use of large data bases. However, the costs of testing and resources required for these tests are frequently regarded as prohibitive, and more limited testing is typical. Consequently, attempts should be made to determine the statistical validity of the tests as an important factor in performance assessment.

Documentation

Proper documentation is an essential component in performance assessment. As recommended in this paper, information should be provided to document the relevant characteristics of the test speaker population, test vocabulary and other test data to establish the relevant context of the testing.

Test material obtained in accurate simulations of field applications may contain noise or speaker artifacts such as coughs, stammers, or false starts. Alternatively these artifacts may have been manually deleted or edited from the test and/or training material. The process of selection or preparation of the test material should be described.

Summary test data that should be documented include those in table 2.

These results are the most frequently cited performance data, however, documentation of confusion pairs (e.g., "five" and "nine") or confusion matrices is informative and provides useful information in designing applications vocabulary.

Response time is a critical factor in successful implementations of large vocabulary systems. There is no

Table 2. Data requiring documentation.

Correct Recognition Percent (Recognition Accuracy) $= \frac{(\textit{\#Correctly Recognized Words}) \times 100}{(\text{\#Test Words})}$ (Percent)

Substitution Percent $= \frac{(\textit{\#Substituted Words}) \times 100}{(\text{\#Test Words})}$ (Percent)

Deletion Percent $= \frac{(\textit{\#Deleted Words}) \times 100}{(\text{\#Test Words})}$ (Percent)

Insertion Percent $= \frac{(\textit{\#Inserted Words}) \times 100}{(\text{\#Test Words})}$ (Percent)

If the reject capability is employed, the following data are important:

Rejection Percent $= \frac{(\textit{\#Rejection Responses}) \times 100}{(\text{\#Test Words})}$ (Percent)

Ratio of Total Errors to Rejections $= \frac{(\textit{\#Substitutions} + \textit{\#Deletions} + \textit{\#Insertions})}{(\text{\#Rejections})}$

Settings of the Reject Threshold or Reject Criteria:

accepted procedure for precise measurement and specification of response time for connected word systems. If the processing time is comparable to or less than the utterance duration, the system response time may be described as "real time." A suggested comparative measure for other than real time systems is a multiple of utterance duration, assuming processing is initiated at the beginning of the utterance and completed with display or return of the recognized word.

In view of the fact that processing times are finite, errors which arise from utterances spoken with insufficient pauses between words—(for isolated word systems, in particular) should be identified and noted in the performance documentation.

The A-weighted sound level (dBA) measured in the vicinity of the test speaker should be specified if environmental noise is believed to be a significant limitation on performance. More thorough documentation of the properties of the environmental noise may be appropriate.

Suggested documentation of the speech signal-to-noise properties of the test material should include the ratio of speech peak level to steady background noise level (in dB) measured according to ANSI S.3.59 [5]. If this is not feasible, at least, the range of typical maximum speech level to background level indication on the VU indicator of a conventional audio tape recorder should be noted and cited.

5. Perspectives on Testing

As explained in the preceding material, there is no simple and completely objective way to test the performance of automatic speech recognition technology. The number and complexity of factors influencing performance is such that in many cases, the relative advantages offered by competing algorithms, commercial products, or integrated systems may be obscured. The approach toward performance assessment in this paper emphasizes the value of benchmark tests and the need to carefully model applications. This is particularly important if the expenses of integrating speech technology into a particular, well-defined application and the benefits to be achieved are appreciable. Attention to detail in planning an applications test should be reflected in greater confidence in the ability of the technology to provide the anticipated benefits. Here also, a poorly-structured applications test or one that does not adequately account for important factors influencing performance will invalidate the test results and may lead to costly and unsuccessful attempts to use this new technology.

There are however, a number of other perspectives toward testing. It is useful to identify some of these perspectives.

Informal Device Testing

Within the past several years, the emergence of low cost commercial products has made an informal approach toward performance testing fairly widespread. Many individual purchasers of speech recognizers undertake an informal test program that primarily consists of familiarization of that purchaser or a designated individual with the technology. In many cases, the primary value of these tests appear to be that the experimenter learns to recognize the constraints imposed by the particular product on system enrollment, environmental factors, user interaction, interface design, etc. If informal testing of this sort results in development of a successful application, the experimenter gains valuable experience in the use of a new technology, insights into the selection of improved second-generation or competitive products and the design of more formal and reliable tests. However, there are real risks that the informal testing may not lead to the development of a successful application, that the purchaser may inappropriately conclude that the technology offers no promise for his application, and that the individuals involved in the testing and systems integration learn little of value in the process. Serious attention should be given to allocating adequate resources to carry out more formal tests to the point at which a serious and detailed investigation has taken place, and at which time the experimenter can demonstrate that he or she has developed an in-depth understanding of the relevant strengths and weaknesses. The information contained in this paper should be valuable in understanding the scope of the issues that should be addressed in these tests.

Workstation or Task Redesign

Another perspective toward studying the performance of this technology is based upon the desire to achieve the productivity benefits that might be offered by redesign of workstations or tasks, by using speech as an alternative or additional data entry or command/control modality. This is perhaps the implicit goal of all efforts to create successful applications. The design of tests to measure productivity benefits is beyond the scope of this report, but extremely important. In tests to measure these productivity benefits, those benefits which are specifically due to the use of speech technology should be compared to those that might be primarily due to redesign of workstations or tasks.

Human Factors Research

The design of successful human-machine dialogs is an active and important research topic at present. Not enough is known at present about the desired properties of automatic speech recognizers and the human-machine interaction to always lead to the design of successful applications. Further research on such issues as the design of optimal error correction protocols and user training and feedback is required and will serve to advance the technology. Attention to the factors contained in this report should serve to increase the value of these studies.

Common Concerns

The differing perspectives are consistent with the different outlooks regarding the purpose of the tests. The testing, like the technology itself, may serve different purposes. Whatever approach is taken, however, there are many common concerns including:

- Recognized complicating factors must be accounted for and carefully controlled. Failure to do so will invalidate the test data.
- Detailed documentation must be made available to indicate experimental design and to provide data and sufficiently detailed data analysis to indicate the significance of the test results to others. Detailed documentation should be part of reporting all performance tests. Failure to do so often leads to meaningless comparisons of product or system performance or misleading citations.
- Benchmark test data, in which severe constraints have been imposed on the test conditions, are extremely valuable. Typical constraints limit the test vocabulary, number of test talkers, format of the input speech, nature of environmental effects, and prohibit feedback between the user and the system. While it can be argued that under these conditions, it is possible to adapt or specially "tune" algorithm or device characteristics to optimize test performance for a specified test data base, benchmark tests provide data that are useful for initial comparisons of algorithm or device performance. Application testing should reveal whether or not the particular device selected for this phase of testing is well suited to a particular application or user's needs.

Agreement should be reached at an early stage of interest in the technology on the purpose of testing and appropriate measures of performance. Once these issues are decided, the nature of the tests can be detemined. In the absence of agreement on these issues, little progress can be made.

An *ad hoc* group met at the National Bureau of Standards in June 1982, to discuss performance assessment of speech recognition systems, following discussions during the NBS/NADC sponsored Workshop on Standardization for Speech I/O Technology. Further discussions were held at the 1982 (Paris), 1983 (Boston) and the 1984 (San Diego) meetings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Following the 1982 ICASSP, the *ad hoc* group was constituted as the Speech I/O Technology Performance Evaluation Working Group, sponsored by the Speech Processing Technical Committee of the IEEE Acoustics, Speech and Signal Processing Society. Material presented in Section IV of this paper is adapted from informal drafts circulated within this Working Group.

Particular appreciation is expressed to Dr. Janet M. Baker, Chairman of the IEEE Speech I/O Technology Performance Evaluation Working Group, for her enthusiastic support and constructive criticism of this material, as well as to many other individuals who have shared their perspectives and expertise in addressing these issues.

6. References

[1] Reddy, R., and V. Zue, "Recognizing Continuous Speech Remains an Elusive Goal," *IEEE Spectrum*, November 1983.

[2] "Automatic Speech Recognition in Severe Environments," Report of the Committee on Computerized peech Recognition Technologies, National Research Council, National Academy Press (Washington, DC), 1984.

[3] Baker, J. M.; D. S. Pallett and J. S. Bridle, Speech Recognition Performance Assessments and Available Data Bases, *Proceedings of ICASSP83*, Boston, MA, April 14–16, 1983, p. 527–530.

[4] See for example, Burrington, R. S., and D. C. May, Handbook of Statistics with Tables (Second Edition) McGraw-Hill, New York, 1970, Sections 14.53–14.64.

[5] American National Standard Method of Measurement of Speaking Levels, ANSI S3.59 (pending).

7. Annotated Bibliography

The papers cited in this annotated bibliography include many of the key papers used for background material in preparing this paper. While there is very limited literature on the topic of performance assessment *per se*, there are valuable discussions of the several issues affecting performance in many of these papers. The interested reader is encouraged to refer to these key papers and the related literature cited therein in order to develop a more thorough understanding of the process of performance assessment.

Automatic-Speech and Speaker Recognition, N. R. Dixon and T. B. Martin, eds., IEEE Press (New York), 1979, p. 33. An important collection of papers that documents the state-of-the-art just prior to widespread use of VLSI technology in automatic speech recognition.

Baker, J. M., "The Performaing Arts—How to Measure Up!" Proceedings of the Workshop on Standardization for Speech I/O Technology, D. S. Pallett, ed., National Bureau of Standards, Gaithersburg, MD, March 18–19, 1982, p. 27–33. Presents the results of a series of tests on isolated word, speaker dependent recognizers, and a discussion of statistical factors in testing.

Baker, J. M., D. S. Pallett, and J. S. Bridle, "Speech Recognition Performance Assessment and Available Data Bases," Proceedings of ICASSP 83, Boston, MA, April 14–16, 1983, p. 527–530. Discusses data bases used in some well-documented benchmark tests.

Chollett, G. F. and C. Gagnoulet, "On the Evaluation of Speech Recognizers and Data Bases Using a Reference System," Proceedings of ICASSP 82, Paris, May 3–5, 1982, p. 2026–2029. Proposes the use of an openly published reference recognition algorithm to provide benchmarks of recognizer performance as well as of the relative difficulty of different speech vocabularies and data bases. Chollett credits R. K. Moore with early advocacy of this approach.

Chollett, G. F., and M. Rossi, "Evaluating the Performance of Speech Recognizers and Data Bases at the Acoustic-Phonetic Level," Proceedings of ICASSP 81, Atlanta, GA, March 30–April 1, 1981, p. 758–761. Advocates comparison of measured confusibility for recognizers with that of humans.

Clark, J., P. Collins, and B. Lowerre, "A formalization of performance specifications for discrete utterance recognition systems," Proceedings of ICASSP 81, Atlanta, GA, March 30–April 1, 1981, p. 753–757. Discusses needs for objective procedures for comparing the performance of isolated word recognizers.

DeMori, R., *Computer Models of Speech Using Fuzzy Algorithms*, Plenum Press (New York), 1983, p. 382–386. Proposes an experimental model for speech recognition and understanding using fuzzy-set theory and concepts of artificial intelligence. Includes a review of several procedures for evaluating language complexity.

Doddington, G. R., and T. B. Schalk, "Speech Recognition: Turning Theory into Practice," IEEE Spectrum, Sept. 1981., p. 26. Describes the results of a series of benchmark tests on isolated word, speaker dependent recognizers, and presents a thoughtful discussion of factors complicating the performance assessment test procedure.

Goodman, R. G., Analysis of Language for Man-Machine Communication, Ph.D. Dissertation, Stanford University, May 1976. An early study of the complexity of imposed grammars and the issues of phonetic ambiguity and measures of the similarity of acoustic events.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K., "Perplexity—a measure of the difficulty of speech recognition tasks," J. Acoust. Soc. Am. Vol. 62, Suppl. No. 1, Fall 1977, p. S63. The abstract of a paper given at the Fall 1977 meeting of the Acoustical Society of America. This paper presented material summarized by Jelinek, Mercer, Bahl, below.

Jelinek, F., Mercer, R. L., Bahl, L. R., "Continuous Speech Recognition: Statistical Methods," Chapter 25 in Handbook of Statistics, Vol 2, P. R., Krishnaiah and L. N. Kanal, eds., North-Holland Publishing Co. (1982) 549–573. These papers describes methods of assessing the performance of continuous speech recognition systems and the relative difficulty of recognition tasks.

Lea, W. A., "What Causes Speech Recognizers to Make Mistakes?", Proceeding of ICASSP 82, Paris, May 3–5, 1982, p. 2030–2033. Discusses factors that affect recognition accuracy and cites other papers by this author on related topics.

Leonard, R. G., "A Database for Speaker-Independent Digit Recognition," Proceedings of ICASSP 84, San Diego, CA, March 19–21, 1984, p. 42.11.1–42.11.4. Describes valuable data base for research and test purposes that represents more than 20 dialects using approximately 100 adult males, 100 adult females and 100 children. Includes a discussion of an experimental procedure to "certify" the data base using human subjects.

Moore, R. K., "Evaluating Speech Recognizers," IEEE Transactions ASSP Vol 25, No. 2, 1977. A significant early contribution proposing a standard for comparing the performance of different recognizers based on a model of human word recognition.

Peckhan, J. B., "Speech Technology Assessment Activities in the U. K.," Proceedings of Speech Tech '85, New York, NY, April 22–24, 1985, pp. 165–169. Describes the activities and goals of the Speech Technology Assessment Group formed in the U. K. within the Speech Group of the Institute for Acoustics.

Poock, G. K., Martin, B. J. and Roland, E. F., The Effect of Feedback to Users of Voice Recognition Equipment, Naval Postgraduate School Report NPS55-83-003, February 1983. One of a series of studies conducted by Poock and his colleagues addressed at measurement of the performance of integrated or interactive systems.

Proceedings of the Workshop on Standardization for Speech I/O Technbology, D. S. Pallett, ed., National Bureau of Standards, Gaithersburg, MD, March 18–19, 1982. Presents approximately thirty technical papers addressing the issue of performance assessment for speech I/O technology.

Proceeding of the Voice Data Entry Systems Applications Conference '84, L. Lerman, ed., American Voice Input/Output Society, Arlington, VA, Sept. 11–13, 1984. Most recent in a series of four Conferences addressing the design of applications of voice-interactive systems for data entry and other purposes.

Rosenberg, A. E., " A probabilistic model for the performance of word recognizers," AT&T Technical Journal, Vol. 63, No. 1, Jan. 1984, pp. 1–32. Develops analytic models of recognizer performance.

Sondhi, M. M., and Levinson, S. E., "Relative difficulty and robustness of speech recognition tasks that use grammatical constraints," J. Acoust. Soc. Am., Vol. 62, Suppl. No. 1, Fall 1977, p. S64. The abstract of a paper given at the Fall 1977 meeting of the Acoustical Society of America. This paper presented material summarized by Sondhi and Levinson, below.

Sondhi, M. M., and Levinson, S. E., "Computing Relative Redundancy to Measure Grammatical Constraint in Speech Recognition Tasks," Proceeding of ICASSP 78, Tulsa, OK, May 1978, pp. 409–412. This paper develops algorithms for computing various statistical properties of finite languages and documents these properties for eight speech recognition task languages that have appeared in the literature.

Spine, T. M., Williges, B. H. and Maynard, J. F., "An Economical Approach to Modeling Speech Recognition Accuracy," Int. J. Man-Machine Studies Vol 21, 1984, pp. 191–202. Discusses the use of a central-composite design methodology as a means to develop empirical prediction equations for speech recognizer performance. Factors manipulated in the experimental design include number of training passes, reject threshold, difference score and size of the active vocabulary.

Wilpon, J. G. and Rabiner, L. R., "On the Recognition of Isolated Digits from a Large Telephone Customer Population," Bell System Technical Journal, Vol. 62, No. 7, September 1983, pp. 1977-2000. Describes the collection and use of an isolated digit data base for use in research and development of speaker independent speech recognition systems.

Wilpon, J.G., "A Study on the Ability to Automatically Recognize Telephone Quality Speech from Large Customer Populations," AT&T Technical Journal, Vol. 64, No. 2, February 1985, pp. 423-451. Provides an important and detailed analysis of procedures used in collecting speech data bases over telephone lines. Together with the preceding paper by Wilpon and Rabiner, valuable information is presented on the complexity of successful implementation of speaker-independent automatic speech recognition over telephone lines.

Appendix: Terminology

Since automatic speech recognition is an emerging technology, a standard terminology has not yet been established. Current activities within the IEEE Acoustics, Speech and Signal Processing Society include a Working Group on Speech I/O Systems Performance Assessment. This Working Group has discussed the desirability of use of a uniform terminology in technical papers, presentations, and vendor's specifications, and have contributed to the suggested definitions of terminology contained in this Appendix and used in this paper.

Active Vocabulary—See "Vocabulary"

Adaptation—The automatic modification of existing internal machine representations (e.g., template sets, word models, etc.) of specific utterances and/or noise.

Artificial Language—See "Constrained Language"

Automatic Speech Recognition—The process or technology which accepts speech as input and determines what was spoken.

Automatic Speech Recognition System—An implementation of algorithms accepting speech as input and determining what was spoken.

Automatic Speech Recognizer—A device implementing algorithms for accepting speech as input, determining what was spoken, and providing potentially useful output depending on word(s) recognized.

Connected Words—Words spoken carefully, but with no explicit pauses between them.

Constrained Language—Lexically and syntactically constrained word sequences (e.g., telephone numbers).

Continuous Speech—Words spoken fluently and rapidly as in conversational speech.

Deletion—An instance in which a spoken word is ignored, and for which the recognizer or system provides no response (e.g., in recognizing a string of digits, if the recognizer returns one less digit than has been input).

Discrete Utterance Recognition—The process of recognizing a word or several words spoken as a single entry.

Enrollment—The process of constructing representations of speech, such as template sets or word models, to be used by a recognizer. Also referred to as "system training," as distinct from "user training."

Enrollment Data—See "Training Data"

False Acceptance—An example of failure to reject properly spoken input utterances that are not part of the active vocabulary, resulting in selection of a word in the active vocabulary.

Grammar—In general, a grammar of a language is a scheme for specifying the sentences allowed in the language, indicating the rules for combining words into phrases and clauses. In automatic speech recognition, task grammars specify the active vocabularies and the transition rules that define the sets of valid statements to complete the tasks. The task grammar and structured vocabulary provide syntactic control of the speech recognition process that can greatly enhance performance.

Insertion—An instance of a recognition occurring due to spurious noise or an utterance other than those that are legitimate on syntactic considerations. In the former case, some input other than an utterance (typically some ambient or electrical noise artifact) is not properly rejected and the system response indicates that some utterance in the recognition vocabulary occurred. In the latter case, a word that has been uttered (but which is not part of the active recognition vocabulary because of current syntactic constraints) is falsely accepted as an utterance from the active recognition vocabulary.

Isolated Words—Words spoken with pauses (typically with duration in excess of 200 ms) before and after each words.

Isolated Word—See "Discrete Utterance Recognition"

Natural Language—Syntactically unconstrained word sequences, typically drawn from a large lexicon and complying with conventional usage.

Practice Data—Any speech material (utterances) used in developing a recognition system prior to a test of that particular recognizer.

Recognition Systems—See "Automatic Speech Recognition Systems"

Recognition Unit—The basic unit of speech on which recognitions being performed, often presumed to be the word. The actual unit used may be smaller (e.g., phones, demisyllables, syllables or features) or larger (e.g., multi-word phrases or utterances).

Recognition Vocabulary—See "Vocabulary"

Recognizer—See "Automatic Speech Recognizer"

Rejection—The property of rejecting inputs. There are three general classes of system response involving rejection: i) noise rejection, ii) rejection of improperly spoken input utterances, iii) rejection of properly spoken input utterances that are part of the active vocabulary, sometimes termed false rejection.

Speaker Dependent Recognition—A procedure for speech recognition which depends on enrollment data from the individual speaker who is to use the device.

Speaker Independent Recognition—A procedure for speech recognition which requires no previous enrollment data from the individual speaker who is to use the device.

Speech Level—A logarithmically based measure of the amplitude of a speech waveform. Accurate specification of speech level is important in specifying the input signal amplitude when testing recognizers and when specifying signal-to-noise ratio. American National Standard ANSI S3.59 provides a well-specified procedure for measurement of speech level.

String—A sequence of spoken words or phrases, often spoken as connected words or continuous speech and intended to provide a single useful input to a recognizer (e.g., a five-digit ZIP Code or a seven-digit telephone number).

Substitution—An instance in which one word in the recognition vocabulary is incorrectly recognized as another word in the recognition vocabulary.

Syntax—Structure by which grammatical word sequences are specified.

Test Data—Any speech material (utterances) used in a particular test of a recognizer not previously used in developing or modifying that recognizer. The same set of test data may be used repeatedly for tests of different recognizers or in production testing, but not for continuing tests of an algorithm or recognizer in development.

Token—A sample speech utterance.

Training—See "Enrollment." "System training" is preferably referred to as "enrollment." "User training" refers to the process of user familiarization with speech technology (e.g., learning how to use an automatic speech recognition device).

Training Data—Speech material used to construct parametric representations of speech such as template sets or word models used by a recognizer. Also referred to as enrollment data. Not to be confused with performance data obtained in training potential users of the technology.

Utterance—A word or multi-word phrase spoken continuously as a single unit.

Vocabulary—The words or phrases to be recognized by a recognizer. Distinctions should be made between the complete set of all words or phrases that a recognizer has been trained or programmed to recognize, sometimes called the total *recognition vocabulary*, and the (instantaneously varying) subset of these that may be active at a given time because of an imposed task grammar or other syntactic constraint, called the *active vocabulary*.

Word—See "Recognition Unit"

Word Model—A parametric (coded) representation of the sound patterns of words as a sequence of units such as phonetic units, syllables, or other speech parameters.

conference report

CHEMICAL KINETICS—THEORY AND EXPERIMENT

The International Conference on Chemical Kinetics held June 17–19, 1985 at the National Bureau of Standards in Gaithersburg, MD, was attended by about 200 scientists from academic, industrial, and government laboratories. The meeting was sponsored by the Air Force Office of Scientific Research, the Environmental Protection Agency, the Gas Research Institute, the National Aeronautics and Space Administration, the NBS Office of Standard Reference Data and the Center for Chemical Physics, and the National Science Foundation. The purpose of the conference was to bring together investigators from a broad range of institutions and backgrounds to review progress and problems in theoretical and experimental chemical kinetics.

The conference in many ways was a continuation of the highly successful symposium held at NBS seven years ago under the chairmanship of Prof. Frederick Kaufman, entitled "Current Status of the Kinetics of Elementary Gas Reactions: Predictive Power of Theory and Accuracy of Measurement." Its scope was considerably broader, however, in that in addition to gas phase phenomena, the program included sections on ionic reactions and reactions in the condensed phase. This is indicative of the exciting developments in chemical kinetics in recent years. The exploitation of new experimental techniques and a deeper theoretical understanding have led to a confluence of the many individual elements in formerly diverse areas of chemical kinetics so that it is now possible to think in terms of the commonality of problems.

The meeting, which was chaired by John T. Herron, consisted of six half-day sessions. The first was held in honor of Prof. Sidney W. Benson for his many outstanding contributions to chemical kinetics. It involved a number of technical contributions by his colleagues. These included H. E. O'Neal, Robin Walsh, A. J. Colussi, S. E. Stein, and K. D. King, along with a presentation by Prof. Benson himself. The session was chaired by R. Srinivasan and was introduced by D. Golden.

Subsequent sessions covered Fundamentals, chaired by R. M. Noyes; Unimolecular Reactions, chaired by M. Bowers; Energy Transfer, chaired by S. H. Bauer; Reactions in the Liquid Phase, chaired by C. Walling; and Bimolecular Reactions, chaired by M. Pilling. Each session began with a series of 30-minute invited talks. This was followed by a 75-minute coffee break where the posters of papers pertaining to that session could be examined, and then an hour discussion session led by the chairman and the invited speakers. The invited speakers and their topics were as follows:

H. F. Schaefer III, "The $F+H_2$ Potential Energy Surface. The Ecstasy and the Agony"; J. Troe, "Elementary Reactions in Compressed Gases and Liquids: From Collisional Energy Transfer to Diffusion Control"; E. M. Arnett, "Comparison of Thermodynamics and Kinetics for Some Important Bond-Making and -Breaking Processes in Organic Chemistry"; W. L. Hase, "Unimolecular and Intramolecular Dynamics. Relationship to Potential Energy Properties"; Thomas Baer, "The Dissociation Dynamics of Energy Selected Ions"; J. J. Gajewski, "Empirical Approach to Substituent Rate Effects in [3,3]–Sigmatropic Shifts Utilizing the Thermochemistry of Coupled Nonconcerted Alternative Paths."; H. M. Frey, "Energy Transfer"; J. R. Barker, "Large Molecule Energy Transfer: New Techniques and New Controversies"; J. C. Stephenson, "Vibrational Relaxation of Chemical Bonds in Liquids and on Surfaces"; J. I. Brauman, "Prediction of Rate Constants for Ionic Reactions in the Gas Phase and in Solution"; "J. A. Howard, "Measurement of Absolute Propagation and Termination Rate Constants for Alkylperoxyls in Solution by the Hydroperoxide Method"; J. H. Espenson, "Kinetics and Mechanisms of Inorganic Reactions in Aqueous Solutions: Intermediates"; T. H. Dunning, Jr., "Theoretical Characterization and Calculation of Potential Energy Surfaces for Chemical Reactions"; and J. Wolfrum, "Laser Stimulation and Observation of Bimolecular Reactions."

Prof. Frederick Kaufman was scheduled to give a talk entitled "Kinetics of Bimolecular Processes" but was unable to attend the meeting because of an illness that was to prove fatal. His loss removes not only one of the world's leading figures in gas kinetics but a man deeply admired for his wit, dedication, and humanity.

The titles of the talks give a clear indication of the scope of the conference. The hour–long discussion periods concluding each session were especially valuable. Under the direction of the chairman, and with the invited lecturers providing guidance, the discussion provided an in-depth look at the current state of each area.

A special feature of this meeting was the bringing together of kineticists from many disparate areas of chemistry.

In all these areas, modern techniques allow us to focus on the fundamental interactions. At this level, the commonality of the effects are clear and we can all learn a great deal from each other. The new methodologies and broadened theoretical understanding have led to two distinct directions in experimental work: the study of truly fundamental processes in the form of state-to-state chemistry, and the study of thermal reactions involving molecules of increasing complexity.

To be truly meaningful, all such studies must be firmly wedded to theory. In the case of state-to-state studies, it appears that the experiments are considerably ahead of theory. It is thus extremely difficult to generalize about the many interesting observations. In the area of thermal reactions, semi-empirical approaches based on transition state theory, and within the framework of thermochemical kinetics, have proved to be an extremely useful concept in explaining observation and clearly will have an increasing predictive capability. Interestingly, it is in this area that theory may play an ever more important role. Thus, while we may continue to have problems using theory to generate accurate potential energy surfaces or to describe motion on these surfaces, the description of a molecule at the bottom of its well in terms of its energetics and energy, at least at a semi-empirical level, is approaching an accuracy that rivals direct measurement, at least for smaller molecules. Since this is a large fraction of the information that is needed for the application of transition state theory, one can foresee many exciting developments in the quantitative aspects of the chemical kinetics of thermal reactions.

Prepared by John T. Herron and Wing Tsang of the Chemical Kinetics Division, National Bureau of Standards, Gaithersburg, MD 20899.