IDIAP RESEARCH REPORT

A MEETING BROWSER EVALUATION TEST

Pierre Wellner, Mike Flynn, Simon Tucker, Steve Whittaker

flynn@idiap.ch, wellner@idiap.ch, s.tucker@sheffield.ac.uk, s.whittaker@sheffield.ac.uk

IDIAP RR-05-02

In *Proceedings of CHI05 Conference on Human Factors in Computing Systems*, New York: ACM Press, 2005

IDIAP Research Institute ● 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

A Meeting Browser Evaluation Test

Pierre Wellner **Mike Flynn**
IDIAP Research Institute
Rue du Simplon 4, CH-1920 Martigny,
Switzerland
{flynn,wellner}@idiap.ch

Simon Tucker **Steve Whittaker**
Department of Information Studies
University of Sheffield, Regent Court, 211
Portobello Street, Sheffield, S1 4DP, UK
{s.tucker,s.whittaker}@sheffield.ac.uk

ABSTRACT
We introduce a browser evaluation test (BET), and describe a trial run application of the test. BET is a method for assessing meeting browser performance using the number of *observations of interest* found in the minimum amount of time as the evaluation metric, where observations of interest are statements about a meeting collected by independent observers. The resulting speed and accuracy scores aim to be objective, comparable and repeatable.

ACM Classification Keywords
H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – evaluation/methodology.

INTRODUCTION
Meetings are an integral part of our working lives. Recent developments in recording and storage techniques have made multimodal meeting recordings readily available, and while it is straightforward to play back such recordings, it is much more laborious for users to browse them. Devising new technology to enhance browsing of recorded meetings has therefore become an active area of research [7].

One critical problem is how to evaluate these different browsers. In previous work, evaluation is either absent or based on informal user feedback *e.g.* [2, 5]. Where objective data has been collected, user tasks and the questions asked vary widely, are often loosely defined, and final scores are therefore open to considerable interpretation. Most importantly, however, it is not current practice to compare overall meeting browser performance objectively.

In many other fields of research, an objective measure of system performance along with a standard corpus and set of reference tasks has been of enormous benefit in helping researchers compare techniques and make progress. For example, in the field of speech recognition, this has made possible the construction of real time, large vocabulary systems that would not have been feasible ten years ago. The text retrieval conference (TREC) has also used standard corpora, tasks and metrics with great success: average precision doubled from 20% to 40% in the last seven years.

Copyright is held by the author/owner(s).
CHI 2005, April 2–7, 2004, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

This work aims to develop similar metrics for meeting browsers, and describes a *browser evaluation test* (or BET) for meeting browsers.

> *We define the task of **browsing** a meeting recording 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 these *observations of interest*. The range of possibilities is enormous and depends upon meeting content and individual user interests. The BET aims to be:

a) an objective measure of browser effectiveness based on user performance rather than satisfaction;
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 allowing different researchers to evaluate their browsers and benchmark them.

This paper first presents an overview of the method, describes each of its significant features in detail, and illustrates results from a trial run of the BET.

OVERVIEW OF METHOD

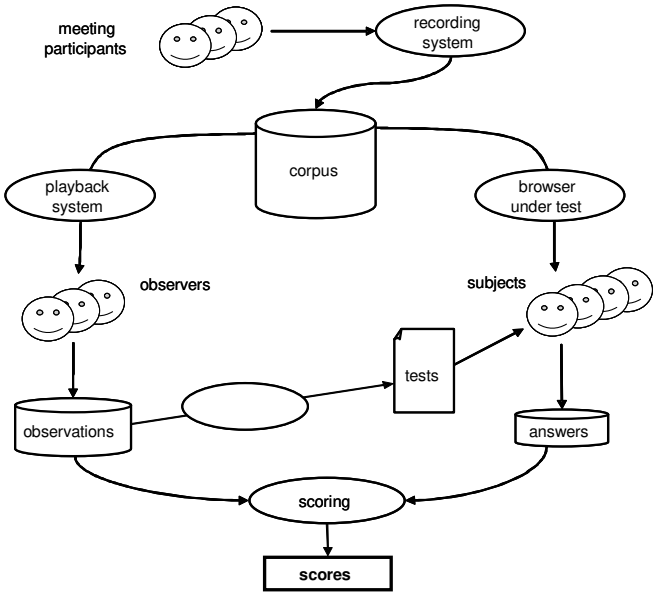

Figure 1. The BET method.

The BET method is illustrated in Figure 1. The significant features are described below, with further detail in subsequent sections:

- The *corpus* is a significant set of media recordings providing the data to be browsed.
- *Observers* watch selected meetings from the corpus, to produce a store of *observations*. Observers are not meeting participants.
- Later, during testing, the observations on some meeting are sampled to produce *tests*.
- *Subjects* use the *browser under test* to review the meeting, answering as many test questions as possible in a short time.
- *Answers* produced by the subjects are stored for scoring and analysis.
- *Scoring* compares the subjects' test answers to the original stored observations, to compute a *score* for the browser.

Using the BET requires one-time investment in creation of the corpus, collection of the observations and running of benchmark tests. Subsequent browser tests take advantage of this one-time work to run tests and produce comparable scores. The BET differs from classic usability testing because tasks are not predetermined by the experimenter, and the BET does not necessarily measure satisfaction.

THE CORPUS

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

Design of the corpus has enormous influence on the test. It determines the observations made, the questions asked, and ultimately the browsing behavior of the subjects. BET results obtained with the use of one corpus are therefore not directly comparable to results obtained with another corpus.

The recorded meeting used for the trial run was made in IDIAP's smart meeting room [6] by A. Lisowska as part of the IM2 project [3]. A 100 hour multi-media meeting corpus collection effort (now underway as part of the AMI project [1]) will provide additional meeting recordings for use in future applications of the BET.

THE OBSERVATIONS

Questions to be used in browser tests are determined by a set of observers, who produce *observations of interest*. Observers have available the full recordings from every media source, including slides. There is no time limit for the observers, but in the trial run, people spent about 4½ times the duration of the meeting to complete their observations. Each observer is instructed to produce observations that the meeting participants appear to consider interesting. This approach is meant to temper undue influence of each observer's own special interests, while avoiding the introduction of experimenter bias regarding the relative importance of particular meeting events.

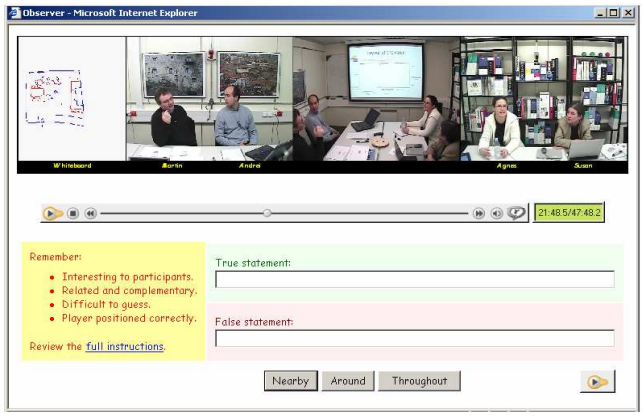

Figure 2. Observation input form.

Each observation is stated as a complementary pair of statements, one true and one false, both of which are later presented to subjects during testing. Observers are instructed to produce observations that should be difficult to guess without access to the recording (difficulty is verified later), and the observations should be simply and concisely stated.

The observer interface is shown Figure 2 above. Observers typically type their true statements first, into the upper text area. Each observation is time-stamped with the media time into the recording, and submitted with an estimate of its locality: *nearby*, *around* or *throughout*. As shown later in the paper, this is used to determine the temporal correspondence between questions and their answers.

Trial Run observations

In the trial run, we collected 294 observations from six observers about one 44-minute meeting, or roughly one observation per meeting-minute per observer. No attempt was made to filter the observations based on validity, as this would reintroduce experimenter's judgment, which the BET attempts to exclude.

A plot of observation density from the trial run (see Figure 3 below) shows the average number of observations made per observer within a one-minute window around each observation. The peaks in this graph identify parts of the meeting that can be interpreted as hot spots [4], where the most observations of interest occur in a short period. Automatic highlighting of these hot spots, should it be possible, could improve browser performance as defined in the introduction.

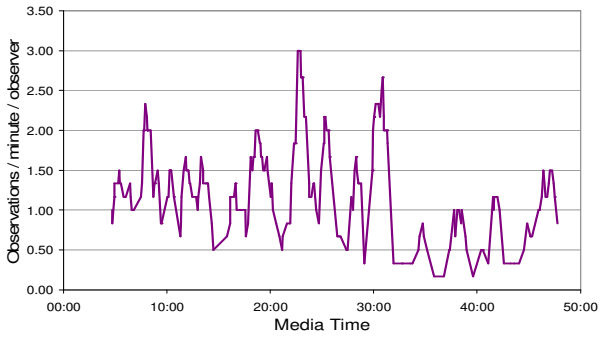

Figure 3. Observation density.

BROWSER TESTING

Test subjects are neither participants nor observers, and can take several tests, each of which requires them to use a browser to examine one of several meetings. 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 one at a time from the observations. Both the true and false statements of an observation pair are presented together in random order and the subject must use the meeting browser to decide which one is correct. Presenting subjects with both statements, rather than just one, gives them more information about what to look for in the meeting, and highlights the crucial facts necessary to determine the answer.

Figure 4. A BET question.

Questions are presented at the bottom of the screen in a window like that illustrated in Figure 4 above. When one of the statements is selected, the *OK* button is enabled, and when pressed, a new pair of statements is immediately presented.

Tests have a time limit of half the duration of the meeting under examination. This is to prevent a simple playback of the whole meeting to answer the questions, and time pressure is required in order to emphasize "the minimum time" stipulation from our definition of browsing. To help remind subjects of their time limit, a continuously running countdown timer is displayed above the *OK* button used to submit answers. Each answer is time-stamped with both the real time of the answer and the media position in the recording.

Trial Run tests

In the trial run, we tested a total of eleven women and thirteen men primarily from academia, whose average age was 35. All subjects were given 22 minutes to answer questions about the 44-minute trial run recording.

There were three test conditions: Guess, Base and F_1. In the Guess condition, subjects saw only the question window illustrated in Figure 4, but in the Base condition, they also had the media player used by the observers (shown in Figure 2). In the F_1 condition, subjects used the Ferret browser, as described below.

Eleven subjects were tested in the Base condition, ten in the F_1 condition, but only three in the Guess condition. Guessers worked so fast that they produced more than fifteen times more answers per subject than in the other conditions, with one subject even exhausting the question set. As a result, more subjects were tested in the Base and F_1 conditions so as not to magnify the imbalance in the number of answers.

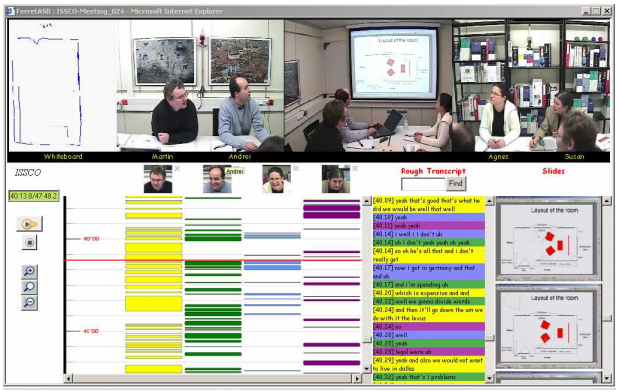

Figure 5. The F_1 condition.

Ferret browser

The experimental Ferret browser [8] can be configured with a range of possible features to assist navigation within a meeting recording. For the trial run, we tested ten subjects using a configuration of Ferret labeled as the F_1 condition, illustrated in Figure 5 above.

The top part of the F_1 screen is the same video and whiteboard player used by subjects in the Base condition. The bottom part of the screen, however, provides three additional navigation aids: speaker segmentations, a rough transcript generated by automatic speech recognition (with approximately 70% error rate), and captured presentation slides, all automatically generated from the meeting recording. Subjects can scroll, zoom, and click any of these elements to navigate in the recording.

RESULTS FROM TRIAL RUN

Figure 6 below shows the number of questions answered by each subject against the proportion answered correctly. Scores for the Guess condition, the Base condition, and F_1 show incrementally increasing accuracy, as expected.

The overall BET scores for each condition are a pair of numbers shown in Table 1: the speed of the browser (in answers per subject per minute), and its accuracy in percentage of questions answered correctly, together with unbiased standard deviations (σ).

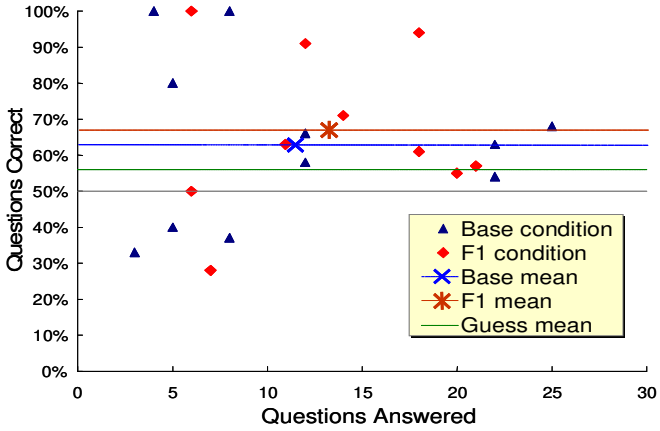

Figure 6. Speed versus accuracy.

Condition	Speed ± σ	Accuracy ± σ
Guess	9.2 ± 2.8	56.7% ± 2.43
Base	0.52 ± 0.36	63.5% ± 22.8
F_1	0.60 ± 0.26	67.7% ± 22.4

Table 1. Overall BET scores.

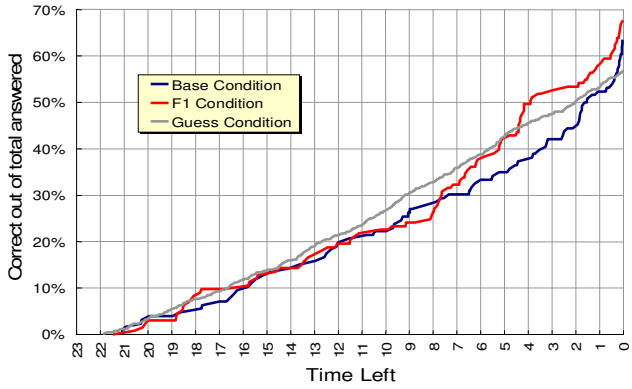

Figure 7. Score increase with time.

Scores over time

Figure 7 above shows how the average score increased over test time, culminating in the final scores of the table above. Although ultimately more accurate, the F_1 and Base conditions were lagging behind the Guess condition for most of the duration of the tests. The F_1 score increases significantly, as subjects become more familiar with the browser and the meeting itself. Both the F_1 and Base condition have last-minute spurts to achieve their final scores.

Media time difference

The time offsets between the subject's player, when the answer was submitted, and the observer's player, when the observation was originally made, are plotted in Figure 8 below. The histogram shows the number of correct and incorrect answers made, excluding *throughout* observations, during one-minute wide intervals centered on zero, for both the Base and F_1 conditions combined.

Answers made within ±30 seconds of the original observation are 93% correct, compared to just 66% overall. Clearly, helping users navigate to the correct point in the meeting helps them to answer more questions correctly.

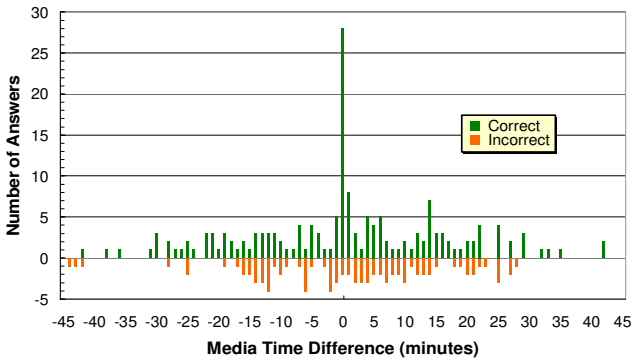

Figure 8. Correct and incorrect answers by media offset.

FUTURE WORK & CONCLUSIONS

Having completed this proof of concept of the BET, we are now extending it to larger corpora and to different styles of browser *e.g.* speech-only browsers, and browsers with access to manually created transcriptions or annotations.

Thanks to the experience gained from the trial run, subsequent applications of the BET will attempt to reduce the variance in scores between subjects and to improve the relevance of test questions. For example, we plan to capture the relative importance of observations (as rated by the observers) and present questions to all subjects in the same order of importance. We also plan to reduce subject variability, by testing all subjects in the Base condition, so that browser scores may be expressed as improvements over a common base.

To conclude, this work is helping to move beyond subjectively evaluated proof-of-concept demonstrations of meeting browsers towards more objective, independent, and repeatable evaluations. The ultimate goal of the BET is to help the research community improve future development of genuinely effective meeting browsers.

ACKNOWLEDGEMENTS

The authors wish to thank our meeting participants, observers and subjects, and our colleagues, in particular David Barber, Samy Bengio, Herve Bourlard, Marge Eldridge, Alison Evans, Paul Fenn, Daniel Gatica-Perez, Maël Guillemot, Peter Holdridge, Martin Karafiat, Agnes Lisowska, Iain McCowan, Andrei Popescu-Belis, Jo Schultz, and Andrew Stones. This work was supported by Swiss and EC projects IM2, M4, and AMI.

Note: An expanded description of this work is available as IDIAP-RR 04-53, and at http://mmm.idiap.ch/bet.

REFERENCES

1. AMI project http://www.amiproject.org.
2. Cutler, R. *et al*. Distributed Meetings: A Meeting Capture & Broadcasting System, *ACM Multimedia* '02.
3. IM2 project http://www.im2.ch.
4. Janin, A. *et. al*, The ICSI meeting project: Resources and research, in *Proc. of ICASSP 2004* Meeting Recognition Workshop.
5. Lee, D. *et. al*, Portable Meeting Recorder, In Proc. *ACM Multimedia 2002*.
6. Moore, D. The IDIAP Smart Meeting Room. IDIAP-COM 02-07, November 2002.
7. Tucker, S., Whittaker, S. Accessing Multimodal Meeting Data: Systems, Problems and Possibilities, In *Proc. of MLMI'04*, Springer-Verlag.
8. Wellner, P., Flynn, M., Guillemot, M. Browsing Recorded Meetings With Ferret, In *Proc. of MLMI'04*, Springer-Verlag.