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-04-53

SEPTEMBER 2004

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

The browser evaluation test (BET) is a method for assessing browser performance on meeting recordings. The number of *observations of interest* found in the minimum amount of time is used as the metric. Observations of interest are statements about the meeting collected by independent observers prior to performing an evaluation. When testing a browser, subjects are presented with questions drawn from the observations, enabling browsers to be assessed in terms of both speed and accuracy. This paper introduces the BET and applies it in a trial run. The resulting scores aim to be objective, independent, and repeatable.

Author Keywords

Meeting, Browser, Evaluation, Testing.

ACM Classification Keywords

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – evaluation/methodology; H.5.2 [Information Interfaces and Presentation]: User Interfaces – evaluation/methodology; H.1.2 [Models and Principles]: User/Machine Systems – human factors, human information processing

INTRODUCTION

Meetings are an integral part of our working lives, where important information is exchanged and decisions are made. Until recently, it was impossible to capture meeting information reliably, but developments in recording and storage techniques are now making this type of data readily available. While it is straightforward to play back such recordings, it is much more laborious for users to browse these recordings for elements of interest. The development of new technology to enhance browsing of recorded meetings has therefore become an active area of research. As surveyed in [26], different designs centered around documents, video frames, transcripts, topic analyses, and user artifacts such as slides [2,4,5,6,7,9,10,14,15,16,22].

Submitted September 2004

One critical problem is to determine how to evaluate these different browsers. Currently there is no standard evaluation procedure for meeting browsers. In some cases, evaluation is absent [2,4,5,7,9,15,16], in others it is based on informal user feedback [6], or focuses on a specific interface element (*e.g.* video key frames [10]). Where objective data has been collected by asking users to carry out tasks, these tasks are often not consistent across studies [13, 22]. In general, user tasks and the questions asked of users vary widely, are often loosely defined, and the final scores are therefore open to considerable interpretation. Most importantly, however, it is not currently possible to compare browsers and browsing techniques objectively.

In many other fields of research, an objective measure of system performance along with a standard corpus and set of reference tasks can be of enormous benefit in helping researchers compare techniques allowing the field to make progress. For example, in the field of speech recognition, the use of standardized tasks, metrics and corpora has made possible the construction of real time, large vocabulary systems that would not have been feasible ten years ago [18,24]. And the text retrieval conference (TREC) has also used standard corpora, tasks and metrics with great success: with average precision doubling from 20% to 40% in the last seven years [27,28]. The aim of this work is to develop equivalent metrics for meeting browsers.

In this paper, we discuss a *browser evaluation test* (or BET) for meeting browsers, originally proposed in [7]. There is considerable breadth in what it means to browse a meeting, and in usage scenarios for meeting browsers. For example, the distinction between searching and browsing is not always clear. We consider search for specific events as a part of browsing, but browsing also includes the rapid assimilation of a meeting overview, and the ability to quickly skim through a meeting to find unexpected points of interest. One of the challenges in designing a good browser evaluation test is to create a task that takes into account these multiple dimensions of browsing.

> *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 method identifies observations of interest based on the impressions of ordinary people. It does not reflect the particular interests of the experimenter or browser designer.

We aim to make the BET:

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

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, and then describes each of its significant features in detail, illustrated by results from a trial run of the BET.

OVERVIEW OF METHOD

The BET objectively measures how well a browser satisfies the goal of finding the most observations of interest in the minimum time, using two groups of people: observers and subjects. Observers have no stake in any particular browser, nor any bias about what is interesting in meetings – unlike the experimenters or browser designers themselves. Several observers watch meeting recordings and produce a set of 'observations'. Subsequently, for each browser under test, a fresh set of subjects is presented with questions based on the observations. Ultimately, their answers determine a score for the browser.

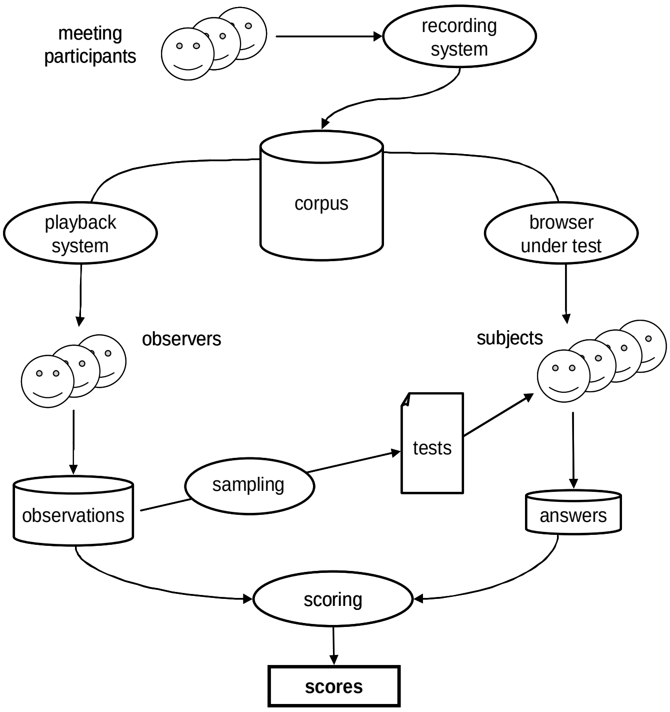


Figure 1. The BET method.

The BET method is summarized in Figure 1 above. 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*.
- 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 they can 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 considerable investment in one-time creation of the corpus and collection of the observations. In order to understand final browser scores, it is also necessary to run benchmark tests for well-known conditions and publish them along with the corpus and observations. This need only be done once, however. Subsequent browser tests take advantage of this one-time work to run tests and produce comparable scores, repeatedly.

Further detail on each of these points is provided in the remainder of the paper. Within each of the following four sections (concerning the corpus, observations, testing and results), the BET method is explained in detail, and then a subsection illustrates its application in a trial run, showing how to construct benchmark tests and scores for a sample meeting browser.

THE CORPUS

The corpus is a set of media recordings consisting of 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 behavior of the subjects.

BET results obtained with the use of one corpus are therefore not directly comparable to results obtained with another corpus. This implies that a shared corpus must be available to anyone performing comparable BETs, 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. For our purposes, the corpus must 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 must be 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 organizational issue is preferable to discussing the mathematics behind a new algorithm.

Trial Run Corpus

The recorded meeting used for the trial run was a 44-minute[1] discussion between four people on how to select and lay out furniture in a university reading room. This recording was made in IDIAP's smart meeting room [19] by A. Lisowska as part of her work in the IM2 project [12,17]. It is available for viewing (along with all other data discussed in this paper) at the BET web site [3]. A large multi-media meeting corpus collection effort (now underway as part of the AMI project [1,19]) will provide additional meeting recordings for use in future applications of the BET.

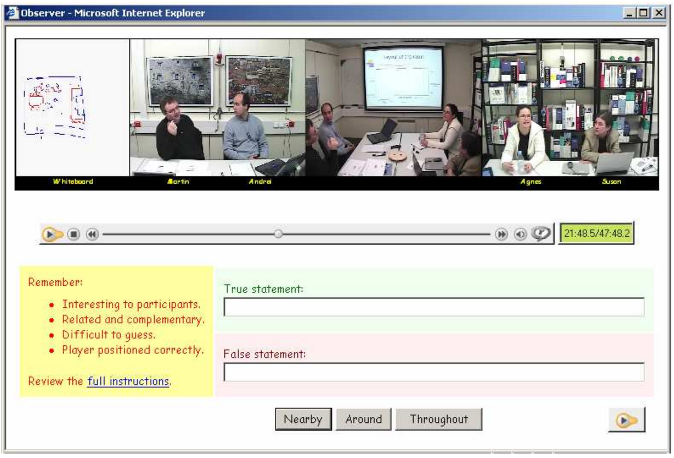
THE OBSERVATIONS

Questions to be used in browser tests are determined by a set of observers, who produce *observations of interest*.

The observers independently (*i.e.* alone) watch selected meetings from the corpus. Observers have available the full recordings from every media source, in parallel, including paper printouts of the slides accompanying the meeting. They may rewind and replay the sources, as they desire. 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.

Instructions are given in a standard manner on a web page made available with the corpus. Each observer is instructed to produce observations that the meeting participants appear to consider interesting. Asking observers to take the perspective of participants is meant to temper undue influence of each observer's own special interests (*e.g.* someone who finds gesticulation more significant than issues discussed). A single observer does not typically make the same observation multiple times, but the most significant features of each meeting are observed multiple times by different people, albeit in slightly 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 selected observations reflects their relative frequency within the meeting.

This approach avoids the introduction of experimenter bias regarding the relative importance of particular meeting events. Instead of looking for pre-determined general categories of events considered to be significant (*e.g.* agreement, disagreement, action items, *etc.*) we sample from the specific details selected by our independent observers within each particular meeting.



Remember:

- Interesting to participants.
- Related and complementary.
- Difficult to guess.
- Player positioned correctly.

Review the full instructions.

True statement:

False statement:

Figure 2. Observer web 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 not be easy to guess without use of a browser (difficulty is verified later), and the observations should be simply and concisely stated. To encourage brevity, observations are collected via a web form (see Figure 2 above), where the box for the observation text is small.

Observers typically type their true statements first in the upper text area. As soon as they begin typing, the media player is paused so that its position can be recorded along with the observation. To encourage consistency between the two complementary statements, the first statement is automatically copied into the other text field for editing before submission.

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. The observer associated with each observation is recorded, and each observer is given a questionnaire, recording personal and professional details, so that these variables are available to be analyzed for possible influences on the score. (Later, subjects are given a similar questionnaire.)

Trial Run Observations

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

[1] Actually a 44-minute segment from a 47-minute recording.

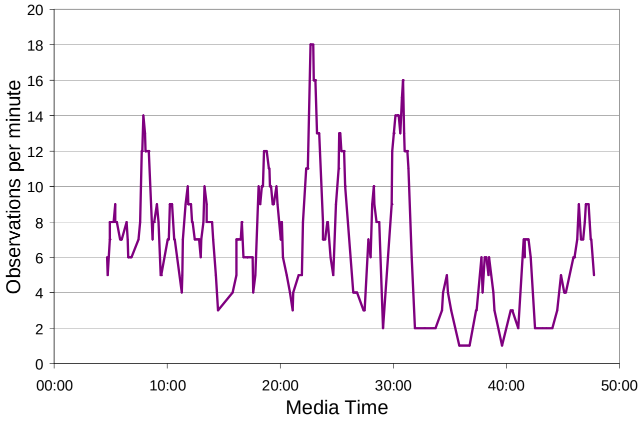


Figure 3. Observation density.

A plot of observation density from the trial run (see Figure 3 above) shows the total number of observations made by all observers within one minute on either side of each observation.

The peaks in this graph identify parts of the meeting that can be interpreted as "hot spots," where the most observations of interest occur in a short period of time. Casual inspection of observations in these peaks shows considerable overlap, *i.e.* most mention the same participant making a point about the same topic. Defining meeting "hot spots" in this way is an alternative method to that used by ICSI [13] but should help browsing performance as defined above, *i.e.* to help people find the maximum observations of interest in the minimum amount of time.

Observations cannot only be used for browser testing, but they can also be used for meeting analysis, and for the development of better browsers in the future. One promising direction, for example, is automatic detection of meeting hot spots using machine-learning techniques such as those proposed in [19]. More hints on useful browser features can be gained by characterizing the observations, especially during peak times. One striking attribute, for example, is that most observations are about individual participants, rather than about overall group actions. Of the top ten most frequently used words in the trial run observations, four were the participants' names, while the remainder were insignificant (*the, of, to, a, is* and *that)*. The name of at least one participant occurred in 81% of our trial run observations (238 out of 294).

These observations support the intuition that spoken words in the audio channel are more important to browser performance than information in the video channels, because most of the observations are about what participants said. For example, the words "want" and "says" appear in the top twenty most frequently used words. This encourages further work on meeting browsers that support navigation through speaker segmentation and speech transcription.

BROWSER TESTING

Test 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 is not critical.

Subjects can 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.

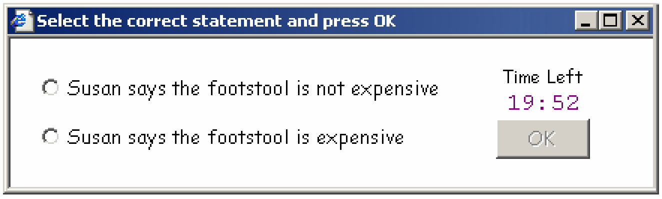
Each test is a set of questions drawn one at a time from the observations. Both the true and false statements of an observation pair are presented together in random order and the subject must use the meeting browser to decide which one is correct. Presenting subjects with both statements, rather than just one, gives them more information about what to look for in the meeting, and highlights the crucial facts necessary to determine the answer.

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

Tests have a time limit of half the duration of the meeting under examination. This is partly to simplify scheduling of subjects, but also to prevent a simple playback of the whole meeting from satisfying the questions. 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.

Observations are selected randomly for each test, but no observation pair is used more than once in the same test. In order to avoid a ceiling effect, the number of questions in a test is practically unlimited.

This testing process is entirely automatic with tests administered via the web. This simplifies use of many possibly unknown subjects, but does not imply that any browser must itself be web-based. The media files may be a local copy to maximize playback performance.


Select the correct statement and press OK

○ Susan says the footstool is not expensive

○ Susan says the footstool is expensive

Time Left
19:52

Figure 4. A BET question.

Discussion of testing options

An alternative approach to presenting questions one at a time is to show each subject a large set of questions *all at once*, and ask them to answer as many as possible within the time limit. This approach can be argued to assess the browsing task (rather than just searching) more accurately and may better reflect the scenario where a person is trying to learn as much as possible about a meeting in a short time, rather than an attempt to find one particular fact. However, there are practical problems with this approach (*e.g.* how many questions to present, and finding screen real-estate for hundreds of questions). It also has the disadvantage that it may encourage too much guessing, and that results will vary due to “exam technique” or the ability of subjects to cherry-pick the easiest questions.

However, sequential presentation does indeed test a blend of both searching and browsing effectiveness – not pure searching. During the later part of the test, subjects have already browsed through large parts of the meeting (up to half). Later questions become progressively easier to answer based on the relevance of material viewed while looking for answers to previous questions. This position is at least partially supported by results of the trial run (presented in the next section) which clearly show an increase of speed and accuracy in the later parts of the tests.

Benchmark tests

Published along with each BET corpus and observation set are also two benchmark scores. These are from two one-time tests that are performed using each of the following conditions:

- *Guess condition*: educated guesses with no media present whatsoever;
- *Base condition*: the same basic playback software used by the observers.

The Guess condition reveals whether observations are too easy to guess, and it provides a lower bound below which no browser should sink, no matter how constrained.

The Base condition provides another useful reference point because we know that all information the observers used was available through this interface, but the observers had unlimited time while the benchmark base test is limited to just half the recording time. A severely restricted browser (*e.g.* video only, without audio) could score lower than the benchmark base, but we would expect most browser designers to consider the Base condition score as a minimum starting point.

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. In the Guess condition, they saw only the question window illustrated in Figure 4, but in the Base condition they also had the media player used by the observers (in Figure 2 above), but without the true and false type-in fields. Eleven subjects were tested in the Base condition, and three subjects were tested in the Guess condition. Guessers worked so fast that they produced more than fifteen times more answers per subject in the Guess condition than in the Base condition, and one subject exhausted the question set. As a result, more subjects were tested in the Base condition so as not to magnify the imbalance in number of answers

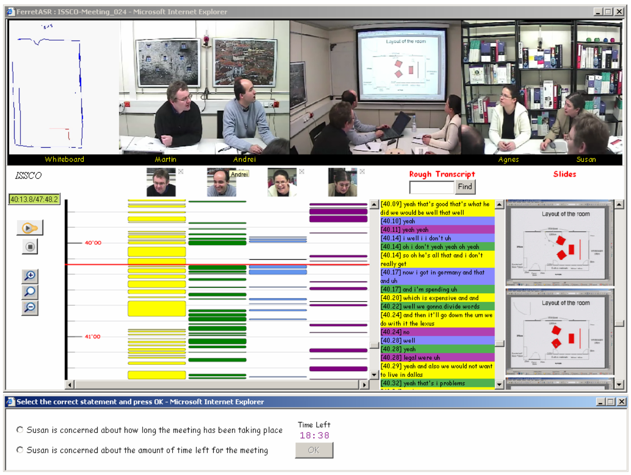
Ferret browser

The experimental Ferret browser [30] 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 below.

The top part of the F_1 screen is the same video and whiteboard player used by observers and the 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 (ASR), and captured presentation slides, all automatically generated from the meeting recording.

1) The speaker segmentation is presented on a scrollable and zoomable timeline displaying a colored column for each participant whenever that person is detected as speaking. A red horizontal cursor moves along this timeline as the media advances, and users can drag this cursor to control playback position, as well as click on any segment to play it.

2) A very rough ASR transcript generated by Karafiat using the M4 recognition system [29] is colored by participant, but has more than a 70% word error rate. At the top of this column is a text field and “Find” button for searching specific words in the transcript. The user can click on text fragments to move playback to the corresponding point in the meeting.



[40.09] yeah that's good that's what he did we would be well that well
[40.10] yeah
[40.11] yeah yeah
[40.14] i well i i don't uh
[40.14] oh i don't yeah yeah oh yeah
[40.14] so oh he's all that and i don't really get
[40.17] now i got in germany and that and uh
[40.17] and i'm spending uh
[40.20] which is expensive and and
[40.22] well we gonna divide words
[40.24] and then it'll go down the um we do with it the lexus
[40.24] no
[40.28] well
[40.28] yeah
[40.28] legal were uh
[40.29] yeah and also we would not want to live in dallas
[40.32] yeah that's i problems

Select the correct statement and press OK - Microsoft Internet Explorer

○ Susan is concerned about how long the meeting has been taking place

○ Susan is concerned about the amount of time left for the meeting

Time Left
18:38

OK

Figure 5. The F_1 condition.

3) Every slide change that occurred during the meeting is captured and displayed in the right column. Subjects can click on these images to navigate the player to the point in the meeting when that slide was first displayed.

F_1 was tested on people drawn from the same subject pool as the observers and benchmark conditions, primarily at the University of Sheffield. Browser software and media playback was running locally, submitting results to servers at IDIAP. To prevent the possibility of communication lags or browser crashes from invalidating a testing session, the countdown timer resumes from the point of last submission.

RESULTS FROM TRIAL RUN

Scores from testing the two benchmark conditions and F_1 are presented first as raw scores, followed by three illustrative graphs, and ending with an overall BET score pair for each of the three conditions.

Raw scores

The score for each subject test is simply the proportion of correct answers obtained. A perfect score for a test would therefore be 100%, while random answers would yield a score of around 50%.

Subject	Answers	Correct	Incorrect	Score
A1	255	142	113	55.7%
A2	220	123	97	55.9%
A3	135	81	54	60.0%
Guess Total	*610*	*346*	*264*	*56.7%*

Table 1. Scores for the Guess condition.

Scores for subjects in the Guess condition are summarized in Table 1 above. The three subjects scored an average of 56.7% correct answers. This is consistent with expectations, showing that observations were not too easy to guess correctly. The subjects answered very different numbers of questions – one of the subjects completed all the questions in the database. However, the slowest subject achieved the highest score.

Subject	Answers	Correct	Incorrect	Score
B1	22	14	8	63%
B2	25	17	8	68%
B3	12	7	5	58%
B4	8	8	0	100%
B5	5	2	3	40%
B6	3	1	2	33%
B7	12	8	4	66%
B8	5	4	1	80%
B9	8	3	5	37%
B10	22	12	10	54%
B11	4	4	0	100%
Base Total	*126*	*80*	*46*	*63.5%*

Table 2. Scores for the Base condition.

Scores for subjects in the Base condition are shown in Table 2 above. Of the eleven subjects in this condition, two scored 100%, but one of them with double the questions of the other. Once again, there is greater accuracy at slower speeds. The average score in this condition was 63.5% – somewhat higher than the Guess condition, but with only a fifth of answers. Surprisingly, three subjects scored less than random, despite watching significant portions of the meeting.

Subject	Answers	Correct	Incorrect	Score
C1	20	11	9	55%
C2	6	3	3	50%
C3	18	17	1	94%
C4	21	12	9	57%
C5	18	11	7	61%
C6	11	7	4	63%
C7	6	6	0	100%
C8	14	10	4	71%
C9	12	11	1	91%
C10	7	2	5	28%
F_1 Total	*133*	*90*	*43*	*67.7%*

Table 3. Scores for the F_1 condition.

Scores for subjects in the F_1 condition are shown in Table 3 above. The ten subjects in this condition achieved a score 67.7%. This is larger than the Base condition, and with a slightly larger number of questions answered in the same time (13.3 questions per subject, versus 11.5 for the Base condition).

Scores over time

There are several times associated with each answer: the real time is recorded, along with the test time remaining for the subject (these are not necessarily directly related, if, for example, a subject needs to switch machines during the test) and the position in the media.

Figure 6 below shows how the average score increases over test time for each condition. The final resting place of the score is that shown in the basic result tables above, with the F_1 condition ahead of the Base condition, ahead of the Guess condition. However, it is interesting to note that both the F_1 and Base conditions were lagging behind the Guess condition for most of the duration of the tests. The gradient of the F_1 score increases significantly with around eight minutes of the test remaining – as subjects become more familiar with either the browser and the meeting itself.

Both the F_1 and Base condition have a final spurt in the last thirty seconds of the test. Intuition and anecdotal evidence suggests that subjects notice their dwindling time remaining, abandon use of the browsers, and simply try to answer as many questions as they can in the final seconds. However, it is interesting to note the high accuracy of these final answers, compared to the earlier answers and the pure Guess condition. This suggests that subjects have learnt much about the meeting content, incidentally, during the test.

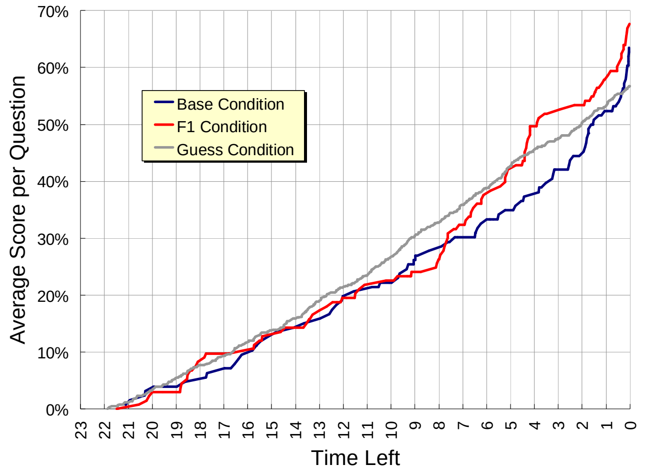


Figure 6. Score increase with time.

Media time difference

The difference between the media time of the observer's player when the observation was made, and the media time of the subject's player when the answer was submitted is plotted as the *media time difference* in Figure 7 below. On the left side of the graph are answers made before their corresponding questions, while answers made later are shown on the right. Correct answers are counted above the axis, and incorrect answers below.

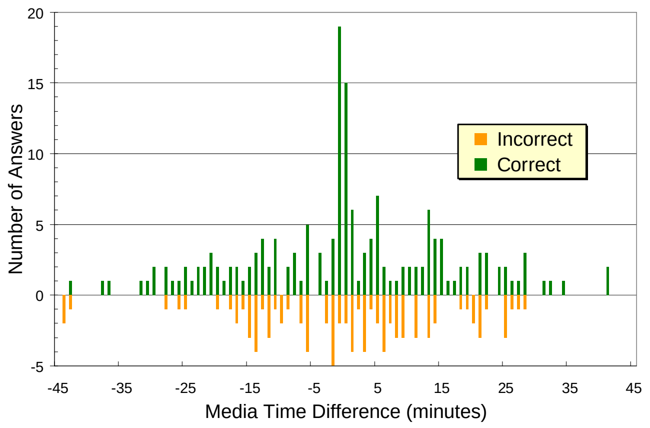


Figure 7. Correct and incorrect answers by media offset.

It can be seen that subjects make many more correct answers (89%) within one minute either side of the original observation[2] – compared to the overall proportion of correct answers (66%). This holds within two percentage points for both the Base and F_1 conditions. The obvious conclusion is that helping users navigate to the correct point in the meeting clearly helps them to answer questions correctly.

Speed versus accuracy

Figure 8 below shows a graph of the number of questions answered by each subject against the proportion answered correctly for the Base and F_1 conditions. Horizontal lines for the Guess condition, the Base condition, and F1 show a progression in accuracy, as expected. The mean values for the two browser conditions are marked, together with one standard deviation on either side on each axis. This shows that F_1 is both faster and more accurate than the Base condition.

It is also evident that both the most accurate and least accurate subjects were amongst the slowest. This suggests that slower subjects were either more diligent, or were presented with more difficult questions. As speed increases, the Base and F_1 subjects tend to become only as accurate as those in the Guess condition. This may be because the browser leads subjects to inappropriate conclusions under pressure, or simply that quickly decided answers degenerate towards guesses.

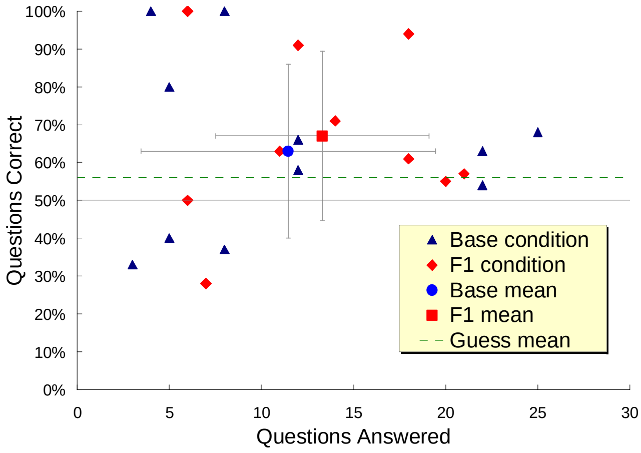


Figure 8. Speed versus accuracy.

The F_1 condition appears to dominate in the mid-range of speeds, where accuracy is highest, with neither a high nor a low speed. The Base condition appears to be either slow or fast, with slower subjects achieving amongst the highest and lowest accuracy, while quicker subjects achieve no better than the Guess condition.

The overall BET scores for each condition are a pair of numbers: one representing the speed of the browser in answers per minute, and the other representing its accuracy, as shown in Table 4 below.

Condition	Speed	Accuracy
Guess	27.7	56.7%
Base	5.7	63.5%
F1	6.0	67.7%

Table 4. BET scores.

[2] Note that a random distribution of questions and answers would naturally yield a simple triangular profile. The small peaks at the extremes are due to answers supplied at the start or end of the recording, concerning the other end of the recording, since the player 'wraps' around at the endpoints.

STATISTICS

This section illustrates the effort required to create a set of corpus observations, the effort to test a single browser, and the statistical significance of the results.

Table 4 below presents figures for observation collection in the trial run, along with projected figures for the planned application of the BET to part of the new AMI corpus. The effort of observation collection is spent only once, since the observations may be used repeatedly to test many browsers.

Observation Collection	**Trial Run**	***Plan***
Number of meetings	1	*10*
Average duration / meeting	44 mins	*40 mins*
Total length of recordings	44 mins	*7 hrs*
Observers / meeting	6	*6*
Meetings observed / observer	1	*2*
Observers	6	*30*
Observation time / observer	3½ hrs	*6 hrs*
Total observation time	20 hrs	*180 hrs*
Observations / observer-hour	14.8	*14.8*
Total observations	294	*2,667*
Observations / meeting	294	*267*
Average test duration	22 mins	*20 mins*
Questions / subject-hour	31	*31*
Answers / test	11.5	*10.4*

Table 5. Observation collection statistics.

Table 5 below presents figures for testing an individual browser. The Base and F_1 condition of the trial run are listed separately, together with the expected figures for testing a browser against the AMI observations. We expect that each browser condition will take more effort to asses, but the reward lies in a tighter confidence interval width[3].

Browser Testing	**Trial Run**		***Plan***
	Base	$\mathbf{F_1}$	X_n
Subjects / meeting	11	10	*5*
Meetings / subject	1	1	*2*
Subject time / subject	22 mins	22 mins	*40 mins*
Subjects	11	10	*25*
Number of tests	11	10	*50*
Total subject time	4 hrs	4 hrs	*17 hrs*
Answers / subject	11.5	13.3	*20.9*
Answers	126	133	*522*
BET score	63%	68%	*68%*
Confidence level	95%	95%	*95%*
Lower confidence limit	54.0%	58.6%	*62.7%*
Upper confidence limit	73.0%	76.7%	*72.1%*
Confidence int. width	19.0%	18.0%	*9.4%*

Table 6. Browser testing statistics.

FUTURE WORK

Having carried out this proof of concept demonstration of the BET technique we now plan to extend to larger corpora and to different styles of browser. The AMI project intends to record 100 hours of meetings, which will serve as a corpus for a larger data set, and extend the set of observations available for experiments.

We also plan to use BET to compare different styles of browser, *e.g.* speech only browsers, or browsers with no video stream. Another possibility is to use BET to determine the effect of various quality parameters for the different UI components. For example, we may investigate the effects of ASR quality, or the quality of speaker detection on browsing performance. These comparisons suggest how BET results can be used to improve future system designs. By comparing the BET scores of multiple system designs, we can look at how browsing is affected by various UI components (*e.g.* video, access to slides), as well as quality parameters (*e.g.* ASR, speaker detection). We can then use this information to inform which components are most important for new browsers (*e.g.* video may be unimportant compared with transcribed speech), and which UI components need most improvement (e.*g.* ASR quality). We might also correlate BET scores with logged user behaviors in order to determine whether use of a particular UI feature improved BET browsing scores, again suggesting directions for future designs. Finally, we want to investigate the relationship between BET scores and the subjective evaluations used in many previous studies.

SUMMARY & CONCLUSION

The browser evaluation test is a method for assessing browser performance on meeting recordings in which the number of *observations of interest* found in the minimum amount of time is used as the metric. Observations of interest are statements about the meeting collected on a meeting corpus by independent observers prior to performing an evaluation. When testing a browser, subjects are presented with questions drawn from the observations, enabling browsers to be scored in terms of both speed and accuracy. This paper introduced the BET and applied it in a trial run.

To conclude, this work aims to help us move beyond pure proof-of-concept technology demonstrations of meeting browsers towards more objective, independent, and repeatable evaluations. The ultimate aim of the BET is to help strengthen the future development of genuinely effective browser technology.

ACKNOWLEDGEMENTS

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

[3] Confidence limits are calculated assuming that answers are independent of one another. In reality, a subject's answers are not independent, especially as a test progresses – they may be asked similar questions more than once, while familiarity with the meeting and browser increase over time.

REFERENCES

1. AMI project http://www.amiproject.org.
2. Bett, M., Gross, R., Yu, H., Zhu, X., Pan, Y., Yang, J., Waibel, A.: Multimodal Meeting Tracker. In: *Proc.* of RIAO, Paris, France (2000)
3. Browser Evaluation Test http://mmm.idiap.ch/bet.
4. Chiu, P., Boreczky, J., Girgensohn, A., Kimber, D.: LiteMinutes: An Internet-Based System For Multimedia Meeting Minutes. In: *Proc.* of 10th WWW Conference, Hong Kong (2001) 140-149
5. Colbath, S., Kubala, F., Liu, D., Srivastava, A.: Spoken Documents: Creating Searchable Archives From Continuous Audio. In: *Proc.* of 33rd Hawaii International Conference On System Sciences, (2000)
6. R. Cutler, Y. Rui, A. Gupta, JJ Cadiz, I. Tashev, Li-wei He, A. Colburn, Zhengyou Zhang, Z. Liu, S. Silverberg. “Distributed Meetings: A Meeting Capture and Broadcasting System”, In *Proc. ACM Multimedia 2002*.
7. Flynn, M., Wellner, P., In Search of a good BET: A proposal for a Browser Evaluation Test, IDIAP-COM 03-11, September 2003, http://www.idiap.ch/.
8. Foote, J., Boreczky, G., Wilcox, L.: An Intelligent Media Browser Using Automatic Multimodal Analysis. In: *Proc. of ACM Multimedia*, Bristol, UK (1998).
9. Geyer, W., Richter, H., Fuchs, L., Frauenhofer, T., Daijavad, S., Poltrock, S.: A Team Collaboration Space Supporting Capture And Access Of Virtual Meetings. In: *Proc. of* 2001 International ACM SIGGROUP Conference On Supporting Group Work, Boulder, Colorado (2001) 188-196
10. A. Girgensohn, J. Boreczky, L. Wilcox, “Keyframe-Based User Interfaces for Digital Video”, In *IEEE Computer*, Vol. 34, No. 9, pp. 61-67, September 2001.
11. M. Guillemot, P. Wellner, D. Gatica-Pérez & J-M. Odobez, “A Hierarchical Keyframe User Interface for Browsing Video over the Internet”, In *Proc. of the 9th IFIP International Conference on Human-Computer Interaction INTERACT 2003*, ETHZ, Zurich, 2003.
12. IM2 project http://www.im2.ch.
13. A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Marcias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, The ICSI meeting project: Resources and research, *in* ICASSP 2004 Meeting Recognition Workshop.
14. Kazman, R., Al-Halimi, R., Hunt, W., Mantei, M.: Four Paradigms for Indexing Video Conferences. IEEE Multimedia 3(1) (1996) 63-73
15. Lalanne, D., Sire, S., Ingold, R., Behera, A., Mekhaldi, D., Rotz, D.: A Research Agenda For Assessing The Utility Of Document Annotations In Multimedia Databases Of Meeting Recordings. In: *Proc. of* 3rd International Workshop on Multimedia Data And Document Engineering, Berlin, Germany (2003)
16. D. Lee, B. Erol, J. Graham, J. Hull and N. Murata, “Portable Meeting Recorder”, In *Proc. ACM Multimedia 2002*, pp. 493-502.
17. A. Lisowska, M. Rajman, and T. Bui (2004), Archivus: A system for accessing the content of recorded multimodal meetings, *Proc. of* MLMI '04.
18. Marcus, M. (1991) *Proc. of* Speech and Natural Language Workshop, San Francisco, Kaufmann.
19. I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling Human Interaction in Meetings. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, April 2003.
20. Meeting recording corpus http://mmm.idiap.ch.
21. D. Moore. *The IDIAP Smart Meeting Room*. IDIAP-COM 02-07, November 2002.
22. Moran, T P., Palen, L., Harrison, S., Chiu, P., Kimber, D., Minneman, S., Melle, W., Zellweger, P.: "I'll get that off the audio": A Case study of salvaging multimedia meeting records. In: *Proc. of* CHI '97, Atlanta, Georgia (1997).
23. C. W. Ng, M. R. Lyu, “ADVISE: Advanced Digital Video Information Segmentation Engine”, In *Proc. of the 11th International World Wide Web Conference 2002*, Honolulu, Hawaii, USA.
24. Price, P. (1991) *Proc. of* Speech and Natural Language Workshop, San Francisco, CA., Kaufmann.
25. B. L. Tseng, C. Lin and J. R. Smith, “Video Summarization and Personalization for Pervasive Mobile Devices”, In *Proc. SPIE 2001*, vol. 4676, 2001.
26. Tucker, S., Whittaker, S. (2004), Accessing Multimodal Meeting Data: Systems, Problems and Possibilities, MLMI04 Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Martigny, Switzerland, 21-23 June 2004.
27. Text REtrieval Conference http://trec.nist.gov
28. Voorhees, E. M., & Harman, D. K., (1998). Overview of the seventh Text Retrieval Conference (TREC-7), in Voorhees, E. M., & Harman, D. K. (Eds.), *Proc. of* the Seventh Text Retrieval Conference (TREC-7), 1998.
29. V. Wan, M.Karafiát and S. Renals, Speech Recognition on M4. Workshop on Multimodal Interaction and Related Machine Learning Algorithms. 21 - 23 June 2004, Martigny, Switzerland.
30. P. Wellner, M. Flynn, M. Guillemot, Browsing Recorded Meetings With Ferret, MLMI04 Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Martigny, Switzerland, 21-23 June 2004.