Extrinsic Summarization Evaluation: A Decision Audit Task

GABRIEL MURRAY
University of British Columbia
THOMAS KLEINBAUER, PETER POLLER and TILMAN BECKER
German Research Center for Artificial Intelligence (DFKI)
and
STEVE RENALS and JONATHAN KILGOUR
University of Edinburgh

In this work we describe a large-scale extrinsic evaluation of automatic speech summarization technologies for meeting speech. The particular task is a decision audit, wherein a user must satisfy a complex information need, navigating several meetings in order to gain an understanding of how and why a given decision was made. We compare the usefulness of extractive and abstractive technologies in satisfying this information need, and assess the impact of automatic speech recognition (ASR) errors on user performance. We employ several evaluation methods for participant performance, including post-questionnaire data, human subjective and objective judgments, and a detailed analysis of participant browsing behavior. We find that while ASR errors affect user satisfaction on an information retrieval task, users can adapt their browsing behavior to complete the task satisfactorily. Results also indicate that users consider extractive summaries to be intuitive and useful tools for browsing multimodal meeting data. We discuss areas in which automatic summarization techniques can be improved in comparison with gold-standard meeting abstracts.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Evaluation/methodology*

General Terms: Algorithms, Measurement, Human Factors

Additional Key Words and Phrases: Summarization, extraction, abstraction, evaluation, browsing, interfaces

ACM Reference Format:

Murray, G., Kleinbauer, T., Poller, P., Becker, T., Renals, S., and Kilgour, J. 2009. Extrinsic summarization evaluation: A decision audit task. ACM Trans. Speech Lang. Process., 6, 2, Article 2 (October 2009), 29 pages.
DOI = 10.1145/1596517.1596518 http://doi.acm.org/10.1145/1596517.1596518

This work is partially supported by the European IST Programme Project FP6-033812 (AMIDA).
Author's address: G. Murray; email: gabrielm@cs.ubc.ca.

DOI 10.1145/1596517.1596518 http://doi.acm.org/10.1145/1596517.1596518

1. INTRODUCTION

Automatically-generated summaries may be evaluated according to *intrinsic* criteria, which directly relate to the quality of summarization, or *extrinsic* criteria, which are concerned with the function of the system in which the summaries are used [Jones and Galliers 1995]. Intrinsic measures for summarization, such as ROUGE [Lin and Hovy 2003], evaluate the performance of the summarization system in terms of how well the information content of an automatic summary matches the information content of multiple human-authored summaries. Such intrinsic measures are invaluable for development purposes, and possess the advantages of being easy to reproduce and automatic to run. In contrast, extrinsic methods usually require human subjects to perform a task using different forms of summaries. Extrinsic evaluations are more expensive, since in addition to the human effort required to perform the extrinsic task, the evaluation of task performance is often subjective. However, as Spärck Jones [1999] wrote, "it is impossible to evaluate summaries properly without knowing what they are for".

This article concerns an extrinsic evaluation of automatic speech summarization in the domain of multiparty meetings. The research was carried out in the context of the AMI and AMIDA projects,[1] whose goal is the support and analysis of multi-modal interactions between people in meetings. We constructed a number of automatic and semi-automatic summarization systems for this domain, employing both extractive and abstractive approaches, and using human and automatic speech transcriptions. The recorded meetings that we investigated took the form of several series of design meetings: within this domain we designed an extrinsic task that modelled a real-world information need. Using a number of experimental conditions, corresponding to the summarization systems, we enlisted subjects to participate in the task. The extrinsic evaluation of the summarization systems was based on this task, using a number of measures to evaluate how well the task was accomplished in the various conditions.

The chosen task was a *decision audit*, wherein a user had to review archived design team meetings in order to determine how a given decision was reached by the team. This involved the user determining the final decision, the alternatives that were previously proposed, and the arguments for and against the various proposals. This task was chosen because it represents one of the key applications for analyzing the interactions of teams in meetings: that of aiding *corporate memory*, the storage and management of a organization's knowledge, transactions, decisions, and plans. An organization may find itself in the position of needing to review or explain how it came to a particular position or why it took a certain course of action. We hypothesize that this task will be made much more efficient if multimodal meeting recordings—and the means to browse the recordings—are available, along with their summaries.

There are many real life examples that demonstrate the value of being able to conduct a decision audit. When the Scottish Parliament opened three years

[1] http://www.amiproject.org

later than scheduled in Edinburgh in 2004, its cost had exceeded initial estimates by at least ten times. Being able to audit how the early estimates were determined and why the construction timeline was overly optimistic would be useful not only to those involved in the design and construction of the building complex, but to outraged taxpayers demanding increased transparency on such matters. As a second example, the delivery of new Airbus A380 passenger jets was delayed significantly because of faulty wiring and configuration management issues between various European factories. The delays caused executive turnover at Airbus, but a decision audit on how the initial wiring plans were agreed upon and why Airbus locations were not all using identical software could have lead to more targeted accountability within the organization. In both cases, the vital information would be spread across multiple meetings, multiple parties, and multiple locations. It is the ability to browse and locate such widely distributed data that we are evaluating in this novel extrinsic task.

The decision audit represents a complex information need that cannot be satisfied with a simple one-sentence answer. Relevant information will be spread across several meetings and may appear at multiple points in a single discussion thread. Because the decision audit does not only involve knowing *what* decision was made but also determining *why* the decision was made, the person conducting the audit will need to understand the evolution of the meeting participants' thinking and the range of factors that led to the ultimate decision. For a particular decision audit task, the decision itself may be a given. Because those conducting the decision audit do not know which meetings are relevant to the given topic, there is an inherent relevance assessment task built into this overall task. Their time is limited and they cannot hope to scan the meetings in their entirety, and so must focus on which meetings and meeting sections seem most promising.

The structure of this article is as follows. In Section 2 we give an overview of relevant research in speech summarization, summarization evaluation, and browser evaluation. In Section 3 we describe the AMI meeting corpus used for these experiments. In Section 4 we provide the experimental setup, describing the decision audit task in detail, introducing the summary conditions evaluated, and presenting a three-dimensional evaluation scheme based on user feedback, annotator ratings, and browsing behavior. In Section 5 we present the results of these evaluations and discuss their ramifications. We show that automatically generated summaries outperform key word baselines in the task, that extractive summaries are considered to be coherent and useful by participants, and that while speech recognition errors impact user satisfaction, users adapt to the errors by modifying their browsing behavior.

2. EVALUATION OF MEETING SUMMARIZATION AND BROWSING

Our extrinsic evaluation was performed by embedding the outputs of each summarization system in a multimodal browser. In this section we provide a concise review of approaches to speech summarization and give an overview of the state-of-the-art in evaluation of both summarization systems and multimodal browsing interfaces.

2.1 Approaches to Speech Summarization

Automatic summarization systems fall into two rough classes: *extractive* and *abstractive*. Extractive summarization involves identifying informative sentences[2] in the source document and concatenating them to form a condensed version of the original, while abstractive summarization operates by generating novel sentences to convey the important information of the source document. There is no clear dividing line between the two approaches, and hybrid approaches are possible: for example, a system may extract the informative sentences and subsequently apply postprocessing techniques such as sentence compression and sentence rewriting in order to create novel summary text.

Extractive summarization may be posed as a binary classification task, in which each sentence must be labelled as informative or not. Thus this approach to the problem is very well suited to statistical pattern recognition approaches in which a classifier is trained on data labeled in this way. In this case each data point corresponds to a sentence that is represented as a feature vector. In previous work on extractive speech summarization, researchers have investigated the usefulness of lexical, prosodic, structural, and speaker-related features, among others [Valenza et al. 1999; Christensen et al. 2004; Kolluru et al. 2005; Koumpis and Renals 2005; Maskey and Hirschberg 2005; Galley 2006; Zhu and Penn 2006]. Such features have also been used in the development of unsupervised speech summarization algorithms [Zechner 2002; Hori et al. 2002; Murray et al. 2006].

The goal of abstractive summarization is the automatic generation of a coherent text that resembles a hand-written summary. Instead of selecting the most informative sentences from the source document, abstractive approaches attempt to mimic the processes of *interpretation*, *transformation*, and *generation* [Spärck Jones 1999] that are performed by human summarizers. Systems following this approach typically utilize a specialized representation formalism which is instantiated during a parsing process. The resulting model may be transformed using heuristic rules [Hahn and Reimer 1999; Paice and Jones 1993] to yield a representation of the final summary contents, but some approaches perform this step implicitly during parsing [Kameyama et al. 1996]. Using natural language generation components, the final text can be generated from the representation thus derived.

Explicit content representations allow for certain features which would be difficult to implement with the extractive approach (e.g., multilingual summarization) [DeJong 1982]. However, the complexity of suitable representation formalisms restricts the generality of abstractive systems and limits them to specific domains [Saggion and Lapalme 2002]. While Endres-Niggemeyer [1998, ch. 5] has reviewed the abstractive summarization of textual documents, less research has been done on the abstractive summarization of spoken discourse, although some promising approaches exist [Kameyama et al. 1996; Alexandersson 2003].

[2] Or, more generally, “sentence-like units”.

2.2 Summarization Evaluation

In this section we review three intrinsic approaches to summarization evaluation—ROUGE, the Pyramid method, and summarization accuracy—followed by several examples of frameworks for extrinsic summarization evaluation.

ROUGE [Lin and Hovy 2003] is a suite of evaluation metrics that matches a candidate summary against a set of reference summaries, and is a variation of the BLEU metric [Papineni et al. 2001] that has become standard in machine translation. Both BLEU and ROUGE are based on comparing n-gram overlap between machine outputs and human references. BLEU is a precision-based metric, whereas ROUGE was developed initially as a recall-based version of BLEU. However, the most recent versions of ROUGE calculate precision, recall, and f-score. There are several metrics within the ROUGE suite, but the most widely used are ROUGE-2 and ROUGE-SU4. ROUGE-2 computes the bigram overlap between the candidate and reference summaries, whereas ROUGE-SU4 calculates the skip bigram overlap with up to four intervening terms. Lin [2004] provided evidence that these metrics correlate well with human evaluations, using several years' worth of data from the Document Understanding Conference (DUC), an annual conference for research on query-based, multi-document summarization with a focus on newswire.[3] Subsequent research has yielded mixed results concerning ROUGE correlations with human evaluations [Dorr et al. 2004; Murray et al. 2005; Dorr et al. 2005; Murray et al. 2006], but ROUGE has become a standard metric for the Document Understanding Conference, and is widely used by summarization researchers, allowing them to directly compare summarization results on particular datasets.

The Pyramid method [Nenkova and Passonneau 2004] uses variable-length subsentential units for comparing machine summaries with human model summaries. These *semantic content units* (SCUs) are derived by human annotators who analyze multiple model summaries for units of meaning, with each SCU being weighted by how many model summaries it occurs in. These weights result in a pyramid structure, with a small number of SCUs occurring in many model summaries and most SCUs appearing in only a few model summaries. Machine summaries are then also annotated for SCUs and can be scored based on the sum of their SCU weights compared with the sum of SCU weights for an optimal summary. Using the SCU annotation, one can calculate both recall-based and precision-based summary scores. The advantage of the Pyramid method is that it uses content units of variable length and weights them by importance according to occurrence in model summaries. The disadvantage is that the scheme requires a great deal of human annotation. Pyramids were used as part of the DUC 2005 evaluation, with numerous institutions taking part in the peer annotation step, and while the submitted peer annotations required a substantial amount of corrections, Nenkova et al. [2007] reported acceptable levels for inter-annotator agreement. Galley [2006] introduced a matching constraint for the Pyramid method, namely that

[3] http://duc.nist.gov

when comparing machine extracts to model extracts, SCUs are only considered to match if they originate from the same sentence in the transcript. This was done to account for the fact that sentences might be superficially similar, in each having a particular SCU, but nevertheless with different overall meanings.

An inherent difficulty of evaluation approaches based on the comparison of n-grams is that they do not account for the fact that relevant information may be conveyed using different wordings. For instance, the source document may contain summary-worthy material multiple times in paraphrased versions. Abstractive summarization approaches, in particular, may score relatively poorly in such evaluations, since their vocabulary may be rather different to the reference summaries. It is a strength of the Pyramid method that it can match content units with varying surface forms.

Zechner and Waibel [2000] introduced an evaluation metric specifically for speech summarization, *summarization accuracy*. The general intuition is that an evaluation method for such summaries should take into account the relevance of the units extracted as well as the recognition errors for the words that comprise the extracted units. Annotators are given a topic-segmented transcript and told to select the most relevant phrases in each topic. For summaries of recognizer output, the words of the ASR transcripts are aligned with the words of the manual transcripts. Each word has a relevance score equal to the average number of times it appears in the annotators' most relevant phrases. Given two candidate sentences, sentence 1 might be superior to sentence 2 when summarizing manual transcripts if it contains more relevant words, but if sentence 1 has a higher word error rate (WER) than sentence 2, it may be a worse candidate for inclusion in a summary of the ASR transcript. Summaries with high relevance and low WER will thereby rate more highly.

A variety of extrinsic evaluation approaches have been proposed for text summarization, based on tasks such as *relevance assessment*, and *reading comprehension*. In relevance assessment [Mani 2001], a user is presented with a description of a topic or event and must then decide whether a given document (summary or full-text) is relevant to that topic or event. Such schemes have been used for a number of years and in a variety of contexts [Jing et al. 1998; Mani et al. 1999; Harman and Over 2004]. Due to problems of low inter-annotator agreement on such ratings, Dorr et al. [2005] proposed a new evaluation scheme that compares the relevance judgement of an annotator given a full text with that same annotator given a condensed text.

In the reading comprehension task [Hirschman et al. 1999; Morris et al. 1992; Mani 2001], a user is asked to read either a full source or a summary text and then completes a multiple-choice comprehension test relating to the full source information. This may then be used to calculate how well a summarization system has performed in terms of the user's comprehension score. The reading comprehension evaluation framework relies on the idea that truly informative summaries should be able to act as substitutes for the full source.

In the DUC evaluations, in which summaries were produced in response to a query, human judges assigned a pseudo-extrinsic *responsiveness* score to each machine summary, representing how well the given summary satisfied the information need in the query. This is not a true task-based extrinsic evaluation, but does give a sense of the potential utility of the summary in light of the query. Daumé and Marcu [2005] have suggested an extrinsic evaluation framework based on a relevance prediction task, pointing out that some of the considerable time and labor required for annotations such as for the Pyramid scheme could be spent implementing a simple task-based evaluation.

2.3 Browser-Based Evaluation

A number of other extrinsic evaluations have used meeting browsing tasks to study how users' performance is influenced by the specific browser setups and the information available to the user within a meeting browser. Most of these evaluations, however, did not focus on meeting summaries in particular.

In attempt to objectively evaluate browser performance, Wellner et al. [2005] introduced the Browser Evaluation Test (BET). In the BET, the subject must decide whether certain *observations of interest*, for example, the observation "Susan says the footstool is expensive." are true or false for a given meeting. Often times, however, the observations are such that their truth value can be found through a simple keyword search ("footstool") without requiring the user to read a summary. Furthermore, in the BET as currently formulated, the annotated observations of interest tend to refer to single points of occurrence. For our own experiment, we chose a more complex information need instead, the decision audit task.

The Task-Based Evaluation (TBE) is an alternative browser evaluation, also developed in the context of the AMI project [Kraaij and Post 2006] which evaluates multiple browser conditions relating to a series of AMI meetings. The subjects of the evaluation are told that they are replacing a previous team and must finish that team's work. The participants are given information (in the form of meeting recordings, documentation, and a meeting browser) related to the previous meetings in the series, and must finalize the previous group's decisions as best as possible given what they know. The TBE relies primarily on postmeeting questionnaire answers for evaluation, which is one of the reasons we have not adopted this evaluation approach. While we do incorporate such questionnaires in our evaluation, we are also very interested in objective measures of participant performance, and in browsing behavior during the task.[4] Furthermore, the TBE is more costly to carry out than our decision audit task, as it requires taking approximately three hours to review previous meetings and to conduct their own meetings, which are also recorded; the decision audit, on the other hand, is an individual task that is completed in less than an hour.

SCANMail [Hirschberg et al. 2001; Whittaker et al. 2002] provides an interface for managing and browsing voicemail messages, using multimedia components such as audio, ASR transcripts, audio-based paragraphs, and extracted names and phone numbers. Both in a think-aloud laboratory study and a larger field study, users found the SCANMail system outperformed a state-of-the-art voicemail system for several extrinsic tasks. The field study in particular yielded several interesting findings. In 24% of the times that users viewed a

[4] It would not be impossible to include in the TBE, but would involve considerable additional instrumentation of individual and group behavior, as well as additional analysis.

voicemail transcript with the SCANMail system, they did not resort to playing the audio. This testifies to the fact that the transcript and extracted information can, to some degree, act as substitutes for the signal, which comments from users confirm. Most interestingly, 57% of the audio play operations resulted from clicking within the transcript. The study also found that users were able to understand the transcripts even with recognition errors, partly by having prior context for many of the messages.

SpeechSkimmer [Arons 1997] is an audio-based browser incorporating skimming, compression, and pause-removal techniques for the efficient navigation of large amounts of audio data. The authors conducted a formative usability study in order to refine the interface and functionality of SpeechSkimmer, recruiting participants to find several pieces of relevant information within a large portion of lecture speech using the browser. Results were gleaned both from a think-aloud experiment structure as well as follow-up questions on ease-of-use. The researchers found that experiment participants often began the task by listening to the audio at normal speed to first get a feel for the discussion, and subsequently made good use of the skimming and compression features to increase search efficiency.

3. THE AMI MEETING CORPUS

The AMI corpus [Carletta et al. 2005; Carletta 2006] consists of about 100 hours of recorded and annotated multiparty meetings. Meeting were recorded using multiple microphones and cameras; in addition, handwritten notes, data-projected presentations and whiteboard events were also captured. The corpus is divided into *scenario* and *nonscenario* meetings. In the scenario meetings, four participants take part in each meeting and play roles within a fictional company. The scenario given to them is that they are part of a company called Real Reactions, which designs remote controls. Their assignment is to design and market a new remote control, and the members play the roles of project manager (the meeting leader), industrial designer, user-interface designer, and marketing expert. Through a series of four meetings, the team must bring the product from inception to market.

The first meeting of each series is the kick-off meeting, where participants introduce themselves and become acquainted with the task. The second meeting is the functional design meeting, in which the team discusses the user requirements and determines the functionality and working design of the remote. The third meeting is the conceptual design of the remote, wherein the team determines the conceptual specification, the user interface, and the materials to be used. In the fourth and final meeting, the team determines the detailed design and evaluates their result.

The participants are given real-time information from the company during the meetings, such as information about user preferences and design studies, as well as updates about the time remaining in each meeting. While the scenario given to them is artificial, the speech and the actions are completely spontaneous and natural. There are 140 meetings of this type in total. The length of an individual meeting ranges from about 15 to 45 minutes, depending on which meeting in the series it is and how quickly the group is working.

The nonscenario meetings are meetings that occur regularly and would have been held regardless of the AMI data collection, and so the meetings feature a variety of topics discussed and a variable number of participants.

The meetings were recorded, in English, at three locations: the University of Edinburgh (UK), TNO (Netherlands), and Idiap Research Institute (Switzerland). The participants included both native and non-native English speakers, and many of them are students. Of the 53% who were non-native English speakers, 53% were native Dutch speakers (i.e., all the TNO participants).

The AMI corpus is freely available,[5] and contains numerous annotations for a variety of multimodal and linguistic phenomena.

4. EXPERIMENTAL SETUP

4.1 Task Data

The data for the extrinsic evaluation was one meeting series (ES2008) from the AMI corpus, comprising four related, sequential meetings. The particular meeting series was chosen because the participant group in that series worked well together on the task. The group took the task seriously and exhibited deliberate and careful decision-making processes in each meeting and across the meetings as a whole.

4.2 Decision Audit Task

The extrinsic evaluation was based on an individual (rather than group-based) decision audit task. We recruited only participants who were native English speakers and who had not participated in previous AMI experiments or data collection. We also checked that each participant had at least moderate familiarity with computers. Many of the participants were graduate-level students at university. The gender breakdown was 27 females and 23 males.

We collected data from five conditions, with ten subjects tested in each condition in a between-subjects design, resulting in a total of fifty subjects. For each condition, six participants completed the task in Edinburgh and four at DFKI.

The experimental setups for the two locations were as similar as possible, with comparable desktop machines running Linux, 17-inch monitors, identical browser interfaces, and the same documents used in each location, as described below.

Each participant was first given a pretask questionnaire (hereafter referred to as the pre-questionnaire) relating to background, computer experience, and experience in attending meetings. In the case that the participant regularly participated in meetings of their own, we asked how they normally prepared for a meeting (e.g., using their own notes, consulting with other participants, etc.).

Each participant was then given general task instructions. These instructions explained the meeting browser in terms of the information provided in the browser and the navigation functions of the browser; the specific information need they were supposed to satisfy in the task; and a notice of the allotted

[5] http://corpus.amiproject.org/

Table I. Experimental Conditions

Condition	Summary	Algorithm	Transcript
KAM	Keyword	Automatic	Manual
EAM	Extractive	Automatic	Manual
EAA	Extractive	Automatic	Automatic
AMM	Abstractive	Manual	Manual
ASM	Abstractive	Semi-Auto.	Manual

time for the task. The total time allotted was 45 minutes, which included both searching for the information and writing up the answer. This amount of time was based on the result of pilot experiments for Condition EAM, extractive summarization on manual transcripts (see below).

The portion of the instructions detailing the specific task read as follows:

> We are interested in the group's decision-making ability, and therefore ask you to evaluate and summarize a particular aspect of their discussion.
> The group discussed the issue of separating the commonly-used functions of the remote control from the rarely-used functions of the remote control. What was their final decision on this design issue? Please write a short summary (1-2 paragraphs) describing the final decision, any alternatives the participants considered, the reasoning for and against any alternatives (including why each was ultimately rejected), and in which meetings the relevant discussions took place.

This particular information need was chosen because the relevant discussion manifested itself throughout the four meetings, and the group went through several possibilities before designing an eventual solution to this portion of the design problem. In the first meeting, the group discussed the possibility of creating two separate remotes. In the second meeting, it was proposed to have simple functions on the remote and more complex functions on a sliding compartment of the remote. In the third meeting, they decided to have an on-screen menu for complex functions; and in the final meeting they finalized all of the details and specified the remote buttons. A participant in the decision audit task would therefore have had to consult each meeting to be able to retrieve the full answer to the task's information need.

While in this case the participant had to determine the decision that was made *and* the reasons behind the decision, in theory the decision audit could be set up in such a way that the decision itself is a given and only the reasoning behind the decision must be determined.

After completing the decision audit task, participants answered a post-task questionnaire (hereafter referred to as the post-questionnaire). The post-questionnaire is described in detail in Section 4.6.

4.3 Experimental Conditions

There were five conditions implemented in total: one baseline condition; two extractive summarization conditions; and two abstractive summarization conditions. Table I lists and briefly describes the experimental conditions. The three-letter ID for each condition corresponds to (i) Keywords/Extracts/Abstracts; (ii) Automatic/Semi-automatic/Manual summarization algorithms; and (iii) Automatic/Manual transcripts. In the extractive summarization condition, the units of speech that were extracted were dialogue acts (DAs).

The baseline condition, Condition KAM, consisted of a browser with manual transcripts, audio/video record, and a list of the top 20 keywords in the meeting. The keywords were determined automatically using the *su.idf* term-weighting scheme [Murray and Renals 2007]. Though this was a baseline condition, the fact that it utilized *manual* transcripts gave users in this condition a possible advantage over users in conditions with ASR. In this respect, it was a challenging baseline. There are other possibilities for the baseline, but we chose the top 20 keywords because we were interested in comparing different forms of derived content from meetings, and because a facility such as keyword search would likely have been problematic for a participant who is uncertain of what to search for because they are unfamiliar with the meetings.

Conditions EAM and EAA presented the user with a transcript, audio/video record, and an extractive summary of each meeting, with the difference between the conditions being that the latter was based on ASR and the former on manual transcripts. The length of the respective extractive summaries was based on the length of the manual extracts for each meeting: approximately 1000 words for the first meeting; 1900 words each for the second and third meetings; and 2300 words for the final meeting. These lengths correlate to the lengths of the meetings themselves and represent compressions to approximately 40%, 32%, 32%, and 30% of the total meeting word counts, respectively.

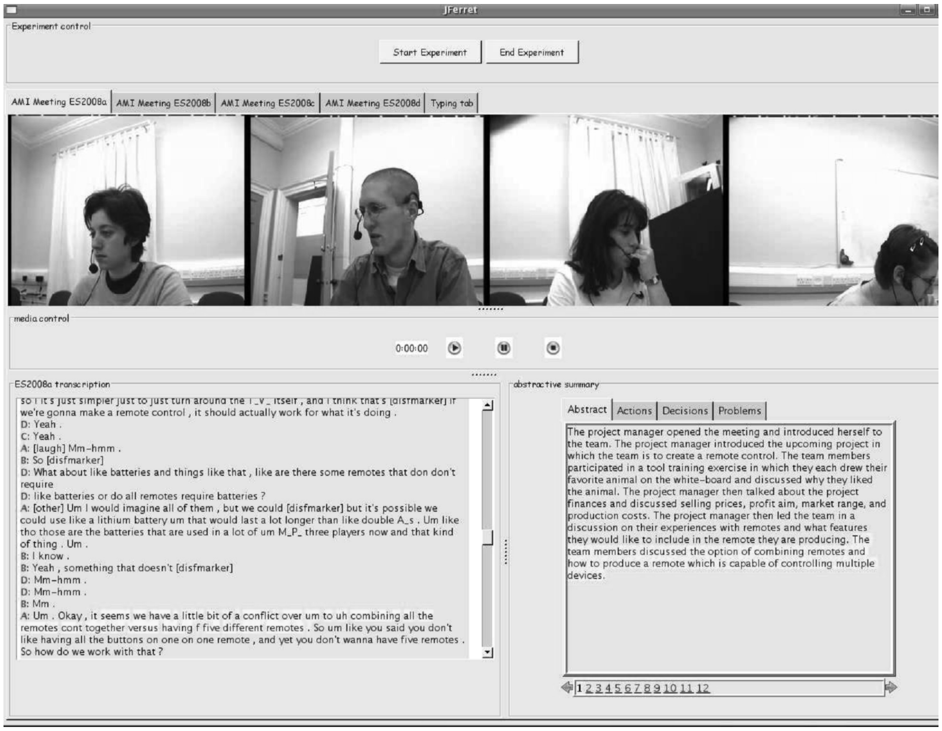
The extractive summarization was performed using a support vector machine (SVM) with a radial basis function (RBF) kernel to classify each DA as extractive or nonextractive, trained on the AMI-labelled training data (90 scenario meetings) using 17 features from 5 broad feature classes: prosodic, lexical, length, structural, and speaker-related. Table II lists all of the features used. The energy and F0 (pitch) features were first calculated at the word level and then averaged over each DA. These were also normalized by speaker. Precedent and subsequent pause refer to pause length before and after a DA. The rate of speech was a rough calculation, using the number of words in a DA divided by the DA duration. Two structural features were used: DA position in the turn, and in the meeting. Two features captured information about speaker dominance, giving the percentage of dialogue acts in the meeting uttered by the current speaker and the percentage of total speaking time represented by the current speaker's utterances. Three features indicated the length or duration of a DA; and we finally included two term-weighting metrics. One is the classic *tf.idf* metric that favors terms that occur frequently in the given document but rarely across a set of documents, while *su.idf* [Murray and Renals 2007] weights terms highly that are used with varying frequency among the meeting participants.

We ran the classifier on the four meetings of interest, ranking dialogue acts in descending order of informativeness according to posterior probability, extracting until we reach the desired summary length. The word error rate for the ASR transcripts in this corpus overall is 38.9%.

Table II. Extractive Summarization Features Key

Feature ID	Description
Prosodic Features	
ENMN	mean energy
F0MN	mean F0
ENMX	max energy
F0MX	max F0
F0SD	F0 stdev.
PPAU	precedent pause
SPAU	subsequent pause
ROS	rate of speech
Structural Features	
MPOS	meeting position
TPOS	turn position
Speaker Features	
DOMD	speaker dominance (DAs)
DOMT	speaker dominance (seconds)
Length Features	
DDUR	DA duration
UINT	uninterrupted length
WCNT	number of words
Lexical Features	
SUI	su.idf sum
TFI	tf.idf sum

Condition AMM was the gold-standard condition, a human-authored abstractive summary. Annotators were asked to write abstractive summaries of each meeting and to extract the meeting dialogue acts that best convey or support the information in the abstractive summary. They used a graphical user interface (GUI) to browse each individual meeting, allowing them to view previous human annotations comprised of an orthographic transcription synchronized to the meeting audio, and topic segmentation. The annotators were first asked to build a textual summary of the meeting aimed at an interested third-party, divided into four sections of "general abstract," "decisions," "actions" and "problems." These abstractive summaries varied in length, but the maximum permitted length for each summary section was 200 words. While it was mandatory that each general abstract section contained text, it was permitted that for some meetings the other three sections could be null; for example, some meetings might not involve any decisions being made. After authoring the abstractive summary, annotators were then asked to create an extractive summary. To do so, they were told to extract the dialogue acts that together could best convey the information in the abstractive summary and could be used to support the correctness of the abstract. They were not given any specific instructions about the number or percentage of dialogue acts to extract, nor any instructions about extracting redundant dialogue acts. They were then required to do a second-pass annotation, wherein for each extracted dialogue act they chose the abstract sentences supported by that dialogue act. The result is a many-to-many mapping between abstract sentences and extracted dialogue acts, that is, an abstract sentence can be linked to more than one dialogue act and vice



so I it's just simpler just to just turn around the T_V_ itself , and I think that's [disfmarker] if we're gonna make a remote control , it should actually work for what it's doing .
D: Yeah .
C: Yeah .
A: [laugh] Mm-hmm .
B: So [disfmarker]
D: What about like batteries and things like that , like are there some remotes that don don't require
D: like batteries or do all remotes require batteries ?
A: [other] Um I would imagine all of them , but we could [disfmarker] but it's possible we could use like a lithium battery um that would last a lot longer than like double A_s . Um like tho those are the batteries that are used in a lot of um M_P_ three players now and that kind of thing . Um .
B: I know .
B: Yeah , something that doesn't [disfmarker]
D: Mm-hmm .
D: Mm-hmm .
B: Mm .
A: Um . Okay , it seems we have a little bit of a conflict over um to uh combining all the remotes cont together versus having f five different remotes . So um like you said you don't like having all the buttons on one on one remote , and yet you don't wanna have five remotes . So how do we work with that ?

The project manager opened the meeting and introduced herself to the team. The project manager introduced the upcoming project in which the team is to create a remote control. The team members participated in a tool training exercise in which they each drew their favorite animal on the white-board and discussed why they liked the animal. The project manager then talked about the project finances and discussed selling prices, profit aim, market range, and production costs. The project manager then led the team in a discussion on their experiences with remotes and what features they would like to include in the remote they are producing. The team members discussed the option of combining remotes and how to produce a remote which is capable of controlling multiple devices.

Fig. 1. Condition AMM browser.

versa. Because of these summary-transcript links, the experimental condition AMM is a hybrid of abstractive and extractive summaries. Since this is a decision audit task and the abstractive summary provided in this condition had a "decisions" subsection, we considered this to be a challenging gold-standard condition, in that decisions were explicitly provided. Figure 1 shows an example of the browser interface for Condition AMM.

Condition ASM presented the user with a semi-automatically-generated abstractive summary, using an approach described by Kleinbauer et al. [2007]. This summarization method utilized hand-annotated topic segmentation and topic labels, and detected the most commonly mentioned content items in each topic. A sentence was generated for each meeting topic indicating roughly what was discussed, and these sentences were linked to the actual DAs in the discussion. These summaries relied on manual transcripts, and so Condition EAA was the only ASR condition in this experiment. The Condition ASM summaries were only semi-automatic, since they relied on manual annotation of propositional content. The summaries in this condition did not feature separate sections for decisions, action items, or problems as in condition AMM. They consist solely of a single paragraph abstract that parallels the structure of the meeting.

While there are other potentially interesting conditions to run, for example, conditions corresponding to KAA and ASA, these five conditions were chosen so that we could evaluate several questions: how summaries compare with simpler derived content, how extractive summaries compare with human-authored abstracts, whether automatic speech recognition significantly decreases the coherence and usefulness of cut-and-paste summaries, and how automatic abstraction techniques compare with human abstracts.

4.4 Browser Setup

The meeting browsers were built so as to exhibit as similar browser behavior as possible across the experimental conditions. In other words, the interface was kept essentially the same in all conditions in an attempt to eliminate any potential confounding factors relating to the user interface.

In each browser, there were five tabs: one for each of the four meetings and a writing pad. The writing pad was provided for the participants to author their decision audit answer.

In each meeting tab, the videos displaying the four meeting participants were laid out horizontally with the media controls beneath. The transcript was shown in the lower left of the browser tab in a scroll window.

In Condition KAM, each meeting tab contained buttons corresponding to the top 20 keywords for that meeting. Pressing the button for a given keyword highlighted the first instance of the keyword in the transcript, and also opened a listbox illustrating all of the occurrences of the word in the transcript to give the user a context in terms of the word's frequency. Subsequent clicks highlighted the subsequent occurrences of the word in the transcript, or the user might choose to navigate to keyword instances via the listbox.

In Conditions EAM and EAA, a scroll window containing the extractive summary appeared next to the full meeting transcript. Clicking on any dialogue act in the extractive summary took the user to that point of the meeting transcript and audio/video record.

In Conditions AMM and ASM, the abstractive summary was presented next to the meeting transcript. In Condition AMM, the abstractive summary had different tabs for *abstract*, *actions*, *decisions*, and *problems*. Clicking on any abstract sentence highlighted the first linked dialogue act in the transcript and also presented a listbox representing all of the transcript dialogue acts linked to that abstract sentence. The user could thus navigate either by repeatedly clicking the sentence, which in turn would take them to each of the linked dialogue acts in the transcript, or else they could choose a dialogue act from the listbox. The navigation options were essentially the same as Condition KAM. The primary difference between Conditions KAM, AMM, and ASM on the one hand and Conditions EAM and EAA on the other was that the extractive dialogue acts linked to only one point in the meeting transcript, whereas keywords and abstract sentences had multiple indices.

Since the writing pad, where the participant typed their answer, was a fifth tab in addition to the four individual meeting tabs, the participant could not view the meeting tabs while typing the answer: they were restricted to tabbing back and forth as needed. This was designed deliberately so as to be able to discern when the participant was working on formulating or writing the answer on the one hand and when they were browsing the meeting records on the other.

After reading the task instructions, each participant was briefly shown how to use the browser's various functions for navigating and writing in the given experimental condition. They were then given several minutes to familiarize themselves with the browser, until they stated that they were comfortable and ready to proceed. The meeting used for this familiarization session was not one of the ES2008 meetings used in the actual task. In fact, it was one of the AMI nonscenario meetings; this was done so that the participant would not become familiar with the ES2008 meetings specifically or the scenario meetings in general before beginning the task. The familiarization time was carried out before the task began so that we were able to control for the possibility that one condition would be more difficult to learn than the others. While participants were offered as much time as they needed, this was typically less than five minutes.

All browsers were built on top of the JFerret[6] framework [Wellner et al. 2004]. Tucker and Whittaker [2004] describe a four-way meeting browser typology: audio-based browsers, video-based browsers, artefact-based browsers, and derived data browsers. In light of this classification scheme, our decision audit browsers—which provide synchronized transcripts, summaries, audio and video—may be regarded as video-based browsers incorporating derived data forms.

4.5 Logging Browser Use

In each condition of the experiment, we instrumented the browsers to enable recording of a variety of information relating to the participant's browser use and typing. In all conditions, we logged transcript clicks, media control clicks (i.e. play, pause, stop), movement between tabs, and characters entered into the typing tab, all of which were time-stamped. In Condition KAM, we logged each keyword click and noted its index in the listbox (e.g., the first occurrence of the word in the listbox). In Conditions EAM and EAA, each click of an extractive summary sentence was logged, and in the abstract conditions each abstract sentence click was logged along with its index in the listbox, analogous to the keyword condition. Because sentences in the extractive summaries have only a single transcript index (i.e., the sentence's original location in the meeting), there was no need for listboxes and listbox indices in the extractive conditions.

To give an example, the following portion of a logfile from a Condition AMM task shows that the participant clicked on the transcript, played the audio, paused the audio, clicked link number 1 of sentence 5 in the Decisions tab for the given meeting, then switched to the typing tab and began typing the word "six".

```
2007-05-24T14:46:45.713Z transcript_jump 687.85 ES2008d.sync.1375
2007-05-24T14:46:45.715Z button_press play state media_d
2007-05-24T14:46:45.715Z button_press play state media_d
2007-05-24T14:47:30.726Z button_press pause state media_d
2007-05-24T14:47:30.726Z button_press pause state media_d
```

[6] http://www.idiap.ch/mmm/tools/jferret

Table III. Decision Audit Evaluation Features

Post-Questionnaire	Human Ratings	Logfile
Q1: *I found the meeting browser intuitive and easy to use*	C1: *overall quality*	Q1: *task duration*
Q2: *I was able to find all of the information I needed*	C2: *conciseness*	Q2: *first typing*
Q3: *I was able to efficiently find the relevant information*	C3: *completeness*	Q3: *amount of tabbing*
Q4: *I feel that I completed the task in its entirety*	C4: *task comprehension*	Q4: *perc. buttons clicked*
Q5: *I understood the overall content of the meeting discussion*	C5: *participant effort*	Q5: *clicks per minute*
Q6: *The task required a great deal of effort*	C6: *writing style*	Q6: *media clicks*
Q7: *I had to work under pressure*	C7: *objective rating*	Q7: *click/writing correlation*
Q8: *I had the tools necessary to complete the task efficiently*		Q8: *unedited length*
Q9: *I would have liked additional information about the meetings*		Q9: *edited length*
Q10: *It was difficult to understand the content of the meetings using this browser*		Q10: *num. meetings viewed*
		Q11: *ave. writing timestamp*

```
2007-05-24T14:47:52.379Z MASCOT (observation ES2008d): selected link
#1 in sentence #5 of tab ‘decisions’
2007-05-24T14:47:53.613Z tab_selection Typing tab
2007-05-24T14:47:54.786Z typed_insert s 316
2007-05-24T14:47:54.914Z typed_insert i 317
2007-05-24T14:47:55.034Z typed_insert x 318
```

4.6 Evaluation Features

To evaluate the decision audit task, we analyzed three types of features: the answers to the users' post-questionnaires, human ratings of the users' written answers, and features extracted from the logfiles that relate to browsing and typing behavior in the different conditions. Table III lists all the features used for the evaluation. Using these three types of evaluations allows us to assess how satisfied users were with the provided tools, how they performed objectively on the task, and whether their browsing behavior was significantly impacted by the experimental condition.

Upon completion of the decision audit task, we presented each participant with a post-task questionnaire consisting of 10 statements with which the participant could state their level of agreement or disagreement via a 5-point Likert scale, such as *I was able to efficiently find the relevant information*, and two open-ended questions about the specific type of information available in the given condition and what further information they would have liked. Of the 10 statements evaluated, some were rewordings of others with the polarity reversed in order to gauge the users' consistency in answering.

In order to gauge the goodness of a participant's answer, we enlisted two human judges to do both subjective and objective evaluations. The judges were familiar with the idea of scenario meetings in the AMI corpus, but were not briefed in particular on the contents of the meeting series that was used in the experiment. For the subjective portion, the judges first read through all 50 answers to get a view of the variety of answers. They then rated each answer using an 8-point Likert-scale on criteria roughly relating to the precision, recall, and f-score of the answer (summarized in the second column of Table III). For the objective evaluation, three judges constructed a gold-standard list of items that should have been contained in an ideal summary of the decision audit. Two judges first drafted a list of gold-standard items they considered to be critical to the issue of separating the remote control's functions, after reviewing the four meetings. This list was then reviewed and edited by a third judge to create the final set of gold-standard items. For each participant answer, they checked off how many of the gold-standard items were contained. An example of a gold-standard item is the group's agreement in meeting ES2008a that the remote control must not have too many buttons.

The remainder of the features for evaluation were automatically derived from the logfiles. These features have to do with browsing and writing behavior as well as the duration of the task. These included the total experiment length; the amount of time before the participant began typing their answer; the total amount of tabbing per user (normalized by experiment length); the number of clicks on content buttons (e.g., keyword buttons or extractive summary sentences) per minute; the number of content button clicks normalized by the number of unique content buttons; number of times the user played the audio/video stream; the number of content clicks prior to the user clicking on the writing tab to begin writing; the document length including deleted characters; the document length excluding deleted characters; how many of the four meetings the participant looked at; and the average typing timestamp normalized by the experiment length.

The total experiment length was included because it was assumed that participants would finish earlier if they had better and more efficient access to the relevant information. The amount of time before typing begins was included because it was hypothesized that efficient access to the relevant information would mean that the user would begin typing the answer sooner. The total amount of tabbing was considered because a participant who was tabbing very often during the experiment was likely jumping back and forth between meetings trying to find the information, indicating that the information was not conveniently indexed. The content clicks were considered because a high number of clicks per minute would indicate that the participant was finding that method of browsing to be helpful, and the number of content clicks normalized by the total unique content buttons indicated whether they made full use of that information source. The number of audio/video clicks was interesting because it was hypothesized that a user without efficient access to the relevant information would rely more heavily on scanning through the audio/video stream in search of the answers. The number of content clicks prior to the user moving to the writing tab indicated whether a content click is helpful in finding a piece of information that led to writing part of the answer. The document length was considered because a user with better and more efficient access to

Table IV. Post-questionnaire Results

Question	KAM	EAM	EAA	AMM	ASM
Q1	3.8	4.0	3.02^{AMM}	4.3_{EAA}	3.7
Q2	2.9^{AMM}	3.8	2.9^{AMM}	4.1_{KAM}	3.0
Q3	2.8^{AMM}	3.4	2.5^{AMM}	$4.0_{KAM,EAA,ASM}$	2.65^{AMM}
Q4	2.3	3.1	2.3	3.2	2.9
Q5	3.8	4.5	3.9	4.1	3.9
Q6	3.0	2.6_{EAA}	3.9^{EAM}	3.1	3.2
Q7	3.3	2.6	3.3	2.7	3.1
Q8	3.1^{EAM}	$4.3_{KAM,EAA}$	3.0^{EAM}	4.1	3.5
Q9	3.0	2.0	2.4	2.6	2.7
Q10	2.1	1.5_{EAA}	2.7^{EAM}	2.0	2.3

For each score in the table, that score is significantly worse than the score for any conditions in superscript, and significantly better than the score for any condition in subscript.

the meeting record would be able to spend more time writing and less time searching. Because the logfiles showed deleted characters, we calculated both the total amount of typing and the length of the final edited answer in characters. The number of meetings examined was considered because a user who had trouble finding the relevant information might not have had time to look at all four meetings. The final feature, which is the average timestamp normalized by the experiment length, was included because a user with efficient access to the information would be able to write the answer throughout the course of the experiment, whereas somebody who had difficulty finding the relevant information might have tried to write everything at the last available moment.

5. RESULTS

The following sections present the post-questionnaire results, the human subjective and objective evaluation results, and the analysis of browsing behaviors.

5.1 Post-questionnaire Results

Table IV gives the post-questionnaire results for each condition. For each score in the table, that score is significantly worse than the score for any conditions in superscript, and significantly better than the score for any condition in subscript. The only significant results listed are those that are significant at the level ($p<0.05$) according to analysis of variance (anova) and a post-hoc Tukey test.

The gold-standard condition AMM scored best on many of the criteria, showing that human abstracts are the most efficient of the methods studied in terms of surveying and indexing into the content of a meeting. For example, participants in this condition found that the meeting browser was easy to use (Q1) and that they could efficiently find the relevant information (Q3).

A striking result is that not only were the automatic extracts in condition EAM also rated highly on many post-questionnaire criteria, this condition was in fact the best overall for several of the questions. For example, on being able to understand the overall content of the meeting discussion (Q5) and having the tools necessary to complete the task efficiently (Q8), Condition EAM scored best.

However, it's clear that extracts of ASR output posed challenges that significantly decreased user satisfaction levels according to several of the criteria. For example, participants in Condition EAA found the browser less intuitive and easy to use (Q1); found it more difficult to understand the meeting discussion (Q5); and used considerable effort to complete the task (Q6). On several criteria this condition rated the same or worse than the baseline Condition KAM, which uses manual transcripts.

Condition ASM incorporating semi-automatic abstracts rated well in comparison with the gold-standard condition on some criteria, scoring not significantly worse than Condition AMM on criteria relating to the ability to understand the meeting discussion and complete the task (Q4 and Q5). However, Condition ASM was rated less highly on Q2, Q9, and Q10, suggesting that users may have liked additional information about the meetings.

Discussion. It can first be noted that participants in general found the task to be challenging, as evidenced by the average answers on questions 4, 6, and 7. The task was designed to be challenging and time-constrained, because a simple task with a plentiful amount of allotted time would have allowed the participants to simply read through the entire transcript or listen and watch the entire audio/video record in order to retrieve the correct information, disregarding other information sources. The task as designed required efficient navigation of the information in the meetings in order to finish the task completely and on time.

The gold-standard human abstracts were rated highly on average by participants in that condition. Judging from the open-ended questions in the post-questionnaire, people found the summaries and specifically the summary subsections to be very valuable sources of information. One participant remarked "Very well prepared summaries. They were adequate to learn the jist [sic] of the meetings by quickly skimming through... I especially liked the tabs (Decisions, Actions, etc.) that categorized information according to what I was looking for." As mentioned earlier, this gold-standard condition was expected to do particularly well considering that it was a decision audit task and the abstractive summaries contain subsections that were specifically focused on decision-making in the meetings.

Condition ASM rated quite well on Q1, regarding ease of use and intuitiveness, but slightly less well in terms of using the browser to locate the important information. It did consistently rate better than Conditions KAM and EAA, however.

The results of the post-questionnaire data are encouraging for the extractive paradigm, in that the users seemed very satisfied with the extractive summaries relative to the other conditions. It is not surprising that the gold-standard human-authored summaries were ranked best overall on several criteria, but even on those criteria the extractive condition on manual transcripts is a close second. Perhaps the most compelling result was on question 8, relating to having the tools necessary to complete the task. Not only was Condition EAM rated the best, but it was *significantly better* than Conditions KAM and EAA. These results indicate that extractive summaries are natural to use as navigation tools, that they facilitate understanding of the meeting content, and allow users to be more efficient with their time.

However, it is quite clear that the errors within an ASR transcript presented a considerable problem for users trying to quickly retrieve information from the meetings. While it has repeatedly been shown that ASR errors do not cause problems for these summarization algorithms according to intrinsic measures, these errors made user comprehension more difficult. For the questions relating to the effort required, the tools available, and the difficulty in understanding the meetings, Condition EAA was easily the worst, scoring even lower than the baseline condition. It should be noted, however, that a baseline such as Condition KAM was not a true baseline, in that it was working off of *manual* transcripts and would be expected to be worse when applied to ASR. Judging from the open-ended questions in the post-questionnaires, it is clear that at least two participants found the ASR so difficult to work with that they tended not to use the extractive summaries, let alone the full transcript, relying instead on watching the audio/video as much as possible. For example, one person responded to the question "How useful did you find the list of important sentences from each meeting?" with the comment "Not at all, because the voice recognition technology did not work properly. The only way to understand the discussion was to listen to it all sequentially, and there simply wasn't time to do that". We will analyze users' browsing behavior in much more detail below.

The ASR used in these experiments had a WER of about 39%; it is to be expected that these findings regarding the difficulty of human processing of ASR transcripts will change and improve as the state-of-the-art in speech recognition improves. The finding also indicates that the use of confidence scores in summarization is desirable. While summarization systems naturally tend to extract units with lower WER, the summaries can likely be further improved for human consumption by compression via the filtering of low-confidence words.

5.2 Human Evaluation Results: Subjective and Objective

Before beginning the subjective evaluation of decision audit answers, the two human judges read through all 50 answers in order to gauge the variety of answers in terms of completeness and correctness. They then rated each answer on several criteria roughly related to ideas of precision, recall, and f-score, as well as effort, comprehension, and writing style. They used an 8-point Likert scale for each criterion. We then averaged their scores to derive a combined subjective score for each criterion.

After the annotators carried out their initial subjective evaluations, they met again and went over all experiments where their ratings diverged by more than two points, in order to form a more objective and agreed-upon evaluation of how many gold-standard items each participant found. There were 12 out of 50 ratings pairs that needed revision in this manner. After the judges' consultation on those 12 pairs of ratings, each experiment was given a single objective rating. The judges mentioned that they found this portion of the evaluation much more difficult than the subjective evaluations, as there was often

Table V. Human Evaluation Results: Subjective and Objective

Criterion	KAM	EAM	EAA	AMM	ASM
C1	3.0^{AMM}	4.15	3.05	4.65_{KAM}	4.3
C2	2.85^{AMM}	4.25	3.05	4.85_{KAM}	4.45
C3	2.55^{AMM}	3.6	2.6^{AMM}	$4.45_{\text{KAM,EAA}}$	3.9
C4	$3.25^{\text{EAM,AMM}}$	5.2_{KAM}	3.65	5.25_{KAM}	4.7
C5	4.4	5.2	3.7	5.3	4.9
C6	4.75	5.65	$4.1^{\text{AMM,ASM}}$	5.7_{EAA}	5.8_{EAA}
Objective	4.25^{AMM}	7.2	5.05	9.45_{KAM}	7.4

For each score in the table, that score is significantly worse than the score for any conditions in superscript, and significantly better than the score for any condition in subscript.

ambiguity as to whether a given answer contained a given gold-standard item or not.

The objective evaluation indicates that this was a challenging task in all conditions. For example, even in the gold-standard Condition AMM there were some people who could only find one or two relevant items while others found 16 or 17. Given a challenging task and a limited amount of time, some people may have simply felt overwhelmed in trying to locate the informative portions efficiently.

Table V gives the results for the human subjective and objective evaluations, formatted analogously to Table IV. As in the case of the post-questionnaires, Condition AMM scored best in the subjective as well as in the objective evaluation for most criteria. But we also observe that neither Condition ASM nor Condition EAM were significantly worse. However, the introduction of ASR had a measurable and significant impact on the subjective evaluation of quality. At the same time, what these findings together help illustrate is that automatic summaries can be very effective for conducting a decision audit by helping the user to generate a concise and complete high-quality answer. Interestingly, the scores on each criterion and for each condition, including Condition AMM, tended to be somewhat low on the Likert scale, due to the difficulty of the task.

Discussion. For the objective human evaluation, the gold-standard condition scored substantially higher than the other conditions in hitting the important points of the decision process being audited. This indicates that there is much room for improvement in terms of automatic summarization techniques. However, Conditions EAM and ASM averaged much higher than the baseline Condition KAM. There is considerable utility in such automatically-generated documents. It can also be noted that Condition EAM was the best of the conditions, with fully-automatic content selection (Condition ASM is not fully automatic).

Perhaps the most intriguing result of the objective evaluation is that Condition EAA, which uses ASR transcripts, did not deteriorate relative to Condition EAM as much as might have been expected considering the post-questionnaire results. What this seems to demonstrate is that ASR errors were annoying for the user, but that the users were able to look past the errors and still find the relevant information efficiently. Condition EAA scored higher than the baseline

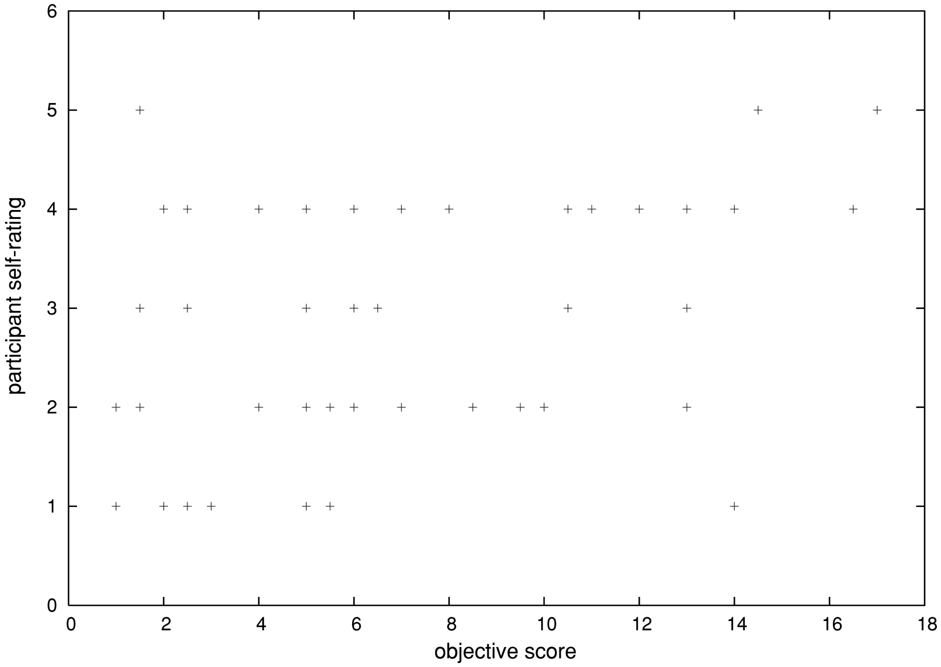


Fig. 2. Objective scores and post-questionnaire scores.

Condition KAM that utilized manual transcripts and not significantly worse than Condition EAM; this is a powerful indicator that summaries of errorful documents are still very valuable. This relates to the previous findings of the SCANMail browser evaluation mentioned in Section 2.3 [Hirschberg et al. 2001; Whittaker et al. 2002], in which participants were able to cope with the noisy ASR data.

To assess the variation in individual performance on this task, we conducted an analysis of variance on the factor "Subject", both within each condition and across conditions. In no case was there a significant main effect on the task evaluation scores. So while individual performance does vary, we can be fairly certain that the participants overall were capable of performing the task and that performance differences are primarily due to the experimental condition.

An interesting question is whether participants' self-ratings on task performance correlated with their actual objective performance according to the human judges. To answer this question, we calculated the correlation between the scores from post-questionnaire Q4 and the objective scores. The statement Q4 from the post-questionnaire is "I feel that I completed the task in its entirety". The result is that there was a moderate but significant positive correlation between participant self-ratings and objective scores (pearson = 0.39, p<0.005).

Figure 2 shows the relationship between the objective ratings and participant self-ratings for all 50 participants. While the positive correlation is evident, an interesting trend is that while there were relatively few people who scored

Table VI. Logfile Feature Results

Feature	KAM	EAM	EAA	AMM	ASM
Q1	45.4	43.1	45.4	45.42	43.2
Q2	16.25	13.9	17.14	8.61	10.22
Q3	0.98	0.81	0.72_{AMM}	1.4^{EAA}	1.13
Q4	$0.39_{\text{EAM,EAA,AMM}}$	0.11^{KAM}	0.08^{KAM}	0.08^{KAM}	0.18
Q5	1.33	2.24	1.47	1.99	0.83
Q6	15.4_{EAA}	14.4_{EAA}	$40.4^{\text{KAM,EAM,AMM}}$	16.6_{EAA}	20.6
Q7	0.03	0.01	0.01	0.01	0.01
Q8	1400	1602	1397	2043	1650
Q9	1251	1384	1161	1760	1430
Q10	3.9	4.0	3.9	4.0	4.0
Q11	0.68	0.73	0.76	0.65	0.65

For each score in the table, that score is significantly worse than the score for any conditions in superscript, and significantly better than the score for any condition in subscript.

highly on the objective evaluation but scored low on the self-ratings, there were a fair number of participants who had a low objective score but rated themselves highly on the post-questionnaire. A challenge with this type of task is that the participant simply may not have had a realistic idea of how much relevant information was out there. After retrieving four or five relevant items, they may have felt that they had completed the task entirely. This result is similar to the finding by Whittaker et al. [2008] in a task-oriented evaluation of a browser for navigating meeting interactions. Participants were asked to answer a general "gist" question and more specific fact questions about a meeting that they could access with a specific meeting browser. There, too, the participants often felt that they performed better than they really had.

5.3 Logfile Results

Table VI gives the results for the logfiles evaluation, formatted analagously to the previous tables. Q1 in the results table refers to the task duration; Q2 is the feature representing the point in the meeting at which the participant began to write the answer; Q3 represents the total amount of tabbing the user did normalized by experiment length; Q4 is the percentage of content buttons clicked normalized by the total number of content buttons; while Q5 is the number of content clicks per minute; Q6 is the number of clicks on the audio/video buttons; Q7 represents how often a content click directly precedes the user moving to the writing tab; Q8 is the total unedited length of the participant's answer; while Q9 is the edited length after any deletion; Q10 is the number of meetings reviewed; and Q11 is the average of the timestamps in the writing tab.

An unexpected result was that the task duration (Q1) did not vary significantly between conditions. Because the task was difficult to complete in 45 minutes, most participants took all or nearly all of the allotted time, regardless of condition.

For many of the logfile features, the impact of the gold-standard Condition AMM is clear: participants in this condition began writing their answer earlier (Q2); did not wait until the end to write the bulk of their answers (Q11); wrote longer answers (Q8); and had more time for editing their answers (Q9).

Condition ASM fares well on these same criteria in comparison with the extractive approaches.

Perhaps the most striking finding from the logfiles analysis is the variation in how participants used the audio/video stream. In Conditions KAM, EAM, ASM, and AMM, the number of media clicks (Q6) averaged around 14–20 per task. For Condition EAA, incorporating ASR, the average number of media clicks was 40.4, significantly higher than all other conditions, with the exception of Condition ASM. Participants in Condition EAA relied much more on audio and video during this task. While they still used the summary dialogue acts to index into the meeting record (Q5), they presumably used the audio and video to disambiguate any ASR errors.

Discussion. It is difficult to derive a single over-arching conclusion from the logfile results, but there were several interesting results on specific logfile features. Perhaps the most interesting was the dramatic difference that existed in terms of relying on the audio/video record when using ASR. This ties together several interesting results from the post-questionnaire data, the human evaluation data, and the logfile data. While the ASR errors seemed to annoy the participants and therefore affected their user satisfaction ratings, they were nonetheless able to employ the ASR-based summaries to locate the relevant information efficiently and thereby scored well according to the human objective evaluation. Once they had indexed into the meeting record, they then relied heavily on the audio/video record, presumably to disambiguate the dialogue act context. It was *not* the case that participants in this condition used only the audio/video record and disregarded the summaries, since they clicked the content items more often than in Conditions KAM and ASM (Q5). Overall, the finding is thus that ASR errors were annoying, but did not obscure the value of the extractive summaries.

It is also interesting that both extractive conditions led to participants needing to move between meeting tabs less than in other conditions. The intuition behind the inclusion of this feature was that a lower number would indicate that the user was finding information efficiently. However, it is surprising that Condition EAA showed the lowest number of tab switches and Condition AMM the highest. It may be the case that participants in Condition AMM felt more free to jump around because navigation was generally easier. A second possible explanation is that extractive summary sentences, in contrast with abstract sentences, form a clear, direct link to the source document sentences, allowing the user to browse the meetings in a more linear, chronological fashion.

Many of the logfile features confirmed that the human abstract gold-standard was difficult to challenge in terms of browsing efficiency. Users in this condition began typing earlier, wrote most of their answer earlier in the task, wrote longer answers, and had more time for editing.

5.4 Extrinsic/Intrinsic Correlation

In order to determine whether available intrinsic evaluation metrics predict the discrepancy in ratings between manual and ASR transcripts, we scored the extractive summaries in both conditions using ROUGE and the weighted

Table VII. Comparison of Extrinsic/Intrinsic Scores for Human and ASR Transcripts

Metric	Man	ASR
Objective	7.2	5.05
PQ4	3.1	2.4
ROUGE-2	0.55	0.41
ROUGE-SU4	0.57	0.47
Weighted F	0.48	0.46

f-score. Table VII shows the results of these intrinsic evaluations along with the objective human results and post-questionnaire statement Q4, "I feel that I completed the task in its entirety." All metrics show a decline on ASR compared with manual transcripts for these four meetings. The difference in scores is most pronounced with ROUGE-2, while the weighted f-score shows the least decline on ASR. This is likely due to the fact that ROUGE evaluations are carried out at the n-gram level, while weighted f-score works only at the dialogue act level. Weighted f-score does not directly take ASR errors into account; the impact of ASR is on whether or not the error-filled dialogue acts are extracted in the first place.

6. GENERAL DISCUSSION

Many of these results are encouraging for the extractive summarization paradigm. Users find extractive summaries to be intuitive, easy-to-use and efficient, are able to employ such documents to locate the relevant information according to human evaluations, and users are able to adapt their browsing strategies to cope with ASR errors. While extractive summaries might be far from what people conceptualize as a meeting summary in terms of traditional meeting minutes, they are intuitive and useful documents in their own right.

To compare abstractive and extractive summaries overall, the main drawback of extracts is not in terms of user satisfaction but in how quickly the relevant information can be retrieved. Conditions AMM and ASM were both superior in terms of participants beginning to write their answers earlier and authoring more comprehensive answers.

The main weakness of the abstractive Condition ASM is in terms of user satisfaction. User satisfaction was generally lower than for Condition EAM, which is somewhat surprising given that the objective scores are slightly higher and that the logfiles indicate a faster retrieval rate. The fact that Condition ASM performs worse than AMM and EAM on Q2, Q9, and Q10 suggests that the semi-automatic abstractive summaries did not contain as much detail as users would have liked.

Perhaps the most interesting result from the decision audit overall is regarding the effect of ASR on carrying out such a complex task. While participants using ASR find the browser to be less intuitive and efficient, they nonetheless feel that they understand the meeting discussions and do not desire additional information sources. In a subjective human evaluation, the quality of the answers in Condition EAA suffers according to most of the criteria, including writing style, but the participants are still able to find many of the relevant pieces of information according to the objective human evaluation. We find that users are able to adapt to errorful transcripts by using the summary dialogue acts as navigation and then relying much more on audio/video for disambiguating the conversation in the dialogue act context. Extractive summaries, even with errorful ASR, are useful tools for such a complex task, particularly when incorporated into a multimedia browser framework.

There is also the possibility of creating browsing interfaces that minimize the user's direct exposure to the ASR transcript. Since we have previously found that ASR does not pose a problem for our summarization algorithms, we could locate the most informative portions of the meeting and present the user with edited audio and video and limited or no textual accompaniment, to give one example.

If the decision audit evaluation is run again in the future, it would be interesting to give participants a longer amount of time to complete the task, as this might yield compelling differences between conditions. As it stands, almost all participants needed the full allotted time. There is also the possibility of exploring the effect of ASR in more detail by artificially varying the word-error rate and determining at which point it begins to become detrimental to performing the task. Additionally, it would be interesting to apply the keywords approach and semi-automatic abstract approach to ASR output and assess their robustness to noisy transcripts.

7. CONCLUSION

We have presented an extrinsic evaluation paradigm for the automatic summarization of spontaneous speech in the meetings domain: a decision audit task. This evaluation scheme models a complex, real-world information need where the relevant information is widely distributed and there is not a simple one-sentence answer. This work represents the largest extrinsic evaluation of speech summarization to date. We found that users considered automatically-generated summaries to be coherent and useful, generally outperforming a keyword baseline. The largely positive results for the extractive conditions in terms of user satisfaction and objective performance justify continued research on this summarization paradigm. However, the considerable superiority of gold-standard abstracts in many respects also support the view that research should begin to try to bridge the gap between extractive and abstractive summarization [Kleinbauer et al. 2007]. These results also indicate that summaries of speech recognition transcripts can be very useful despite considerable noise, particularly when the text is supplemented by, and linked to, the audio/video record.

It is widely accepted in the summarization community that there should be increased reliance on extrinsic measures of summary quality. It is hoped that the decision audit task will be a useful framework for future evaluation work. We believe that this evaluation scheme would be appropriate for various cases where groups are holding ongoing discussions across multiple meetings, and could be generalized to conversations in other modalities such as email. For development purposes, it is certainly the case that intrinsic measures are indispensable: as mentioned before, in this work we use intrinsic measures to evaluate several summarization systems against each other and use extrinsic measures to judge the usefulness of the extractive methods in general. Intrinsic and extrinsic methods should be used hand-in-hand, with the former as a valuable development tool and predictor of usefulness and the latter as a real-world evaluation of the state-of-the-art.

ACKNOWLEDGMENTS

We thank the AMI-ASR team for providing the speech recognition transcripts. Three anonymous reviewers provided very valuable feedback and suggestions.

REFERENCES

ALEXANDERSSON, J. 2003. Hybrid discourse modelling and summarization for a speech-to-speech translation system. Ph.D. dissertation, Universtität des Saarlandes, Germany.

ARONS, B. 1997. Speechskimmer: A system for interactively skimming recorded speech. *ACM Trans. Comput.-Hum. Interact. 4*, 1, 3–38.

CARLETTA, J. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything *Lang. Resour. Eval. 41*, 2, 181–190.

CARLETTA, J., ASHBY, S., BOURBAN, S., FLYNN, M., GUILLEMOT, M., HAIN, T., KADLEC, J., KARAISKOS, V., KRAAIJ, W., KRONENTHAL, M., LATHOUD, G., LINCOLN, M., LISOWSKA, A., MCCOWAN, I., POST, W., REIDSMA, D., AND WELLNER, P. 2006. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*. Lecture Notes in Computer Science, vol. 3869, Springer, Berlin, 28–39.

CHRISTENSEN, H., KOLLURU, B., GOTOH, Y., AND RENALS, S. 2004. From text summarisation to style-specific summarisation for broadcast news. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, vol. 2997, Springer, Berlin, 223–237.

DAUMÉ, H. AND MARCU, D. 2005. Bayesian summarization at DUC and a suggestion for extrinsic evaluation. In *Proceedings of the Document Understanding Conference.*

DEJONG, G. 1982. An overview of the FRUMP system. In *Strategies for Natural Language Processing*, W. G. Lehnert and M. H. Ringle Eds., Lawrence Erlbaum, Mahwah, NJ, 149–176.

DORR, B., MONZ, C., OARD, D., ZAJIC, D., AND SCHWARTZ, R. 2004. Extrinsic evaluation of automatic metrics for summarization. Tech. Rep. LAMP-TR-115,CAR-TR-999,CS-TR-4610,UMIACS-TR-2004-48, University of Maryland, College Park and BBN Technologies.

DORR, B., MONZ, C., PRESIDENT, S., SCHWARTZ, R., AND ZAJIC, D. 2005. A methodology for extrinsic evaluation of text summarization: Does ROUGE correlate? In *Proceedings of the ACL05 Workshop.*

ENDRES-NIGGEMEYER, B. 1998. *Summarizing Information*. Springer, Berlin.

GALLEY, M. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (EMNLP'06). Association for Computational Linguistics, 364–372.

HAHN, U. AND REIMER, U. 1999. Knowledge-based text summarization: Salience and generalization operators for knowledge base abstraction. In *Advances in Automatic Text Summarization*, I. Mani and M. Maybury Eds., MIT Press, Cambridge, MA, 215–232.

HARMAN, D. AND OVER, P. EDS. 2004. *Proceedings of the Document Understanding Conference*.

HIRSCHBERG, J., BACCHIANI, M., HINDLE, D., EISENHOWER, P., ROSENBERG, A., STARK, L., STEAD, L., WHITTAKER, S., AND ZAMCHICK, G. 2001. SCANMail: Browsing and searching speech data by content. In *Proceedings of the 7th European Conference on Speech Communication and Technology*. 1299–1302.

HIRSCHMAN, L., LIGHT, M., AND BRECK, E. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. 325–332.

Hori, C., Furui, S., Malkin, R., Yu, H., and Waibel, A. 2002. Automatic speech summarization applied to english broadcast news speech. In *Proceedings of the International Conference in Acoustics Speech and Signal Processing*. 9–12.

Jing, H., Barzilay, R., McKeown, K., and Elhadad, M. 1998. Summarization evaluation methods: Experiments and analysis. In *Proceedings of the AAAI Symposium on Intelligent Summarization*. 60–68.

Jones, K. S. and Galliers, J. 1995. Evaluating natural language processing systems: An analysis and review. Lecture Notes in Artificial Intelligence, vol. 1083, Springer, Berlin.

Kameyama, M., Kawai, G., and Arima, I. 1996. A real-time system for summarizing human-human spontaneous dialogues. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, Vol. 2, 681–684.

Kleinbauer, T., Becker, S., and Becker, T. 2007. Combining multiple information layers for the automatic generation of indicative meeting abstracts. In *Proceedings of the European Natural Language Generation Workshop*. 151–154.

Kolluru, B., Gotoh, Y., and Christensen, H. 2005. Multi-stage compaction approach to broadcast news summarisation. In *Proceedings of the Interspeech Conference*. 69–72.

Koumpis, K. and Renals, S. 2005. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Trans. Speech Lang. Process. 2*, 1–24.

Kraaij, W. and Post, W. 2006. Task based evaluation of exploratory search systems. In *Proceedings of the SIGIR Workshop, Evaluation Exploratory Search Systems*. ACM, New York, 24–27.

Lin, C.-Y. 2004. Looking for a few good metrics: Automatic summarization evaluation: How many samples are enough. In *Proceedings of the NTCIR-5 Workshop*. 1765–1776.

Lin, C.-Y. and Hovy, E. H. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the HLT-NAACL03 on Text Summerization*. 71–78.

Mani, I. 2001. Summarization evaluation: An overview. In *Proceedings of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*. 77–85.

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., and Sundheim, B. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Proceedings of the EACL'99*. 77–85.

Maskey, S. and Hirschberg, J. 2005. Comparing lexical, acoustic/prosodic, discourse and structural features for speech summarization. In *Proceedings of the Interspeech Conference*. 621–624.

Morris, A., Kasper, G., and Adams, D. 1992. The effects and limitations of automated text condensing on reading comprehension performance. *Inform. Syst. Resear. 3*, 1, 17–35.

Murray, G. and Renals, S. 2007. Term-weighting for summarization of multi-party spoken dialogues. In *Proceedings of the MLMI Conference*. 155–166.

Murray, G., Renals, S., Carletta, J., and Moore, J. 2005. Evaluating automatic summaries of meeting recordings. In *Proceedings of the ACL MTSE Workshop*. 33–40.

Murray, G., Renals, S., Moore, J., and Carletta, J. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the HLT-NAACL Conference*. 367–374.

Nenkova, A. and Passonneau, B. 2004. Evaluating content selection in summarization: The Pyramid method. In *Proceedings of the HLT-NAACL Conference*. 145–152.

Nenkova, A., Passonneau, R., and McKeown, K. 2007. The Pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Comput. Logic 4,* 2, 1–23.

Paice, C. D. and Jones, P. A. 1993. The identification of important concepts in highly structured technical papers. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*. ACM, New York, 69–78.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.

Saggion, H. and Lapalme, G. 2002. Generating indicative-informative summaries with sumum. *Comput. Linguist. 28*, 4, 497–526.

Sparck-Jones, K. 1999. Automatic summarizing: Factors and directions. In *Advances in Automatic Text Summarization*, I. Mani and M. Maybury Eds., MITP, 1–12.

Tucker, S. and Whittaker, S. 2004. Accessing multimodal meeting data: Systems, problems and possibilities. In *Proceedings of the MLMI Conference*. 1–11.

VALENZA, R., ROBINSON, T., HICKEY, M., AND TUCKER, R. 1999. Summarization of spoken audio through information extraction. In *Proceedings. of the ESCA Workshop on Accessing Information in Spoken Audio*. 111–116.

WELLNER, P., FLYNN, M., AND GUILLEMOT, M. 2004. Browsing recorded meetings with Ferret. In *Proceedings of the MLMI Conference*. 12–21.

WELLNER, P., FLYNN, M., TUCKER, S., AND WHITTAKER, S. 2005. A meeting browser evaluation test. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 2021–2024.

WHITTAKER, S., HIRSCHBERG, J., AMENTO, B., STARK, L., BACCHIANI, M., ISENHOUR, P., STEAD, L., ZAMCHICK, G., AND ROSENBERG, A. 2002. Scanmail: A voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 275–282.

WHITTAKER, S., TUCKER, S., SWAMPILLAI, K., AND LABAN, R. 2008. Design and evaluation of systems to support interaction capture and retrieval. *Person. Ubiquit. Comput. 12*, 3, 197–221.

ZECHNER, K. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Comput. Linguis. 28*, 4, 447–485.

ZECHNER, K. AND WAIBEL, A. 2000. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the NAACL Conference*. 186–193.

ZHU, X. AND PENN, G. 2006. Summarization of spontaneous conversations. In *Proceedings of the Interspeech Conference*. 1531–1534.

Received October 2008; revised August 2009; accepted August 2009