DETC2008-49669

A COMPREHENSIVE METRIC FOR COMPARING TIME HISTORIES IN VALIDATION OF SIMULATION MODELS WITH EMPHASIS ON VEHICLE SAFETY APPLICATIONS

H. Sarin, M. Kokkolaras,* **G. Hulbert, P. Papalambros**
{harshit,mk,hulbert,pyp}@umich.edu

Department of Mechanical Engineering
The University of Michigan, Ann Arbor

S. Barbat, R.-J. Yang
{sbarbat,ryang}@ford.com

Passive Safety, Research and Advanced Engineering
Ford Motor Company, Dearborn, MI

ABSTRACT

Computer modeling and simulation are the cornerstones of product design and development in the automotive industry. Computer-aided engineering tools have improved to the extent that virtual testing may lead to significant reduction in prototype building and testing of vehicle designs. In order to make this a reality, we need to assess our confidence in the predictive capabilities of simulation models. As a first step in this direction, this paper deals with developing a metric to compare time histories that are outputs of simulation models to time histories from experimental tests with emphasis on vehicle safety applications. We focus on quantifying discrepancy between time histories as the latter constitute the predominant form of responses of interest in vehicle safety considerations. First we evaluate popular measures used to quantify discrepancy between time histories in fields such as statistics, computational mechanics, signal processing, and data mining. Then we propose a structured combination of some of these measures and define a comprehensive metric that encapsulates the important aspects of time history comparison. The new metric classifies error components associated with three physically meaningful characteristics (phase, magnitude and topology), and utilizes norms, cross-correlation measures and algorithms such as dynamic time warping to quantify discrepancies. Two case studies demonstrate that the proposed metric seems to be more consistent than existing metrics. It is also shown how the metric can be used in conjunction with ratings from subject matter experts to build regression-based val-idation models.

1 Introduction

Vehicle safety has become a major concern in modern society. Automotive manufacturers have to meet several regulations and mandatory Federal Motor Vehicle Safety Standards (FMVSS). Additionally, consumer information programs such as the New Car Assessment Program (NCAP) and the Insurance Institute of Highway Safety (IIHS) impose further requirements for vehicle safety. Currently, testing whether these requirements are satisfied is conducted through numerous, costly and time-consuming physical experiments.

Computer modeling and simulation-based methods for virtual vehicle safety analysis and design verification could make this process more cost-efficient. Moreover, virtual testing (VT) can improve real-world vehicle safety beyond regulatory requirements since computer predictions can be used to extend the range of protection to real-world crash conditions at speeds and configurations not addressed by current regulations.

To achieve the promises of VT, computer predictions need verification and validation (V&V), so that the designs obtained using simulation models can be cleared for production with minimized prototype testing. The AIAA guide for verification and validation of computational fluid dynamics simulations defines verification and validation as follows [1].
Verification is the process of determining that a model implementation accurately represents the developer's conceptual descrip-

---

*Corresponding author, Phone/Fax: (734) 615-8991/647-8403

1

tion of the model and the solution to the model.
Validation is the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.

The American Society of Mechanical Engineers Standards Committee on verification and validation in computational solid mechanics describes model validation as a two step process [2]:
1. Quantitatively comparing the computational and experimental results for the response of interest.
2. Determining whether there is acceptable agreement between the model and the experiment for the intended use of the model.
Moreover, Oberkampf and Barone proposed in [3] six properties that a validation metric should satisfy. These six properties form a generic guideline and act as a set of requirements for the development of a new validation metric.

A comprehensive metric for measuring the discrepancy between simulation model responses represented by time histories is necessary to accomplish the first step of the validation process as defined above. In this paper, we review existing metrics and discuss their advantages and limitations. We then propose a new metric that is based on components associated with three physically meaningful error characteristics: phase, magnitude and topology. The proposed metric utilizes measures such as cross-correlation and L_1 norm and algorithms such as dynamic time warping to quantify the discrepancy between time histories. We use two vehicle safety case studies to demonstrate that the proposed metric seems to be more consistent than existing metrics. We then show how the metric can be used to build regression-based validation models in cases where subject matter expert data are available.

2 Review of selected metrics and algorithms

In this section we review popular metrics and algorithms used currently to quantify discrepancies between time histories in various fields such as voice, signature or pattern recognition, computational mechanics, data mining and operations research. We review these metrics with respect to their advantages and disadvantages in order to propose a new comprehensive metric that utilizes strengths and avoids weaknesses of existing tools and techniques. We provide references only for the less commonly used metrics.

To illustrate some limitations of the reviewed metrics, we consider an example with three time histories that shall be referred to as "test 1", "test 2" and "test 3." Time histories test 2 and test 3 are compared to test 1 to determine which one has the smallest discrepancy and is thus the best prediction of test 1 (Figure 1).
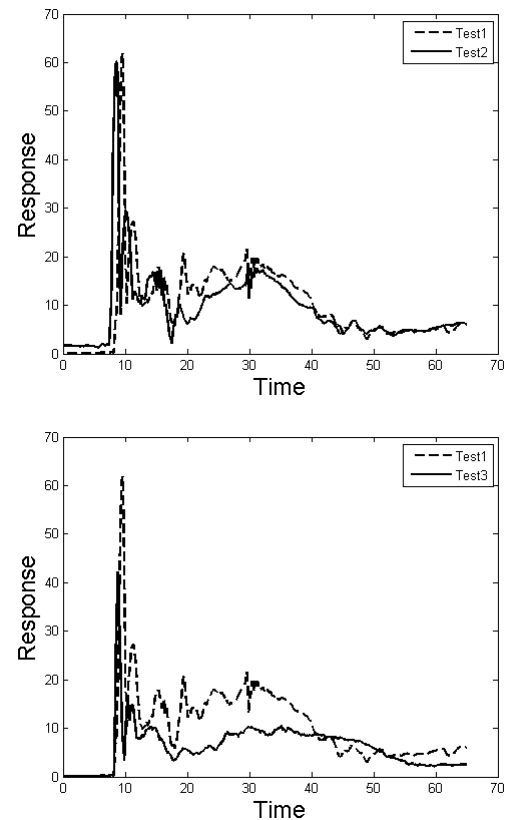

Figure 1.   Time history examples

2.1 Vector norms

When time histories are discretized (i.e., finite-dimensional), the most popular measure for quantifying their difference is to use vector norms. Assuming two time history vectors A and B of equal size N, the L_p norm of the difference of the two is

$$\|A - B\|_p = \left(\sum_{i=1}^{N} |a_i - b_i|^p \right)^{\frac{1}{p}}. \tag{1}$$

The three most popular norms are L_1, L_2 (Euclidean) and L_∞. The results obtained when using these three norms for measuring the discrepancy between test 1 and test 2 and test 1 and test 3 are presented in Table 1, and confirm the known fact that norm choice may lead to different conclusions: One would conclude that test 2 is "closer" to test 1 when using the L_1 and and L_∞ norms, while the use of the L_2 norm would lead to the conclusion that test 3 is in fact closer to test 1. The major limitation of using norms (and the reason of the illustrated differences) is that they are not capable of distinguishing error due to phase from error due to magnitude. Even with this limitation, norms form the

2   Copyright © 2008 by ASME

Table 1. Results for the L_1, L_2 and L_∞ norms

Norm	test 1 and test 2	test 1 and test 3
L_1	0.3	0.45
L_2	0.6	0.58
L_∞	0.82	0.85

foundation for quantifying discrepancy between time histories.

2.2 Average residual and its standard deviation

The average residual measures the mean difference between two time histories:

$$\overline{R} = \frac{\sum_{i=1}^{N}(a_i - b_i)}{N}. \qquad (2)$$

A distinct disadvantage is that positive and negative differences at various points may cancel each other out. The standard deviation of residuals is defined as the square root of the sample variance of the residuals:

$$S_{N-1} = \sqrt{\frac{\sum_{i=1}^{N}(R_i - \overline{R})^2}{N-1}}, \qquad (3)$$

where $R_i = (a_i - b_i)$.

The results for the time history examples shown in Figure 1 are presented in Table 2. The results cannot lead to conclusive

Table 2. Results for average residual and its standard deviation

Measure	test 1 and test 2	test 1 and test 3
$\overline{R}$	0.8	3.85
S_{N-1}	7.7	6.4

statements regarding which test (2 or 3) is closer to test 1, as the measures of average residual and its standard deviation are conflicting.

2.3 Coefficient of correlation and cross-correlation

The coefficient of correlation is a measure that indicates the extent of linear relationship between two time histories, i.e., to what extent can A be represented as $mB + c$. The coefficient of correlation can range from -1 to $+1$. The value of $+1$ represents a perfect positive linear relationship between the time histories,

which implies that they are both identical in topology (or shape[1]). A value of -1 would indicate a perfect negative linear relation which would indicate that the two time histories are mirror images of each other. The coefficient of correlation is computed as

$$\rho = \frac{\sum_{i=1}^{N}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{N}(a_i - \overline{a})^2 \sum_{i=1}^{N}(b_i - \overline{b})^2}} \qquad (4)$$

The square of the coefficient of correlation is called the coefficient of determination and is commonly known as R-square.

The results of applying this measure to the previous time history examples are presented in Table 3, and indicate that test 3 is better correlated to test 1 than test 2. However, the R-

Table 3. Results for coefficient of correlation and R-square

Measure	test 1 and test 2	test 1 and test 3
ρ	0.5	0.6
R-square	0.25	0.36

square values for test 2 and test 3 are very low and hence neither seems to be close to test 1. This is mainly because these measures are sensitive to phase difference and cannot distinguish between error due to phase from error due to magnitude.

A modification to the concept of coefficient of correlation used in signal processing is called cross-correlation. It is sometimes called the sliding dot product, and has applications in the fields of pattern recognition and cryptanalysis. It can be used to measure the phase lag between two time histories. Cross-correlation is a series defined as

$$\rho(n) = \frac{(N-n)\sum_{i=1}^{N-n}a_i b_{i+n} - \sum_{i=1}^{N-n}a_i \sum_{i=1}^{N-n}b_{i+n}}{\sqrt{(N-n)\sum_{i=1}^{N-n}a_i^2 - \left(\sum_{i=1}^{N-n}a_i\right)^2}\sqrt{(N-n)\sum_{i=1}^{N-n}b_{i+n}^2 - \left(\sum_{i=1}^{N-n}b_{i+n}\right)^2}} \qquad (5)$$

for $n = 0, 1, \ldots, N-1$. To compute the phase difference between two time histories we determine the maximum value $\rho(n_*)$; n_* would then be a measure for phase lag. This concept has been used by Liu et al. [4] and Gu and Yang [5], and is also included as a metric in ADVISER, a commercial software package that contains a simulation model quality rating module[6, 7], for vehicle safety applications.

2.4 Sprague and Geers (S&G) metric

Geers proposed an error measure for comparing time histories that combined the errors due to magnitude and phase differences [8]. Recently, Sprague and Geers updated the phase error

---

[1]We use the terms topology and shape interchangeably.

portion of the metric [9, 10]. The error in magnitude and phase are computed for the time histories by using Equations (6) and (7), respectively. The combined error $C_{S\&G}$ is then used to provide an overall error measure between the two time histories.

$$M_{S\&G} = \sqrt{\frac{\psi_{AA}}{\psi_{BB}}} - 1 \qquad (6)$$

$$P_{S\&G} = \frac{1}{\pi} \cos^{-1}\left(\frac{\psi_{AB}}{\sqrt{\psi_{AA}\psi_{BB}}}\right) \qquad (7)$$

$$C_{S\&G} = \sqrt{M_{S\&G}^2 + P_{S\&G}^2} \,, \qquad (8)$$

where

$$\psi_{AA} = \frac{\sum_{i=1}^{N} a_i^2}{N} \qquad \psi_{BB} = \frac{\sum_{i=1}^{N} b_i^2}{N} \qquad \psi_{AB} = \frac{\sum_{i=1}^{N} a_i b_i}{N} \,.$$

The results of applying the S&G metric to the time history examples are presented in Table 4. The S&G metric quantifies a lower magnitude error for test 2 and a lower phase error for test 3. The combined error is lower for test 2, indicating that test 2 is closer to test 1 than test 3. The limitation of the S&G metric is that it is not symmetric. The results depend on the time history that is used as a reference in Equation (6).

Table 4.   Results for S&G metric

	t 1 vs t 2	t 2 vs t 1	t 1 vs t 3	t 3 vs t 1
$M_{S\&G}$	0.0824	-0.0761	0.6745	-0.4028
$P_{S\&G}$	0.2014	0.2014	0.1744	0.1744
$C_{S\&G}$	0.2176	0.2153	0.6967	0.4389

The separation of the error into magnitude and phase components is an advantage when more detailed investigation of the error sources is necessary. But, the metric lumps the entire information of the time histories into ψ_{AA}, ψ_{BB} and ψ_{AB}. Consequently, this metric cannot consider the shape of the time histories. This limitation is illustrated by the example in Figure 2: The two simple time histories have the same value for ψ_{AA} and ψ_{BB} but differ from each other in magnitude, phase and shape. Even though there exists an error in magnitude, the S&G metric quantifies it as zero.
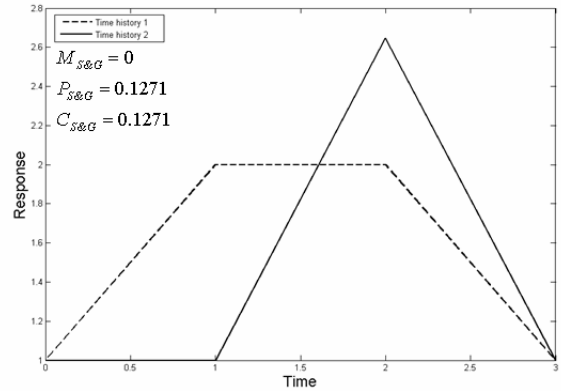

Figure 2.   Failure of S&G metric to quantify error due to magnitude

2.5 Russell's error measure

Russell proposed a set of magnitude, phase and comprehensive error measures to provide a robust means for quantifying the difference between time histories [11, 12]. The metric is similar to the S&G metric with a modification in the magnitude error factor. The magnitude error factor is defined such that it has approximately the same scale as the phase error when there exists an order-of-magnitude difference in amplitude of the responses. These are then combined to form the comprehensive error facto, similar to the S&G metric. The magnitude error factor is given by

$$M_R = \text{sign}(\psi_{AA} - \psi_{BB}) \log^{10}\left(1 + \left|\frac{\psi_{AA} - \psi_{BB}}{\sqrt{\psi_{AA}\psi_{BB}}}\right|\right). \qquad (9)$$

Even though Russell's error measure overcomes the limitation of asymmetry as observed in the S&G metric, it still fails in identifying and quantifying the magnitude error of the example shown in Figure 2.

2.6 Normalized Integral Square Error (NISE)

The Normalized Integral Square Error (NISE) is used to quantify the difference between acceleration histories from repeated tests, e.g., see [13]. It measures the difference between two time histories and is related in principle to the concept of cross-correlation. It considers three aspects: phase shift, amplitude (magnitude) difference and shape difference.

It uses the cross-correlation principle from Section 2.3 to compute n_*. It then shifts one of the time histories (A or B) relative to the other by n_* "steps" to compensate for the error in phase. The quantity $\psi_{AB}(n_*)$ is computed after this adjustment. The equations for the phase, magnitude and shape error are given

4   Copyright © 2008 by ASME

in (10), (11) and (12), respectively.

$$P_{NISE} = \frac{2\psi_{AB}(n_*) - 2\psi_{AB}}{\psi_{AA} + \psi_{BB}} \qquad (10)$$

$$M_{NISE} = \rho(n_*) - \frac{2\psi_{AB}(n_*)}{\psi_{AA} + \psi_{BB}} \qquad (11)$$

$$S_{NISE} = 1 - \rho(n_*) \qquad (12)$$

The overall NISE for two time histories is given by

$$C_{NISE} = P_{NISE} + M_{NISE} + S_{NISE} = 1 - \frac{2\psi_{AB}}{\psi_{AA} + \psi_{BB}}. \qquad (13)$$

Even though NISE accounts for error in shape, it can be observed that the overall measure (C_{NISE}) is independent of $\rho(n_*)$ and hence does not account for the effect of shape.

2.7 Dynamic Time Warping (DTW)

Dynamic time warping (DTW) is an algorithm for measuring discrepancy between time histories and was first used in context with speech recognition in the 1960's [14]. Since then, it has been used in a variety of applications: computer vision (e.g., [15]), data mining (e.g., [16]), signature matching (e.g., [17]), and polygonal shape matching (e.g., [18]).

The ability of time warping measurement to identify that two time histories with time shifts are a "match," makes it an important similarity identification technique [19] in speech recognition, since human speech consists of varying durations and paces. The time warping technique aligns peaks and valleys as much as possible by expanding and compressing the time axis accordingly [20].

The definition of DTW is based on the notion of a warping path. Let d be the $N \times N$ matrix of pairwise squared distances between the components of A and B: $d[i,j] = (a_i - b_j)^2$ (see Figure 3). The matrix d is called the cost matrix. The function used to calculate the value for each cell of d is called the cost function (in this example, the cost function is $(a_i - b_j)^2$; filled matrix entries represent cost of considered steps while non-filled entries are not considered as a step option). A warping path $W = \langle w_1, w_2, \ldots, w_K \rangle$ is a sequence of $K(N \le K \le 2N-1)$ matrix cells, $w_k = [i_k, j_k](1 \le i_k, j_k \le N)$, such that the following conditions are satisfied:

Boundary conditions: $w_1 = [1,1]$ and $w_K = [N,N]$, i.e., W starts in the lower-left cell and ends in the upper right cell.
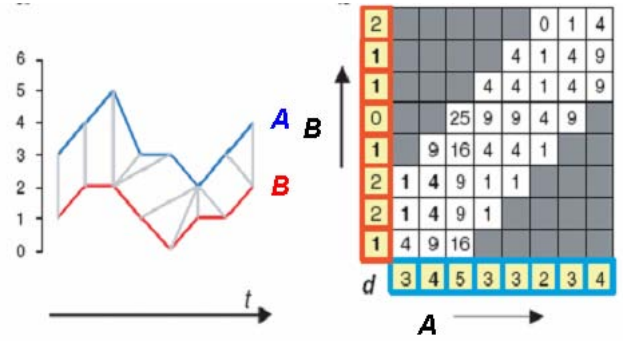

Figure 3. Warping path example

Continuity: given $w_{k-1} = [i_{k-1}, j_{k-1}]$ and $w_k = [i_k, j_k]$, then $i_k - i_{k-1} \le 1$ and $j_k - j_{k-1} \le 1$. This ensures that the cells of the warping path are adjacent.

Monotonicity: given $w_{k-1} = [i_{k-1}, j_{k-1}]$ and $w_k = [i_k, j_k]$, then $i_k - i_{k-1} \ge 0$ and $j_k - j_{k-1} \ge 0$, with at least one strict inequality. This forces W to progress over time.

Any warping path W defines an alignment between A and B and, consequently, a cost to align the two time histories. The squared DTW distance is the minimum of such costs, i.e., the cost of the optimal warping path W_{opt}:

$$(DTW(A,B))^2 = \min_W \sum_{[i_k,j_k] \in W} d[i_k, j_k] = \sum_{[i_k,j_k] \in W_{opt}} d[i_k, j_k] \quad (14)$$

The DTW distance can be recursively computed using an $O(N^2)$ dynamic programming approach that fills the cells of a cumulative cost matrix D using the recurrence relation

$$D[i,j] = d[i,j] + min\left(D[i-1, j-1], D[i-1, j], D[i, j-1]\right) \qquad (15)$$

and setting $DTW(A,B) = \sqrt{D[N,N]}$.

Using the cost function defined in the previous example, the DTW distance for test 2 and test 3 was 786 and 5636, respectively: Test 2 seems to be a closer representation of test 1.

The advantage associated with DTW is its capability to map many-to-one or one-to-many points relative to the Euclidean norm or the coefficient of correlation, which map one-to-one [21].

One of the concerns about the use of DTW distance is that it does not satisfy the triangle inequality, i.e., $DTW(A,B) + DTW(B,X)$ is not always $\ge DTW(A,X)$. Hence, it cannot be treated as a metric [22]. Moreover, DTW computation can be a time consuming operation due to the recursive dynamic programming approach discussed earlier. Modifications like Stream-DTW [23], Regression Time Warping [21] and Vector

5   Copyright © 2008 by ASME

Quantization-DTW [17] have been proposed to reduce the computational time of DTW at the expense of accuracy.

3 Proposed metric

Several measures used to quantify discrepancy (or error) between time histories have been discussed in the previous section. Each of them has its own advantages and limitations. The problem in quantifying error associated with three major features (phase, magnitude and topology) separately is that there exists a strong interaction among them. For example, to quantify the error associated with magnitude, the presence of a phase difference between the time histories may result in a misleading measurement. Thus, it is important to minimize the influence of the other two features when quantifying the error due to the third one.

In this section we propose a new metric by combining existing measures and algorithms to achieve independent measures for phase, magnitude and topology. The magnitude and phase error components are discussed adequately in the literature. We introduce a third component, the topology component to quantify error due to difference in number of peaks and valleys in the time histories.

The three aforementioned error components quantify the overall (or global) discrepancy between time histories. In certain applications, it may be beneficial to also distinguish error associated with localized areas of interest in the time histories. Such "target points" consider only a part of the entire time history and do not indicate an overall performance of the time history. When quantifying target point errors, we consider magnitude and phase only.

3.1 Phase error

To quantify the error due to phase, we considered the phase measure used by S&G and Russell in their metric (Equation (7)) and the cross-correlation technique presented in Section 2.3. The geometric interpretation of the S&G phase measure is related to the correlation coefficient. The cross-correlation based method for quantifying error in phase (used in [24]) shifts one of the time histories in order to maximize the correlation coefficient. This shift is considered to be the measure for error in phase.

We compared the performance of the cross-correlation method versus the S&G phase error, and concluded that the cross-correlation method has greater potential. An example to illustrate this is presented in Figure 4. It is evident that there exists a much larger phase difference between the CAE-1 and Test time histories than between the CAE-2 and Test time histories. However, the S&G phase error quantification was identical for both cases. The cross-correlation quantification, however, was able to distinguish between the two cases. Thus, we will use the cross-correlation technique to quantify error in phase in our metric.
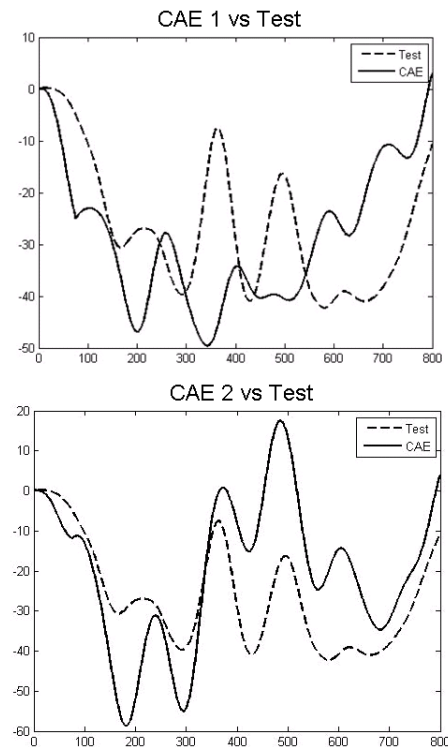

Figure 4. Example to compare S&G phase measure to cross-correlation

If we consider the number of time steps shifted n_* as the measure for phase error, we essentially consider a linear penalty. In most practical cases, small time step differences are treated as local errors, and need not be penalized at the same rate as large time step differences. In order to account for this, we propose a penalty function that can be modified and tuned to suit a particular application:

$$Error_{phase} = e^{\left(\frac{n_* - c}{r}\right)}, \qquad (16)$$

where c and r are parameters that define the rise start point and rate of increase for the function.

3.2 Magnitude error

To quantify the error only associated with magnitude, we need to first minimize the discrepancy between the time histories caused by error in phase and topology. We can compensate for global time shift by shifting the time history by the number of steps (n_*) computed in Section 3.1. The resultant time histories after time shift are referred to as time-shifted histories and are represented by A^{ts} and B^{ts}. However, time-shifted histories may still exhibit local "timing" errors. Moreover, errors due to dif-

6

Copyright © 2008 by ASME

ference in slope should not be treated as magnitude errors. To address these issues, we use Dynamic Time Warping (DTW).

The cost function for warping is defined such to penalize for distance and difference in slope between the two points:

$$d[i,j] = ((a_i^{ts} - b_j^{ts})^2 + (t_i - t_j)^2) \left| \left(\frac{dA^{ts}}{dt} \right)_{t=t_i} - \left(\frac{dB^{ts}}{dt} \right)_{t=t_j} \right|$$

(17)

This ensures the mapping of a point to the closest point having similar slope on the other time history.

Figure 5 depicts two time history examples before and after warping. The warped time histories are now represented as
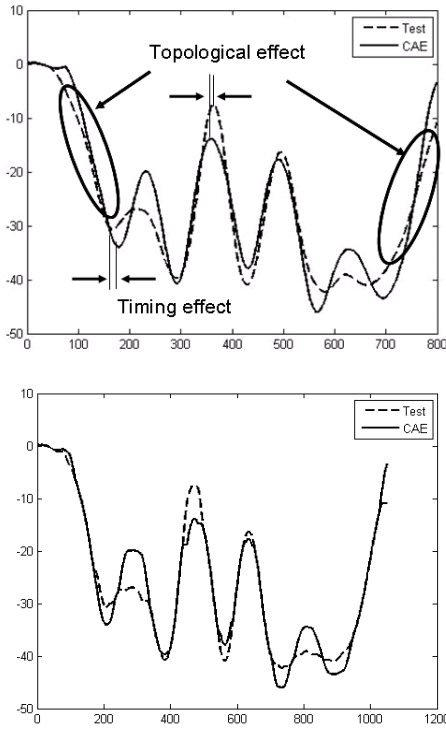

Figure 5.   Illustration of effect of warping on time histories

$A^{[ts+w]}$ and $B^{[ts+w]}$. It can be observed that warping minimizes the local phase and topology effects. We then use the L_1 on the warped time shifted histories to isolate the relative magnitude error between the two time histories:

$$Error_{magnitude} = \frac{\left\| A^{[ts+w]} - B^{[ts+w]} \right\|_1}{\left\| B^{[ts+w]} \right\|_1}$$

(18)

3.3 Topology error

Topology error is a measure of discrepancy in the shape of the two time histories. The topology of a time history could be classified by the number of times the curve crosses the mean magnitude value. To provide a more local measure of topology, and to enable a means to distinguish between phase/magnitude and topology, we employ the slope of the time history at each time point. Therefore, the topology error is computed on the derivative of the time histories. In order to ensure that the effect of global time shift is minimized, the slope is calculated for the time shifted histories. Thus, by taking the derivative at each point, we obtain "derivative time-shifted histories" represented by $A^{[ts+d]}$ and $B^{[ts+d]}$. Considering the derivative information ensures that the effect of magnitude is compensated for, as the derivative depends on the slope and not on the amplitude. The effect of localized time shifts still exist. Thus, we use the same methodology to evaluate the magnitude error on the derivative time shifted histories. The L_1 norm of the warped derivative time shifted histories is then used to quantify the isolated contribution of topology error:

$$Error_{topology} = \frac{\left\| A^{[ts+d+w]} - B^{[ts+d+w]} \right\|_1}{\left\| B^{[ts+d+w]} \right\|_1}$$

(19)

It should be mentioned that the topology error component will be hard to compute and perhaps even meaningless for highly noisy signals. Dealing with time histories rich in simulation noise is out of the scope of the work presented in this paper.

4 Example

In this section we demonstrate the application of the proposed metric using data from a case study provided by an International Standards Organization (ISO) working group on Virtual Testing (ISO technical committee (TC) 22, subcommittees (SC) 10 and 12, working group (WG) 4). An experimental test setup used available crash pulses to record acceleration time histories at different locations of a dummy during impact: head, thorax and tibia. The test setup for the head impact is shown in Figure 6. Three experiments were conducted to collect sets of test data. Figure 7 depicts a typical head acceleration history in the x-direction. Three computational models were developed to simulate these tests. Computational results and test data were used to rank the predictive capability of the computational models [25]. We present here calculations and rankings for three responses of the head impact case: head impactor, head acceleration in the x-direction and neck force in the x-direction.

We quantify error between the different tests and the computational models for each response individually. For each response, we compare tests among themselves to obtain the error
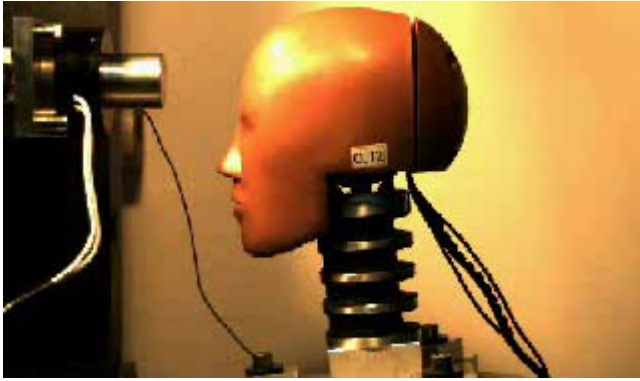
7

Copyright © 2008 by ASME

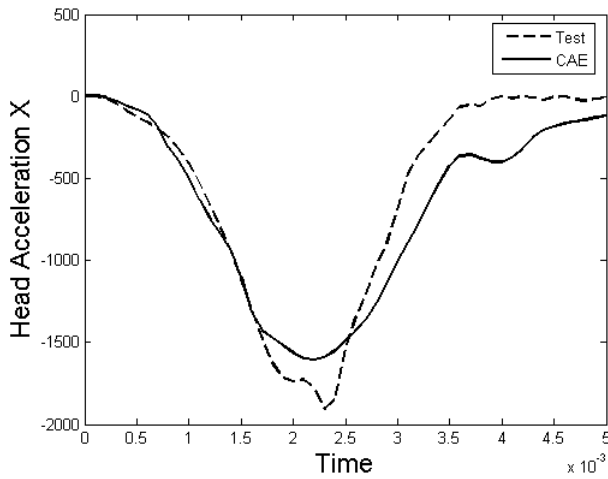Figure 6. Head impact test setup for ISO case study

Figure 7. Typical head acceleration time history in the x-direction

between test repetitions. We then compare the computational model predictions to each of these tests to obtain a measure for the discrepancy between test and computational data. If the error between tests is greater than or equal to the error between the computational model and the tests, we can conclude that the computational model is adequate.

To demonstrate this idea, we assume that test 1 represents "reality." We compare the remaining two tests and the three computational models to test 1. We then have three cases:

1. Looking at one response at a time, if the error associated with all three components (phase, magnitude and topology) for a computational model is less then or equal to the respective errors measured for the tests, then we can conclude that the computational model is a good representation of reality.
2. Looking at one response at a time, if all the three error com-

ponents for one computational are less than all the three for another, we can conclude that the first model is better than the second.

3. Looking across all responses, if we find that one computational model performs well for all of the responses, we can conclude that it is better than the other models collectively.

Figure 8 depicts the results for the three considered responses. Looking at the head impactor response, we can see that all three error components for all three computational models relative to test 1 are less or equal to the errors of test 2 and test 3 relative to test 1. Thus, we can conclude that they are all adequate. Moreover, there are no differences between all three models, so they are all ranked equal. Looking at the head accel-
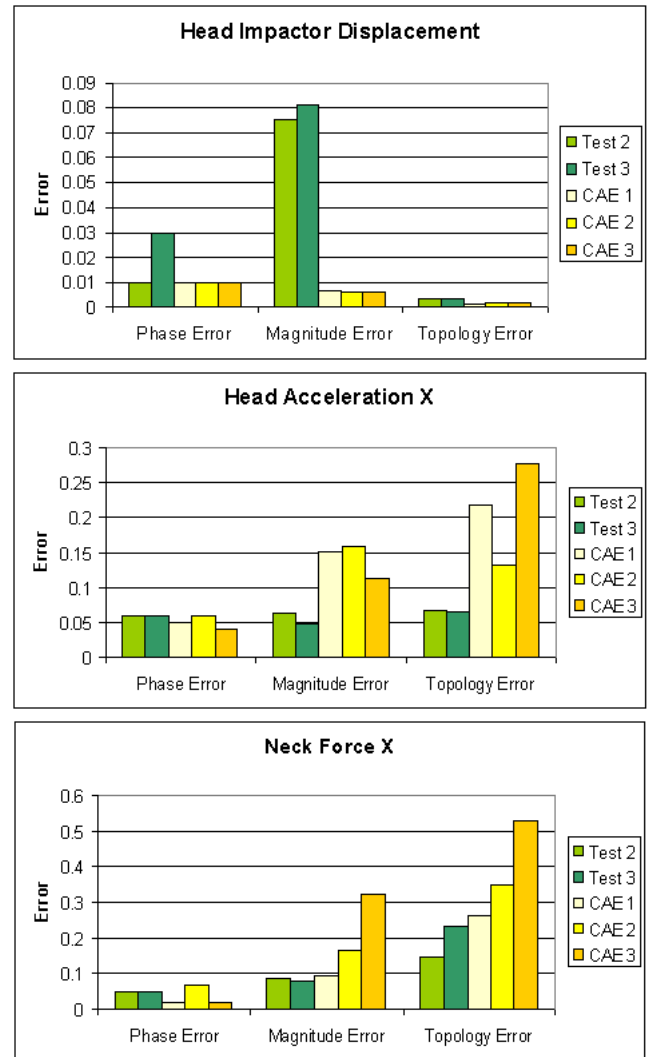

Figure 8. Sample of results for head impact case

8

eration in the *x*-direction, we see that the computational models have acceptable error only in the phase component; they are not adequate in regard to magnitude and topology errors. Among themselves, the three computational models do not exhibit consistently better or worse errors, so no conclusive ranking can be made. A similar situation exists for the neck force response in the *x*-direction: Again, only the phase error is acceptable for all three models. However, the computational models exhibit consistent magnitude and topology errors, with model 1 being the best and model 3 being the worse. In this case we can rank the models without concluding whether their prediction capabilities are acceptable or not.

5   Building regression-based validation models using ratings of subject matter experts

We now consider a case where ratings of subject matter experts (SMEs) are available. Subject matter experts are individuals with long experience in a particular discipline. They are thus trusted to evaluate and rank the predictive capability of computational models by (mostly visual) inspection of comparison plots. We use SME ratings of computational models and the three component errors as quantified by our proposed metric to build a regression-based validation model that can validate and/or rank other computational models.

We consider a case (previously reported in [26]), where a deceleration time history from a crash is known by means of measurement (physical experiment). Fifteen computational models have been developed to predict the deceleration time history for this crash (note that these are not necessarily different computational models; they may be fifteen different substantiations of the same computational approach due to different parameters in the numerical models). Six SMEs have been presented with fifteen comparison plots (one for each model) and their average ranking of the models has been recorded. Ratings range from 1 (worst match) to 10 (excellent match). Figure 9 depicts a typical plot shown to the SMEs.

We used ten of the available fifteen data sets and ratings to build our regression-based validation model. We then used the remaining five data sets to test our model. Obviously, there are many combinations of which ten data sets to use to build the regression model, but this discussion is out of the scope of this paper (please refer to [25] for a full discussion of this case study). Table 5 presents the individual and average SME ratings (in ascending order) for the time histories associated with the training and test data sets. Each computational model (CAE) is identified with an ID number.

The error components computed using our proposed metric are summarized in Table 6 in the order of Table 5 to facilitate comparisons. Note that the relatively large phase error for CAE 1189 is not a typing error: this model exhibits a large phase error as reflected by the low SME rating. We now can combine all
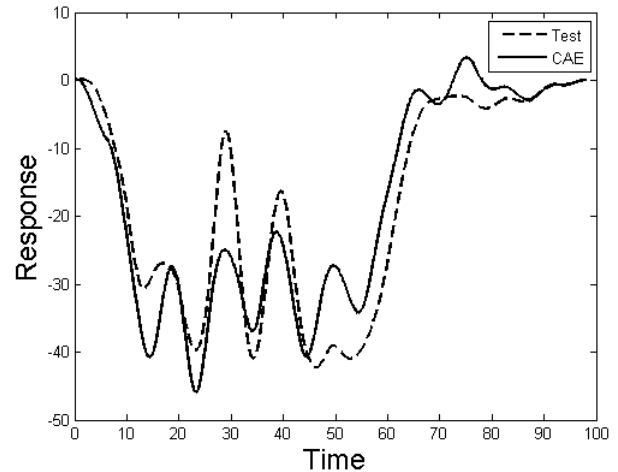

Figure 9.   A typical plot presented to the SMEs

three error components into a regression model that can predict predict the average expert ratings. We built a linear regression model, i.e., used the following first-degree polynomial to fit our data to the SME average ratings

$$R_p = 10 - \left(c_1 Error_{phase} + c_2 Error_{magnitude} + c_3 Error_{topology} \right), \tag{20}$$

where R_p denotes predicted rating.

The plots in Figure 10 depict the regression model rating predictions (denoted by the acronym EARTH – Error Assessment of Response Time Histories) on the ten time histories used to build the model and on the five remaining time histories relative to the average SME ratings (the bars for the SME ratings represent the range of the SME ratings). It can be seen that the validation model predictions are good and always within the range of the individual SME ratings. this is the case for all regression models we built using different combinations of training and test time histories, i.e., data sets [25].

Lastly, we compare the rating predictions of our regression-based validation model to the rating predictions of four existing metrics used currently for this particular application [26]:

1. Wavelet decomposition method
   This metric uses a pre-processing operation on the original time histories to split into two base functions called the scale and envelop functions. These functions can be thought of as component time histories that, when added together, would result in the original time history. The discrepancy for both the component histories is computed using the Russell's measure. The error for both the scale and envelop function is then combined using a regression approach similar to EARTH.

9   Copyright © 2008 by ASME

Table 5. SME ratings for the fifteen CAE models (first ten models used to build the regression-based validation model; last five used to test it)

CAE ID	SME 1	2	3	4	5	6	Average
1188	5	3	1	2	3	4	3.00
1189	3	4	4	3	3	3	3.33
1130	4	4	4	3	4	5	4.00
1047	5	4	5	4	5	5	4.67
1020	6	4	5	6	7	4	5.33
1041	6	5	5	6	6	5	5.50
1028	7	6	5	5	7	6	6.00
1005	8	7	7	6	8	6	7.00
1083	7	7	7	9	9	7	7.67
1052	8	7	8	10	10	8	8.50
1042	4	3	4	3	4	4	3.67
1100	5	4	3	3	4	6	4.17
1009	7	6	6	7	7	6	6.50
1016	7	7	7	8	9	5	7.17
1022	8	9	8	10	10	8	8.83

Table 6. Error components of proposed metric for the CAE models

CAE ID	Phase	Magnitude	Topology
1188	0.52	0.42	0.43
1189	51.94	0.20	0.33
1130	1.73	0.31	0.51
1047	0.67	0.22	0.41
1020	0.61	0.19	0.48
1041	0.50	0.22	0.31
1028	0.52	0.19	0.27
1005	0.50	0.16	0.23
1083	0.52	0.12	0.33
1052	0.52	0.09	0.25
1042	0.64	0.33	0.45
1100	1.16	0.28	0.69
1009	0.70	0.16	0.28
1016	0.61	0.17	0.44
1022	0.52	0.08	0.22

2. Step function
Most crash pulses for this application can be characterized as step functions; each period of a step function reflects a certain impact event in the vehicle structure. The parameters for the step function that are used to quantify the discrepancy are the slope for the first edge, the value for the first and second steps, the duration of the first and second steps, the peak value and its timing, maximum crush distance and variation computed for the value of the first and second steps. All of these measures are combined using the regression approach similar to EARTH.

3. ADVISER model evaluation criteria
This metric is a combination of target point error measures. For this particular application, ten target point errors are considered. These target point errors correspond to timing of first and second extreme, time and value for first and second edge, $\rho(n_*)$, n_* and time and value score of the shape corridor. The time and value score of the shape corridor is evaluated by creating a corridor around the experimental time history and comparing the CAE time history to this corridor. The time score is based on the duration hat the CAE time history lies "outside" the corridor. Similarly, the value score is based on the maximum distance between the

CAE time history and the corridor when a violation occurs. All the different error scores are used to form a regression model that combines them.

4. Corridor Violation Plus Area (CVPA)
This metric evaluates the error associated with the integral of the time histories. It uses a similar concept as the shape corridor and evaluates time score and an area score. The area score is computed based on the area enclosed by the CAE time history outside the corridor. The time and area scores are computed for multiple corridor widths. These different scores are combined using linear regression.

The performance of EARTH is compared to these metrics in Figure 11: The top plot illustrates that the EARTH and wavelet decomposition metrics predict the SME ratings very well. The bottom plot shows results for a different EARTH regression model (built using a different combination of data sets): in this case the EARTH and step function metrics predict the SME ratings very well. In all the regression model we built, EARTH consistently predicted SME ratings well. This indicates that EARTH is capable of recognizing the key features associated with the time histories for this application and provide an over all error measure by combining them. EARTH is capable of bringing in an objective method to evaluate time histories even though it is de-
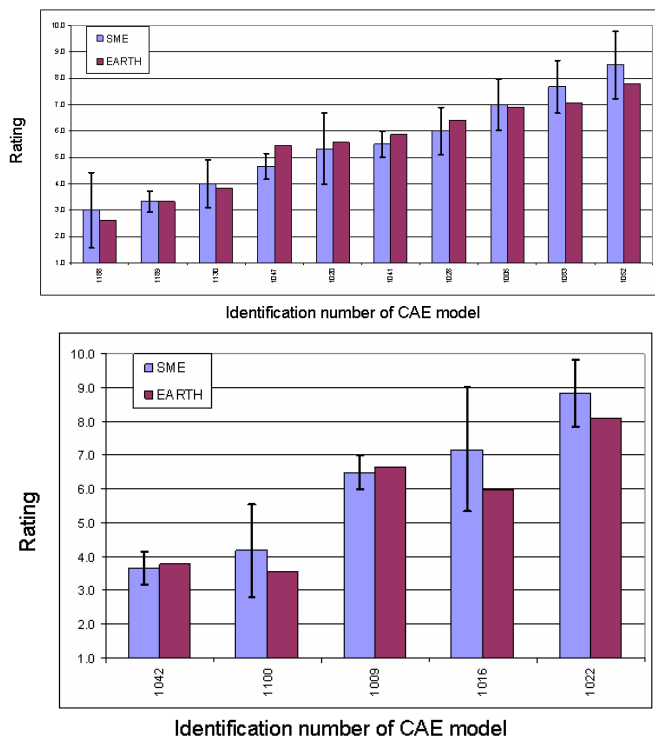
Copyright © 2008 by ASME

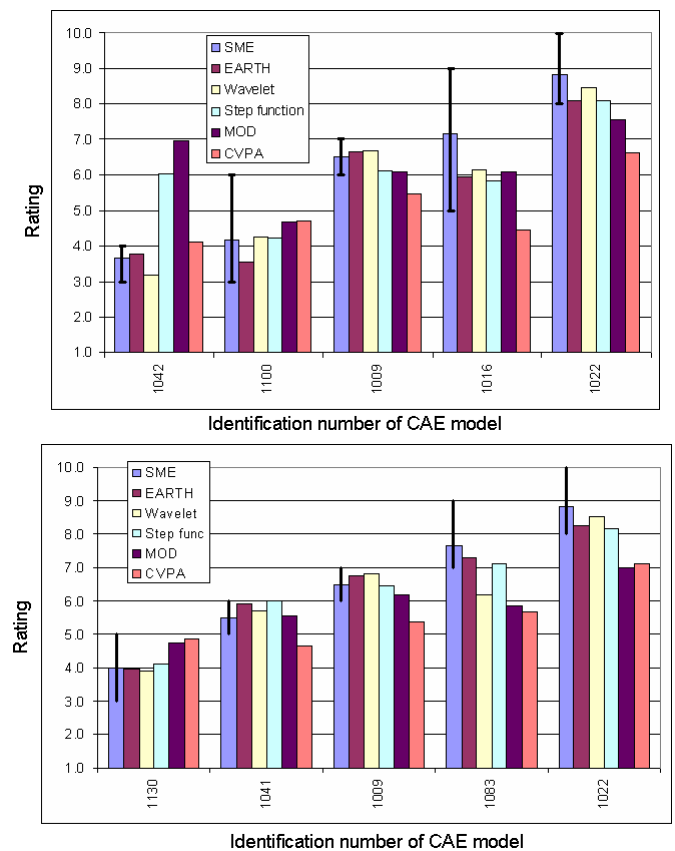Figure 10.   Regression-based validation model: data fit and test

Figure 11.   Comparison of EARTH to other metrics

veloped on the subjective opinion of SMEs.

6   Concluding remarks

The objective of the research presented in this paper was to evaluate existing measures for assessing the error between time histories and to introduce a combined approach to remedy limitations of these metrics when applied to vehicle safety.

We proposed a new metric that quantifies error associated with three components: phase, magnitude and topology. We adopted the idea of classifying error into phase and magnitude based on existing metrics like S&G, Russell's error measure and NISE criteria. At the same time, identifying their inability to quantify error due to difference in shape, we introduced the concept of topology error to account for discrepancy in shape. Lastly, we introduce the use of the dynamic time warping (DTW) algorithm to process time history data so that the three error components can be isolated and quantified.

The applicability of the proposed metric has been demonstrated through two case studies pertaining to vehicle safety. The first case study illustrates how the proposed metric can be used to assess predictive capability of computational models. The second case study showed how the metric can be used in conjunction with subject matter expert (SME) data to develop regression-based models to validate simulation models. A comparison with four existing metrics for model validation in vehicle safety applications demonstrated that the proposed metric agrees consistently with SME ratings.

Acknowledgements

REFERENCES

[1] American Institute of Aeronautics and Astronautics. *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, 1998.

11

[2] American Society of Mechanical Engineers. *Council on Codes and Standards, Board of Performance Test Codes: Committee on Verification and Validation in Computational Solid Mechanics*, 2003.

[3] W.L. Oberkampf and M.F. Barone. Measures of agreement between computation and experiment: validation metrics. *Journal of Computational Physics*, 217(1):5–36, 2006.

[4] X. Liu, F. Yan, W. Chen, and M. Paas. Automated occupant model evaluation and correlation. In *Proceedings of the 2005 ASME International Mechanical Engineering Congress and Exposition*, Orlando, Florida, 2005.

[5] L. Gu and R.J. Yang. CAE model validation in vehicle safety design. *SAE Technical Paper Series*, 2004.

[6] *ADVISER Reference Guide*, 2.5 edition, Jan 2007.

[7] C. Jacob, F. Charras, X. Trosseille, J. Hamon, M. Pajon, and J.Y. Lecoz. Mathematical models integral rating. *IJCrash*, 5(4):417–431, 2000.

[8] T.L. Geers. Objective error measure for the comparison of calculated and measured transient response histories. *Shock and Vibration Bulletin*, pages 99–107, Jun 1984.

[9] M.A. Sprague and T.L. Geers. A spectral-element method for modeling cavitation in transient fluid-structure interaction. *International Journal for Numerical Methods in Engineering*, 60(15):2467–2499, 2004.

[10] L.E. Schwer. Validation metrics for response histories: A review with case studies. Technical report, Schwer Engineering & Consulting Service, 6122 Aaron Court Windsor CA 95492 USA, Feb 2005.

[11] D.M. Russell. Error measures for comparing transient data: Part I, development of a comprehensive error measure. In *Proceedings of the 68th Shock and Vibration Symposium*, Hunt Valley, MD, 1997.

[12] D.M. Russell. Error measures for comparing transient data: Part II, error measures case study. In *Proceedings of the 68th Shock and Vibration Symposium*, Hunt Valley, MD, 1997.

[13] B.R. Donnelly, R.M. Morgan, and Eppinger. Durability, repeatability and reproducibility of the NHTSA side impact dummy. In *27th Stapp Car Crash Conference*, 1983.

[14] L.R. Rabiner and B.H. Huang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[15] M. Munich and P. Perona. Visual identification by signature tracking. *IEEE Transactions PAMI*, 25(2):200–217, 2003.

[16] T. Oates, L. Firoiu, and P. Cohen. *Using dynamic time warping to bootstrap HMM-based clustering of time series*, pages 35–52. Springer-Verlag, 2000.

[17] M. Faundez-Zanuy. On-line signature recognition based on VQ-DTW. *Pattern Recognition*, 40:981–992, 2007.

[18] E.M. Arkin, L.P. Chew, D.P. Huttenlocher, K. Kedem, and J.S.B. Mitchell. An efficiently computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):209–216, 1991.

[19] A. Efrat, Q. Fan, and S. Venkatasubramanian. Curve matching, time warping, and light fields: New algorithms for computing similarity between curves. *Journal of Mathematical Imaging and Vision*, 27(3):203–216, Apr 2007.

[20] F. Chan, A. Fu, and C. Yu. Haar wavelets for efficient similarity search of time-series: With and without time warping. *IEEE Transactions on Knowledge and Data Engineering*, 15(3), 2003.

[21] H. Lei and V. Govindaraju. Synchronization of batch trajectory based on multi-scale dynamic time warping. In *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, Nov 2003.

[22] E.V. Ruiz, F.C. Nolla, and H.R. Segov. Is the DTW "distance" really a metric? An algorithm reducing the number of DTW comparisons in isolated word recognition. *Speech Communication*, 4:333–344, 1985.

[23] P. Capitani and P. Ciaccia. Warping the time on data streams. *Data & Knowledge Engineering*, 62:438–458, 2007.

[24] Y. Chang and P. Seong. A signal pattern matching and verification method using interval means cross correlation and eigenvalues in the nuclear power plant monitoring systems. *Annals of Nuclear Energy*, 29:1795–1807, 2002.

[25] H. Sarin. Error Assessment of Response Time Histories (EARTH): A metric to validate simulation models. Master's thesis, University of Michgan, Ann Arbor, MI, 2008.

[26] R.J. Yang, G. Li, and Y. Fu. Development of validation metrics for vehicle frontal impact simulation. In *Proceedings of the ASME 2007 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Las Vegas, Nevada, 2007.