IDIAP COMMUNICATION

IN SEARCH OF A GOOD BET

A PROPOSAL FOR A BROWSER EVALUATION TEST

Mike Flynn & Pierre Wellner

flynn@idiap.ch, *wellner@idiap.ch*

IDIAP COM-03-11

SEPTEMBER 2003

Dalle Molle Institute
for Perceptual Artificial
Intelligence ● P.O.Box 592 ●
Martigny ● Valais ● Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

1. Introduction

It is becoming ever simpler and economical to capture multi-modal multi-media digital recordings of meetings. While it is straightforward to play back such recordings, it is much more laborious for users to browse for elements of interest. The development of new technology to enhance browsing of recorded meetings has therefore become an active area of research.

In previous work, the evaluation of meeting browsers is either absent [1] [3] [4] [7] [8] or very subjective [2] [5]. The evaluation in [6] attempted to measure performance objectively, but using a task selected for simplicity. In general, user tasks and the questions asked of users vary widely, are often loosely defined and final scores are open to considerable interpretation. Most importantly, however, it is not currently possible to compare browsers and browsing techniques objectively.

As with many other fields of research, an objective measure of system performance can be of enormous benefit in helping researchers make progress towards solving a particular problem. For example, word error rate is used to evaluate speech recognition systems; compression ratio is used to evaluate compression systems; and a corresponding *browser evaluation test* (BET) is needed to evaluate media browsers.

> For this proposal, we consider the task of ***browsing*** media as an attempt to find a maximum number of ***observations of interest*** in a minimum amount of time.

A key problem in testing browsers, therefore, is identifying the relevant 'observations of interest'. The range of possibilities is enormous and depends upon meeting content and individual user interests. The method described here identifies observations of interest based on the impressions of ordinary people and not based on the particular interests of the experimenter or browser designer.

We aim to make our test:

a) an objective measure of browser effectiveness based on user performance rather than judgement;

b) independent of experimenter perception of the browsing task and meeting structure;

c) produce directly comparable numeric scores, automatically; and

d) replicable, through a publicly accessible web site.

The remainder of this document details the method we propose. We would be grateful for any constructive criticism.

2. Overview of Proposed Method

With the BET, illustrated in *Figure 1* below, observations of interest within a recording are identified not by experimenters, but by ordinary people whose interests are likely to be relatively diverse. Collectively, these people generate a wide range of observations (simple statements about the meeting), providing greater diversity than would be achieved by experimenters alone.

Although a single person is unlikely to make the same observation multiple times, we expect the most significant features of each meeting to be observed by multiple people, albeit in different forms. Thus, samples drawn from the set of all observations can include multiple instances of common points of interest, and the statistical distribution of observations should therefore reflect their relative importance.

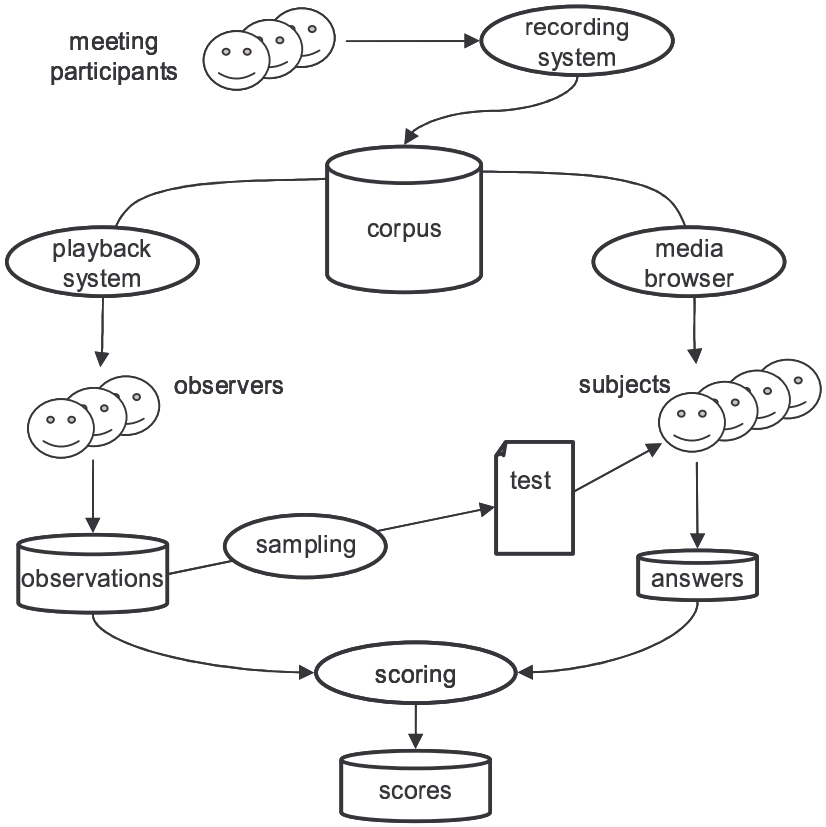

Figure 1 – Overview of the BET

To this end, the BET involves three distinct groups of people, in addition to the experimenters:

- *Participants* are the people in the original recorded meeting;
- *Observers* determine the questions and answers for the test;
- *Subjects* use a browser to answer the questions, thereby determining a score.

There are also several bodies of data:

- a *corpus* of media recordings;
- a collection of *observations*, made by the observers;
- a collection of *questions*, drawn from the observations;
- a collection of *tests*, embodying such questions.

These roles and data are examined in more detail in the following sub-sections, which describe how tests are constructed, administered and scored.

3. The Corpus

A significant set of media recordings, the *corpus*, provides the data to be browsed. The BET can be applied to a number of different types of corpus (*e.g.* news videos, home videos, or meeting recordings), but our initial application is meeting recordings.

Design of the corpus has enormous influence on the BET. The corpus determines the observations made, the questions asked, and ultimately the browsing behaviour of the subjects.

BET results obtained with the use of one corpus are not directly comparable to results obtained with another corpus. This implies that the corpus must be available to anyone performing a BET, so should not contain sensitive information. It also implies that the relevance of BET scores to real browser applications is dependent on the relevance of the corpus to these applications.

The primary purpose of this proposal is to describe the BET, not the corpus, but ideally the corpus should contain recordings of *real meetings*.

To facilitate the selection of diverse observers and subjects, the content of the corpus should also be comprehensible to a wide audience. Both observers and subjects need to able to follow discussions, reasoning and conflicts within a meeting, although not necessarily in every detail. For example, planning a social event or a common organisational issue may be preferable to discussing the mathematics behind a new algorithm.

4. Collecting Observations

Determination of the questions to be used in a test is performed, indirectly, by a set of observers.

The observers independently (*i.e.* alone) watch selected meetings from the corpus. Observers have available the full recordings from every source, in parallel. They may choose a source for closer scrutiny, and rewind and replay the sources, as they desire. There is no time limit for the observers, but initial experience indicates that six times the length of the original meeting may be required.

Each observer is instructed to produce observations that the meeting participants, or absentees, might have thought 'significant'. Asking observers to take the perspective of participants may help temper undue influence of each observer's own special interests (*e.g.* someone who finds hairstyles more significant than issues discussed). We expect that the types of observations collected in this manner will reflect the same types of observations that typical users would like to use a browser to find. Instructions are given in a standard manner (a written document, or perhaps a video), made available with the corpus.

Each observations is stated as a matched pair, one true and one false, only one of which is later presented to a particular subject during testing. Observers should aim to produce observations that are neutral, in that it should not be easy to guess their truth without the use of a browser. In order to ensure this, guessable observations are identified and removed during the control-score setting procedure, detailed in the next section.

The observations should be simply and concisely stated. To encourage this, observations are typed and submitted via a web form, where the box for the observation text is small, without scrolling. The observer associated with each observation is also recorded.

The number of observation pairs made by one observer for a given meeting is not determined, but is expected to be between ten and twenty per hour. Playback is halted while each observation pair is written. The observations are time-stamped with the offset into the recording at which it was made. Later, this may be useful for determining the temporal correspondence between questions and their answers.

Finally, each observer is given a questionnaire, recording personal and professional details, so that these variables are available to be checked for possible influences on scoring.

5. Establishing Control Scores

It may prove useful to the scientific community to establish and publish some 'control' scores, for reference purposes. Three separate scores would be collected from subjects using each of the following conditions:

a) educated guesses with no media present;

b) the same software as did the observers;

c) well-known basic media applications, such as Microsoft's Windows Media Player, or RealPlayer;

In the process of running condition (a) above, statistics are gathered concerning how easy it is to guess the veracity of each observation. Those observations that are easiest to decide are removed (along with their converse) in order to decrease the number of subjects ultimately needed for a statistically sound result.

6. Testing a Browser

New multi-media browser designs are tested as follows.

Subjects are neither participants nor observers, and preferably have no direct or vested interest in the content of the corpus. Their task is well defined and effectively determined by the observers, so the precise background and interests of each individual subject should not be critical.

Subjects take several tests, each of which requires them to use *the same* browser, to examine one of several meetings, one per test. That is, the test is administered "between-subjects" – a necessity, as other researchers may later test other browsers elsewhere. The order in which each meeting is presented is counterbalanced across subjects, to avoid any sequence effect.

Each test is a set of *questions* drawn from the observations by asking whether the observed statement is true or false (or unknown). All the questions in each test are presented to the subject simultaneously, in one web page, in a random order. The subject may answer the questions in any order, and revise answers as necessary.

Tests have a time limit of half the duration of the meeting under examination. This is partly to ease timetabling of subjects, but also to prevent a simple playback of the whole meeting from satisfying the questions. While such a time limit may prevent finding the ultimate capabilities of a browser, nevertheless, some time pressure is required in order to emphasize "the minimum time" stipulation from our definition of browsing. Each answer to a test is time-stamped and subsequent analysis may show an increase in speed as familiarity is gained and a slow-down as relatively easy questions are picked off first.

Only a small sample of observations is used in a particular test, drawn randomly, and no observation is used more than once in the same test, nor the complementary observation. In order to avoid a ceiling effect, the number of questions in a test is chosen to make it difficult and unusual to finish the test within the time limit without randomly guessing the answers. This number is determined as the average number of observations made by each observer for the meeting under consideration.

A test is finished either by expiry of the time limit, or by the subject submitting their answers to be scored. There is no benefit to a subject in finishing before the time limit has expired, except that answers may be double-checked for accuracy. However, there is effectively a penalty for not completing questions, in that their score is likely to be lower.

The order in which questions are answered or revised is recorded, along with the time taken. This is to allow the measurement of any speed increase, as subjects become familiar with the browser. The time at which each answer is made, both in real time and in the media offset, is recorded too, for later analysis. In addition, the number of meetings and total time with the browser are recorded.

This testing process is entirely automatic, and potentially extending over considerable time. Any new subjects with any new browser may take a test for some meeting in the corpus. The media material may be a local copy, but tests are administered via the web. This is in order to simplify using many possibly unknown subjects, but does not imply that any browser must, itself, be web-based.

Subjects are given their score, and the average score to date, to satisfy their curiosity. Finally, each subject is given a background questionnaire, as for the observers, for later analysis.

7. Scoring and Analysis

Scoring is performed over the web in the following manner.

Each question in a test has four possible answers: `True`, `False`, `Unknown` or `Unanswered`. The 'correct' answer is that originally given by the observer, either `True` or `False`. Correct answers are rewarded, and incorrect answers penalized, so that arbitrary answers produce no gain. Specifically, a numerical score is determined for each question as shown in *Table 1* below.

Table 1 - Scores for each question

Subject's Answer	Observer's Answer	
	`True`	**`False`**
`True`	1	-1
`Unknown/Unanswered`	0	0
`False`	-1	1

The score for a test is thus the difference between the number of correct answers, and the number of incorrect answers, expressed as a percentage of the number of questions in the test. A perfect score for a test would therefore be 100%, while random answers would yield around 0%. Negative scores could only be achieved by a subject performing consistently worse than random.

The score for a browser is the average of the test scores obtained through many subjects using it, for many meetings. However, it is to be expected that subjects would learn to use the browser more effectively with time. Thus the number of tests each subject has performed is noted so that increases in score with use can be registered. In particular, browser scores should only be compared across equally experienced subjects, in the absence of knowledge of users' learning curves.

In addition to the overall score for each browser, it may prove worthwhile to plot the increase in score over time, to its final value, both within a single test, and over the course of multiple tests. Increasing familiarity with a browser should be evident, along with how quickly familiarity with a new meeting recording is achieved.

The results of the background questionnaires, given to both observers and subjects, may help determine whether subjects could answer questions set by observers of similar background any better than average. This might inform further development of the BET.

8. Statistics

Because the BET makes use of human subjects, it is critical to understand how much effort is consumed in the process of producing statistically significant results. This section illustrates (a) the effort required to create a set of corpus observations, (b) the effort to test a single browser, and (c) the statistical significance of the results.

For this illustration, assume a corpus containing 20 meetings, each 40 minutes long on average, the total duration of recorded media is 13 hours and 20 minutes.

a) *Observers & observations*

Observation by inexperienced observers appears to take as long as annotation, if not longer. Initial experience indicates a factor of 6 over real time is required, or a total of 240 hours of observation time.

Assume that each observer watches just 1 meeting, and that each meeting is observed by 3 observers. Then, the number of observers required for the whole corpus is 60. If each observer makes an average of 18 observation-pairs per hour, then the corpus will contain 4,320 observation-pairs, or 216 per meeting.

b) *Testing*

A test is half as long as the meeting upon which it is based, so each will have a time limit of 20 minutes, on average. During that time subjects are expected to complete, at most, around 40 questions – more typically 20 questions or less.

If each subject is tested upon 8 meetings, with each meeting tested by 4 subjects, then 10 subjects are required for each browser under test.

The 10 subjects will spend 2 hours 40 minutes each, or 26 hours and 40 minutes between them, yielding around 160 answers each, or 1,600 in total. A quicker browser, or more guessing, may yield more answers.

c) Significance

If 90% of all questions are answered ‘correctly’ and assuming a binomial distribution of the results, then for 1,600 independent answers, the true result lies within the confidence limits of 88.2% and 91.6%, with a confidence level of 95%.

It is possible that the same subject may be asked similar questions about the same meeting from different observers, bringing into question the independence of the answers. Analysis of observations in the corpus may be required to determine the impact of this on confidence intervals.

9. Conclusion

This document introduced an evaluation test for multi-media meeting browsers. The aims of this test, as laid out in the introduction, are satisfied as follows.

a) An objective measure of browser effectiveness based on performance rather than judgement.

Scores for a browser are produced automatically from the answers to simple questions. The questions are clear binary choices, without need for users to judge degree, or please the experimenters.

b) Independent of experimenter perception of the browsing task and meeting structure.

The questions are not produced by the experimenters, but by a significant number of distinct observers. These people do not have, necessarily, any direct connection with the experiment, browser design, or the field of human-computer interaction.

c) Produce directly comparable numeric scores, automatically.

The final result for a browser is a simple numeric score with a confidence interval. Richer data may underlie this, such as how quickly familiarity was gained.

d) Replicable, through a publicly accessible web site.

The corpus, questions, test instructions and mechanism will be freely available to the public.

In summary, this BET eliminates much of the subjectivity in typical multi-media browser tests. It may be possible to apply the same technique to areas other than media browsers, where it is desirable to test actual performance rather than user perceptions.

10. Acknowledgements

The authors acknowledge financial support provided by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2. The NCCR is managed by the Swiss National Science Foundation on behalf of the Federal Authorities.

The authors also wish to thank their colleagues, in particular Samy Bengio, Marge Eldridge, Daniel Gatica-Perez, Maël Guillemot, and Iain McCowan for discussion and comments on earlier drafts of this document.

11. References

[1] Belle L. Tseng, Ching-yung Lin and John R. Smith, “Video Summarization and Personalization for Pervasive Mobile Devices”, In *SPIE Proceedings*, vol. 4676, 2001, p. 359-70, 2001.

[2] Ross Cutler, Yong Rui, Anoop Gupta, JJ Cadiz, Ivan Tashev, Li-wei He, Alex Colburn, Zhengyou Zhang, Zicheng Liu, Steve Silverberg. “Distributed Meetings: A Meeting Capture and Broadcasting System”, *ACM Multimedia*, 2002.

[3] A. Divakaran and R. Cabasson, “Content Based Browsing System for Personal Video Recorders”, ICCE, Los Angeles, June 16-20, 2002.

[4] Divakaran, A.; Radhakrishnan, R.; Peker, K.A., “Motion Activity-Based Extraction of Key-Frames from Video Shots”, *IEEE International Conference on Image Processing (ICIP)*, ISSN: 1522-4880, Vol. 1, pp. 932-935, September 2002.

[5] Andreas Girgensohn, John Boreczky, Lynn Wilcox, “Keyframe-Based User Interfaces for Digital Video”, In *IEEE Computer*, Vol. 34, No. 9, pp. 61-67, September 2001.

[6] M. Guillemot, P. Wellner, D. Gatica-Pérez & J-M. Odobez, “A Hierarchical Keyframe User Interface for Browsing Video over the Internet”, In *Proceedings of the 9th IFIP International Conference on Human-Computer Interaction INTERACT 2003*, ETHZ, Zurich, Switzerland, 2003.

[7] Dar-Shyang Lee, Berna Erol, Jamey Graham, Jonathan J. Hull and Norihiko Murata, "Portable Meeting Recorder", *ACM Multimedia 2002*, pp. 493-502.

[8] Chung Wing Ng, Michael R. Lyu, “ADVISE: Advanced Digital Video Information Segmentation Engine”, In *Proceedings of the 11th International World Wide Web Conference*, Honolulu, Hawaii, USA, May 2002.