*IEEE ICASSP–2003*
*Hong Kong*

# AUDIO INFORMATION ACCESS FROM MEETING ROOMS

*Steve Renals*

University of Sheffield,
Department of Computer Science,
Sheffield S1 4DP, UK
s.renals@dcs.shef.ac.uk

*Dan Ellis*

Columbia University,
Department of Electrical Engineering,
New York NY 10027, USA
dpwe@ee.columbia.edu

**ABSTRACT**

We investigate approaches to accessing information from the streams of audio data that result from multi-channel recordings of meetings. The methods investigated use word-level transcriptions, and information derived from models of speaker activity and speaker turn patterns. Our experiments include spoken document retrieval for meetings, automatic structuring of meetings based on self-similarity matrices of speaker turn patterns and a simple model of speaker activity. Meeting recordings are rich in both lexical and non-lexical information; our results illustrate some novel kinds of analysis made possible by a transcribed corpus of natural meetings.

## 1. INTRODUCTION

Audio information access may concern an archive of spoken documents (eg indexing, retrieval and browsing), a single document (eg segmentation, identifying named entities) or some mixture of the two (eg summarization). Most current approaches to both archive-wide and single-document audio information access are based on the lexical transcription alone ("speech-as-text"), without incorporating complementary non-lexical information such as prosody, dialog turn structure, etc. Spoken document retrieval (SDR) is a successful example of speech-as-text approaches. Evaluation in the broadcast news domain (within the SDR track of the Text Retrieval Conference, TREC) resulted in the important—and surprising—outcome that retrieval performance on speech recognizer output was similar to that obtained using human-generated reference transcripts, with little or no dependence on word error rate (WER) [1].

Accessing information in meetings is more challenging than in domains such as broadcast news for several reasons: Meetings consist of spontaneous overlapping speech across multiple channels; the information content may be thinly spread across a multi-party discussion; and, the desired information required from a meeting recording (eg agreements and disagreements, decision points, actions items) is somewhat different to the types of information extracted from text or broadcast news. Waibel et al [2] have described a meeting browser that operates primarily on the text generated by an automatic speech recognition (ASR) system. This work included a novel summarization system [3] which, in addition to the usual text-based methods, employed methods specific to spontaneous spoken dialog—detection of disfluency/repetition and question-answer pairs. Baron et al [4] used lexical and prosodic features to detect sentence boundaries and disfluency interruption points in meetings. In this paper, we investigate spoken document retrieval, segmentation based on speaker turn patterns, and a model of speaker activity. Additionally, we present some approaches to meeting visualization using both lexical and non-lexical features.

## 2. EXPERIMENTAL DATA

We have used a corpus of meetings that was recorded and annotated by ICSI [5]. Each meeting contained 3–8 participants, who were equipped with close-talking, head-mounted, wireless microphones. In addition there were six far-field microphones: four high quality PZM microphones, and two lower quality microphones (to simulate a possible palmtop device). The meetings were transcribed at the word level, with additional marks for non-speech sounds, non-vocal sounds, emphasis, etc. Additionally, for six meetings, boundaries between topics were annotated.

In this work we used the human generated transcripts, and speech recognition output obtained from the close talking microphones [4] (using the SRI recognizer trained on the Switchboard database of telephone conversations [6]). An automatic segmenter [7] was used to detect regions of speech activity in each channel; however since missed speech regions and crosstalk result in a 10% degradation in WER, hand-corrected segments were used. We used 32 meetings held by ICSI research groups in speech and language, totalling 31.9 hours of multi-channel audio, and about 307 000 transcribed words. The WER was about 45% for native speakers, and 72% for non-native speakers. The out-of-vocabulary (OOV) rate for the recognizer ranged from 2.8% to 6.1% per meeting, averaging 3.9%. Since the recognizer vocabulary was not adapted to the specific domain of these meetings, the OOV list included several important content words such as "Java", "recognizer", "Linux" and "spectral".

## 3. LEXICAL METHODS

### 3.1. Spoken document retrieval

We have ported the spoken document retrieval system that we developed for broadcast news [8] to the meetings domain. In our experiments we investigated indexing both the hand transcripts and the ASR transcripts. Rather than attempting an *a priori* topic segmentation, we indexed each meeting as a sequence of overlapping 30 s excerpts. Each "document" was Porter stemmed, but no stop list was used. The query-document match used Okapi BM25 term weighting [9], which is based on the product of term frequency and inverse document frequency (tf.idf). Excerpts longer than 30 s were obtained in a query-dependent manner, by merging adjacent, potentially relevant excerpts. We experimented using ten queries,

| | Hand Trans | ASR |
|---|---|---|
| P1 | 0.8 | 0.8 |
| P5 | 0.62 | 0.48 |
| P10 | 0.37 | 0.34 |
| R-P | 0.57 | 0.45 |

Table 1: Spoken document retrieval results for the archive of 32 meetings, transcribed by hand and by speech recognizer. Results are given as precision-at-*n*-documents (P*n*) and as *R*-precision (R-P). The results are an average over the ten queries used in the experiment. P10 values are lower than P1 or P5, since most of the queries had fewer than 10 relevant excerpts.

some of which contained words known to be OOV with respect to the recognizer[1]. We evaluated the system using precision-at-$n$[2] and $R$-precision[3]. The experimental results (table 1) indicate that there is some degradation when the hand transcripts are replaced by ASR output, but that retrieval on the ASR transcripts is usable (eg the top ranked document was relevant 8 out of 10 times). The R-precision on hand transcripts was greater than on the ASR transcripts on 5 out of 10 queries, the same on 3 out of 10, and lower on the remaining 2 queries.

The SDR system returned an excerpt as a time index to a meeting. An excerpt was marked as relevant if the time index fell within a meeting segment manually marked as relevant. This is the same criterion used in the TREC SDR track. Other returned time marks corresponding to the same segment were counted as non-relevant. A weakness of this approach, that becomes quite apparent when dealing with meeting data, is that the duration of relevant excerpts is not explicitly considered. For some queries, a relevant segment may be a complete 60 minute meeting, for others a 1 minute piece, but both cases are evaluated as a single time mark, with no information about segment duration. The issue of accurately segmenting relevant topics is also problematic, complicated by the existence of non-relevant sub-segments within a relevant segment (more generally, topic nesting).

### 3.2. Visualization

We have been exploring methods of visualizing the topic structure within meetings. Foote [10] has used self-similarity matrices for the visualization of music and video, and we have applied this approach to meeting transcripts. The left of figure 1 shows a similarity matrix for a 60 minute meeting: each cell $(i, j)$ of the matrix represents the similarity of minute $i$ of the meeting to minute $j$ of the meeting. We used tf.idf as a similarity measure, with the idf of term $t$ estimated as $\mathrm{idf}(t) = \log[N/n(t)]$, where there are a total of $N$ non-overlapping extracts of 60s in the archive, of which $n(t)$ contain term $t$. The block diagonal structure may be interpreted as chunks of lexical similarity. Off diagonal blocks may be interpreted as returns to a "topic" previously discussed. As before, words were stemmed; also, a 268 word stop list was applied which, in addition to the usual function words, contained filled pauses and disfluencies. Without using a stop list the structure tended to be speaker-specific—a "style" similarity—based on characteristic filled pauses, etc., rather than the desired content similarity. In the next section we discuss self-similarity matrices based on speaker turn patterns.

[1]Example queries included "room acoustics reverberation", "SpeechDAT car", and "Bayesian belief networks".

[2]The retrieval precision considering the $n$ top ranked documents.

[3]The precision at $R$, when there is a total of $R$ documents relevant to the query. In this case, we estimated recall by manual use of the SDR system.

## 4. NON-LEXICAL METHODS

Recorded meetings contain much more information than just the word sequences. One rich source is the pattern of speaker turns: many meetings have frequent and rapid alternations among speakers, and these patterns can vary significantly throughout and between meetings, indicating different episodes or modes of discussion. Figure 2 shows an example of the turn patterns for a complete meeting of 6 participants, clearly suggesting a level of significant structure.

Our preliminary investigations into extracting this information consist of modelling the structure *within* a given meeting by finding segments whose speaker transitions are approximately constant, and also looking for variations *between* meetings to characterize how each participant behaves in each meeting

### 4.1. Speaker turn pattern segmentation

Some meetings may be viewed as a series of component discussions between different subsets of the participants, perhaps corresponding to different topics being discussed. If this were true, one approach to topic segmentation would be to look at the patterns of speaker turns, without considering what was being said, and place boundaries between episodes that consisted of different sets of speakers, or different patterns of dialog.

We attempted to make such a segmentation using the Bayesian Information Criterion (BIC), analogous to acoustic speaker segmentation [11]. In this procedure, every possible division of a given segment of the signal is tested to see if the likelihood gain from using separate models for each half is large enough to justify the additional parameters expended over using a single model for the entire segment, according to the following test:

$$\log L(X_1; M_1) + \log L(X_2; M_2) \gtrless \log L(X; M_0) + \frac{\lambda}{2} \log(N)\#(M)$$

where $\log L(X; M_0)$ is the log-likelihood of the entire data segment under a single model, to be compared against $L(X_1; M_1)$ and $L(X_2; M_2)$, the likelihoods of the two parts under separate models. $N$ is the total number of data points, $\#(M)$ is the number of parameters in each model (i.e. the additional parameter count in the two-model explanation), and $\lambda$ is a weighting constant theoretically equal to 1.

Our likelihoods were based on the following model: A meeting was reduced to a sequence of dominant speakers within each quarter-second window; overlaps were resolved by preferring the speaker who started most recently. For a given stretch of time, this speaker index sequence was modelled with an $S \times S$ transition matrix, where $S$ is the number of speakers, and the transition probabilities (including self-loops) were based on simple counts. Thus, any segment can be described by its transition matrix. The right of figure 1 shows the self-similarity matrix between every minute of an example meeting as the symmetric Kullback-Leibler (KL) distance between the transition matrices.

The overall likelihood of a segment was the product of the probabilities from the matrix for each transition and self-loop within the segment. Leaving $\lambda$ at 1 but using the relatively large

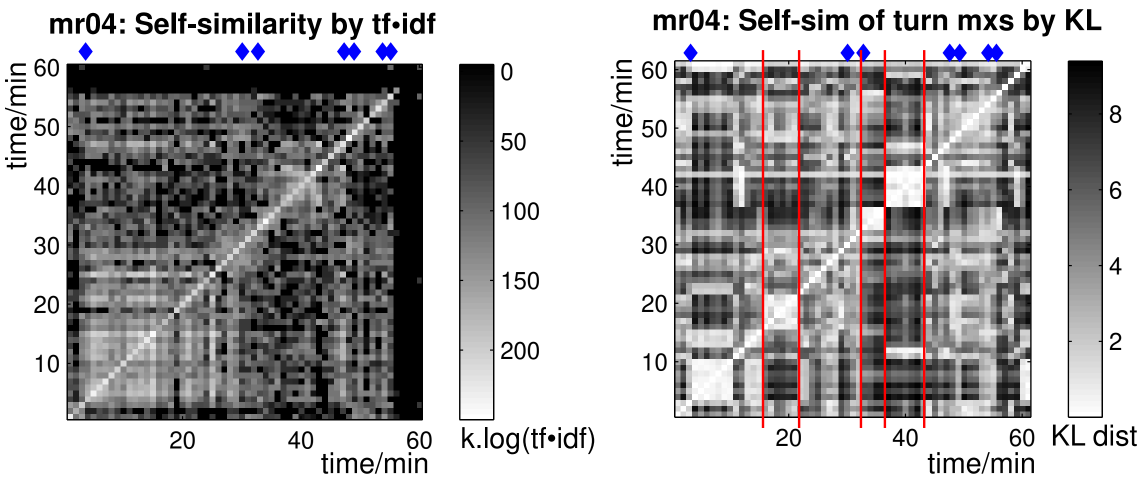


Figure 1: Visualization using self-similarity matrices for an example meeting, derived from a hand transcription. Cell $(i, j)$ represents the similarity of minute $i$ with minute $j$, bright regions indicating high self-similarity. The similarity measure in the left matrix was a log-scaled, normalized tf.idf score; the right matrix used the KL distance between speaker transition matrices. Diamonds indicate manually assigned topic boundaries, vertical lines the boundaries found by BIC segmentation.

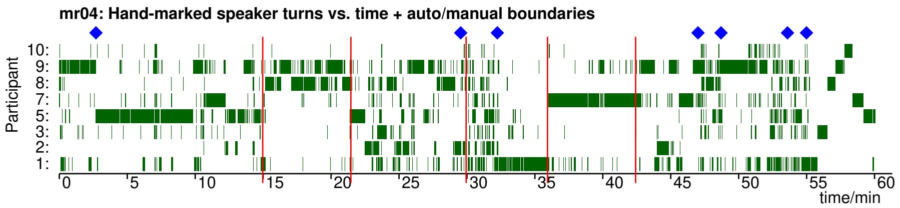


Figure 2: Patterns of speaker turns within a 60 minute meeting. Each row corresponds to a different participant. The meeting ended with most participants separately reading a series of digits (56 minutes onwards). Vertical lines indicate boundaries from BIC segmentation; blue diamonds show the hand-marked topic boundaries for this meeting.

parameter count of a second $S \times S$ matrix yielded around 4 to 10 segments per meeting, as shown by the vertical lines in figures 1 and 2.

The BIC segmentation structures a meeting according to speaker turn patterns. We compared this segmentation with the manual topic segmentations (an example is shown by the diamonds in figure 2). Of the 36 manually-marked topic boundaries over 6 meetings, turn-based segmentation agreed with only 15 (42%) to within 2 minutes; in addition, 16 turn-based boundaries were found that had no corresponding topic boundary. Thus it seems that turn-pattern boundaries are not directly related to discussion topics, although they may provide an important alternative perspective on the temporal structure of meetings.

### 4.2. Modelling speaker "talkativity"

The proportion of a given meeting filled by each participant varies extensively among the participants, in line with our informal sense that some people are more 'talkative' than others. At the same time, we expect that certain participants will speak more in some meetings than in others, depending on their interest in the topics being discussed as well as other factors such as competition for the floor from more talkative colleagues. This is a potentially useful basis for indexing, to answer questions like "which was the meeting when A and B were so vocal?". However, there is a question over how to separate the baseline "talkativity" of each speaker from their particular performance in a single meeting.

The ICSI meeting corpus contains several sets of regular meetings which are composed of a roughly constant set of participants recorded approximately once a week. We focused on the MR set–meetings that discussed the meeting recorder project itself. This set consists of 26 meetings with an average duration of 48.8 minutes; the meetings were recorded over a 9 month period. Ten speakers participated in at least six of these meetings; the average number of meetings per participant was 18.1 and the average number of these participants in each meeting was 6.9. For each meeting, we calculated the proportion of the total meeting time during which each participant was speaking. This data is illustrated in the upper panel of figure 3, where each row is a participant and each column is a meeting; a white cell indicates that a participant was not present in a particular meeting, whereas a gray level shows how much of that meeting was taken up with that participant's speech.

To factor this into baseline talkativity and meeting-specific variability, we fit the following model: Each speaker $s$ has an innate 'talkativity index' $T_s$; in a given meeting $m$ consisting of a set of speakers $S_m$, the predicted proportion of the meeting time 'occupied' by speaker $s$ is:

$$P_{sm} = \frac{T_s}{\sum_{t \in S_m} T_t}$$

i.e. each participant's speech expands to fill the available time in the meeting in proportion to their innate talkativity.

We fit this model to the recorded meeting data by initializing the $T_s$ values to the global proportion of speech for each participant (which does not account for the fact that some people attended more meetings than others), then iteratively re-estimated the $T_s$ for each participant with:

$$T_s = avg_{m \in M_s} \frac{P_{sm} \sum_{t \in S_m, t \neq s} T_t}{1 - P_{sm}}$$

where $M_s$ is the set of meetings in which speaker $s$ participated. This converged very rapidly to the marginal values shown on the left of the upper panel of figure 3. Given this model, we can then compare the actual proportion of the each meeting occupied by each participant with the predicted values; these ratios then give us our 'factored' activity level of a particular speaker in a particular meeting, with innately talkative speakers suitably discounted. This is shown in the lower panel of figure3.

The results indicate that the model can capture effects such as a speaker dominating a particular meeting (eg talker 2 speaking for 46% of meeting 16, in the absence of talkers 8 and 9). However, there are some large deviations from the model, some of which are clear from the transcript (eg new talker 6 giving an introductory presentation in meeting 26).

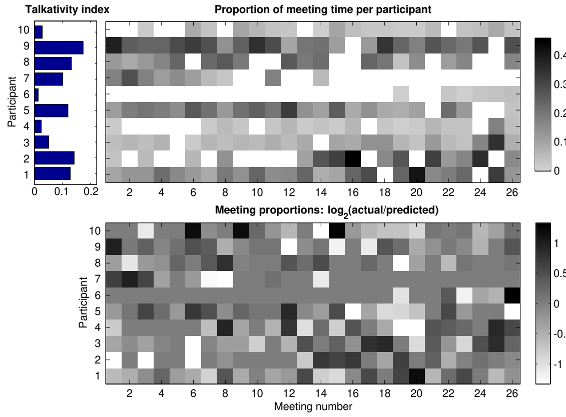


Figure 3: Analysis of proportions of the meeting time taken up by each participant's speech for a collection of 10 participants (rows) who were present in 26 meetings (over a 9 month period) (columns). Top pane shows the raw proportion of each meeting occupied by each participant; white cells indicate that a particular speaker was absent; participant 7 left the project half way through, and participant 6 joined a couple of months to the end of the recording period. On the left of the top pane are the estimated 'talkativity' indices for each speaker, obtained by fitting the simple linear model to the data. Lower pane shows the log-ratio of actual speaking proportions relative to the predictions of the linear model. Thus, even though participant 6 occupied about the same proportions in meetings 25 and 26, the added competition from 'talkative' speakers 5, 8 and 9 in meeting 26 make participant 6's actual-to-predicted ratio much larger in that meeting.

## 5. CONCLUSIONS

Meeting recordings contain rich information, in both lexical and non-lexical forms. Spoken document retrieval ports well to this domain, even in the presence of a relatively high word error rate. We are investigating structures based on interactions within a meeting using models of turn structure and speaker activity. Although our results are preliminary, they illustrate some novel kinds of analysis made possible with such a transcribed corpus of natural meetings. Our current investigations include more powerful models of speaker activity (based on a notion of variable speaker rate), and their relation to other higher level events, such as dispute and decisions.

## ACKNOWLEDGMENTS

We would like to thank members of the ICSI meeting project for much help and discussion. The speech recognition transcriptions were provided by Don Baron, Andreas Stolcke and Elizabeth Shriberg, and the manual topic boundaries were provided by Michel Galley. This work was supported by the European Union IST project *M4* (IST-2001-34485) and US NSF project *Mapping Meetings* (IIS-0121396).

## REFERENCES

[1] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC spoken document retrieval track: A success story. In *Proc. RIAO 2000*, 2000.

[2] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. In *Proc IEEE ICASSP*, 2001.

[3] K. Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proc. ACM SIGIR*, 2001.

[4] D. Baron, E. Shriberg, and A. Stolcke. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *Proc. ICSLP*, 2002.

[5] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Proc. HLT*, pages 246–252, 2001.

[6] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng. The SRI March 2000 Hub-5 conversational speech transcription system. In *Proc. NIST Speech Transcription Workshop*, 2000.

[7] T. Pfau, D. Ellis, and A. Stolcke. Multispeaker speech activity detection for the ICSI Meeting Recorder. In *Proc. IEEE ASRU Workshop*, 2001.

[8] S. Renals, D. Abberley, D. Kirby, and T. Robinson. Indexing and retrieval of broadcast news. *Speech Communication*, 32:5–20, 2000.

[9] S. E. Robertson and K. Spärck Jones. Simple proven approaches to text retrieval. Technical Report TR356, Cambridge University Computer Laboratory, 1997.

[10] J. Foote. Visualizing music and audio using self-similarity. In *Proc. ACM Multimedia*, pages 77–80, 1999.

[11] S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In *Proc. DARPA Broadcast News Workshop*, 1998.