Towards Methods for Efficient Access to Spoken Content in the AMI Corpus

Gareth J. F. Jones
Centre for Digital Video Processing
School of Computing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

Maria Eskevich
Centre for Digital Video Processing
School of Computing
Dublin City University
Dublin 9, Ireland
meskevich@computing.dcu.ie

Ágnes Gyarmati
Centre for Digital Video Processing
School of Computing
Dublin City University
Dublin 9, Ireland
agyarmati@computing.dcu.ie

ABSTRACT

Increasing amounts of informal spoken content are being collected. This material does not have clearly defined document forms either in terms of structure or topical content, e.g. recordings of meetings, lectures and personal data sources. Automated search of this content poses challenges beyond retrieval of defined documents, including definition of search items and location of relevant content within them. While most existing work on speech search focused on clearly defined document units, in this paper we describe our initial investigation into search of meeting content using the AMI meeting collection. Manual and automated transcripts of meetings are first automatically segmented into topical units. A known-item search task is then performed using presentation slides from the meetings as search queries to locate relevant sections of the meetings. Query slides were selected corresponding to well recognised and poorly recognised spoken content, and randomly selected slides. Experimental results show that relevant items can be located with reasonable accuracy using a standard information retrieval approach, and that there is a clear relationship between automatic transcription accuracy and retrieval effectiveness.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital LibrariesI.2Artificial IntelligenceI.2.7 Natural Language Processing [Speech recognition and synthesis]

General Terms

Measurement, Experimentation

Keywords

Speech search, information retrieval, automatic speech recognition

.

1. INTRODUCTION

Increasing amounts of spoken content are being captured and archived from a wide variety of sources. If this material is to realise its full potential value, effective automated search is required to enable users to access relevant content in an efficient manner. The appropriate way to index and search such spoken content depends on its form and nature. While searching a collection of well defined spoken documents clearly articulated in a spoken environment for which high accuracy automated transcripts can be derived, can generally be handled in the same manner as a standard text document retrieval task [4], other sources of spoken content pose much greater challenges presenting a range of potential barriers to effective automated search. For example, the content may be informally structured meaning that retrieval units cannot be easily defined, it may be casually spoken in a noisy environment with significant amounts of cross-talk between multiple speakers often greatly reducing the transcription accuracy of automated speech recognition, and the actual spoken content may assume much knowledge of the subject under discussion meaning that there is an absence of content words to facilitate reliable retrieval.

Our project IISSCoS[1] is focused on the development of methods to improve search quality for challenging spoken data sources incorporating poor speech quality, lack of structure and informal content. In order to do this we are first seeking to fully understand the extent and impact of these issues on retrieval effectiveness and establish search baselines against which we can demonstrate the effectiveness of novel techniques as we develop them.

In this paper we focus on the task of searching recordings of meetings based on a study using the AMI corpus [2]. Search of meetings is an interesting task for spontaneous speech search since it incorporates all the issues highlighted above. Content may be spoken in a wide range of often informal spontaneous styles, topic boundaries will generally not be clearly defined, and participants frequently know that others in the meeting are fully cognisant of the topic under discussion. Additionally meetings are often very long, covering multiple topics meaning that identifying meaningful search units within them is needed to facilitate fine granularity retrieval and content access efficiency. In our experiments we describe an initial investigation into content segmentation and retrieval effectiveness comparing behaviour of manually and automatically generated transcripts from

[1] `http://www.cdvp.dcu.ie/IISSCoS/`

the AMI corpus. Experimental results demonstrate significant differences between segmentation behaviour for manual and automatic transcripts. We also show that word recognition accuracy within segments affects retrieval accuracy in a known-item search task. This work establishes a framework for us to begin exploration of novel methods for tasks such as meeting search.

This paper is structured as follows: section 2 reviews relevant existing work in spoken content search, section 3 outlines the task of searching audio recordings of meetings and the features of the AMI corpus, section 4 describes use of the AMI corpus in developing our search task, section 5 gives results and analysis of our initial investigation of search of the AMI corpus, and finally section 6 concludes and outlines directions for our future work.

2. BACKGROUND

In this section we briefly review some key results of existing work in speech search and the relationship between recognition word error rate (WER) and retrieval effectiveness.

2.1 Existing Research in Speech Search

Existing speech retrieval research has predominantly focused on search for relevant spoken contents where the retrieval units are clearly defined document units. This work is probably best exemplified by the Spoken Document Retrieval (SDR) task at TREC in the late 1990s [4]. In this task TV and radio news broadcasts were manually segmented into carefully defined story units prior to retrieval. While the SDR track had an unknown story boundaries condition to explore search without fixed story boundaries, the source of this data meant that the underlying content was explicitly divided into story units. Examination of the results of the TREC SDR tracks declared speech search a largely solved problem [4].

More recent work has explored search of less formally structured and spoken material. One example is work investigating search of the Malach collection [8]. This consists of interviews with survivors and witnesses of the Holocaust [1]. Interviews were manually divided into meaningful segments, and were augmented to include a number of pieces of manually and automatically generated metadata for each "document" unit, providing description of the spoken content. Experiments showed that even with ongoing improvements in speech recognition accuracy, speech search for more complex speech sources such as this still presents significant challenges. While the spoken content itself may not always be sufficient to enable highly effective search without augmentation with additional metadata, retrieval effectiveness is greatly enhanced by including indexing of the manually generated metadata and marginally so using the automatically generated metadata.

To the best of our knowledge work to date has not explored search of multi-topic discursive material with automated segmentation.

2.2 Recognition Word Error Rate and Retrieval

An important component for any speech search application is automatic speech recognition (ASR). The cost of manual transcription means that in practice speech search must rely on automated indexing methods. We are interested in exploring the relationship between ASR accuracy and the effectiveness of speech search for data such as recordings of meetings. The basic relationship between average transcription quality and accurate (or near accurate) transcription is reported in most speech retrieval studies. For example, the TREC SDR track illustrated how the relatively low recognition error rates on the radio and TV news material used for these studies resulted in little loss in retrieval effectiveness [4].

A more interesting and careful examination of the differences in retrieval behaviour of documents with different speech transcription accuracy levels for the results of the TREC 7 SDR task is described in [11], [12]. The analysis of the distribution of the errors shows a general tendency for documents with low WERs to be retrieved at higher ranks, independent of document relevance to the search query.

The impact of the errors according to their types was measured using different quality metrics on the level of the document or on the level of the whole collection. Such metrics as Named Entity WER and Named Entity Mean Story WER for Cross-Recognizer Results have shown the best correlation with retrieval performance [5]. The global semantic distortion metric based on the vector space model and focusing on various types of substitutions (frequent vs infrequent, semantically similar vs dissimilar) revealed a higher impact of the infrequent and semantically dissimilar substitution errors [7].

3. MEETING SEARCH

3.1 Searching in a Meeting Corpus

As outlined in Section 1, recordings of meetings are a challenging speech search environment encapsulating many of the challenges of search of informal unstructured content. The most simple search scenario would be simply to take transcripts of complete meetings and use these as the search unit. However, since meetings are often very long, lasting anything from a few minutes to several hours, and will often cover many topics, some related and some very distinct, it is more sensible to think in terms of breaking them into smaller focused units and use these as the search unit. We can then hopefully retrieve relevant search units and direct the user effectively to this content.

Traditionally one participant in a formal meeting is assigned to take minutes which summarise the activities and conclusions of the meeting, in the case of more informal meetings there is often no record of the proceedings. Even when taken, minutes often record only the key elements of the discussions and decisions reached, as understood at the time of the meeting by the person taking the minutes. Thus, minutes may be deficient if the minute taker misunderstands some elements of the discussion or the future significance of some part of the meeting is not apparent to the participants and no record is kept. When recordings are made of meetings, participants and others can potentially play back parts of a meeting to access specific details or to revisit how a decision was made. Finding the right meeting or part of a specific meeting will be very inefficient if undertaken manually, thus effective automatic search has great potential value for use of recordings of meetings. Since a key element of searching meetings is identifying suitable search units, in addition to automatic transcription using ASR, some means of segmenting the resulting transcripts is required prior to search.

Meetings can be searched using standard interactive manner by users posing queries to find relevant content. An alternative, as introduced by Popescu-Belis et. al. is *query-free* or *just-in-time* retrieval (the AMIDA Automatic Content Linking Device) [9]. This system transcribes an on-going meeting automatically, and uses the transcript to perform searches of previous meetings and additional material at regular intervals. Another possible scenario is to use query-based offline search, to gather additional information and/or find the relevant (parts of) meetings where a certain topic was being discussed. In this paper, we adopt the latter scenario. Thus, relevant content from previous meetings can dynamically be presented to participants. This is an interesting mode of use since it means that material which may have been forgotten about, possibly since it was not regarded as of interest during previous meetings, can be made available during a later discussion. This may result in greater efficiency in avoiding repeating arguments, insights from previous discussions and ultimately potentially better decision making.

3.2 The AMI Corpus

Investigation of meeting search requires a suitably rich and well planned experimental dataset. Construction of such a dataset is a complex and expensive process requiring not only the actual planning and conduct of the meetings to be recorded, but also if an analysis is to be made of the impact of speech recognition errors on search effectiveness, a very expensive full accurate manual transcription of the meetings is needed. Considering these factors our current experiments are carried out on the AMI corpus, collected as part of the AMI project and made publicly available for research purposes.

The AMI Corpus[2] contains 100 hours of annotated recordings of meetings [2]. Meetings last about 30 minutes each, 70% of them simulate a project meeting on product design. Meetings usually involve 4 participants, and were recorded using 6 cameras and 12 microphones: 1 headset microphone for each speaker, and an 8-element circular microphone array. For the majority of the meetings, both manual and automatic transcripts are provided, for the latter the developer of the corpus created a system that “makes use of a standard ASR framework employing hidden Markov model (HMM) based acoustic modeling and n-gram based language models (LMs)” [10]. The dataset also includes several types of additional material, eg. slides projected in scenario meetings, e-mails related to meetings, handwritten notes, and also various sets of annotations, from facial gestures to topic segmentation. In this study we used the AMI automatically-derived annotations release 1.4

In the next section we describe our initial use of the AMI corpus for our investigations of meeting search.

4. PROCESSING THE AMI DATASET

For our current investigation we are interested in the scenario of a meeting participant wanting to find locations in a meeting (or potentially multiple meetings) where the topic of a PowerPoint slide used in the meeting was being discussed. For this initial study we are only interested in the location where the slide was presented associated with a specific spoken segment. The slide may also have been presented elsewhere in the same meeting or in another meeting. Also the topic covered in the slide may have been discussed at other locations in meetings when the slide was not being displayed, and at which the searcher may or may not have been present. However, for these initial experiments we consider only content associated with a specific projection of the slide as described in section 5.1. In our experiments we make use of the manual and automatic transcripts provided with the AMI corpus along with slides taken from the set of slides provided with the corpus.

In this section we describe the preprocessing steps applied to the supplied meeting datasets to form the search collection for this investigation.

4.1 Basic Corpus Pre-Processing

Transcripts of the meetings in the AMI corpus are published separately for each speaker. We automatically merged the per speaker transcriptions using the time marking data to form a single transcript file for each meeting. In these cases due to technical problems, the ASR transcript for one or more speakers of some meetings is missing. In our research, we omitted incompletely transcribed meetings, and used only the fully transcribed ones in this study, since we wished to work with only complete meeting transcripts. This left us with a total of 160 meetings for our experiments.

4.2 The AMI Corpus and Transcription Accuracy

The length of the manual and ASR transcripts are slightly different since the speech recognition system was run over all the audio data, including for example regions before meetings commenced where microphones were being checked, which were ignored by manual transcribers. Since we wished to investigate the effect of ASR quality on retrieval, we carefully aligned the two transcripts using the timestamp information. We were then able to count the total number of correctly recognized words for the whole meeting (ranging 2–85%, average 71%), and for separate files for each speaker taking part in the meeting (ranging 2–89%, average 70%). Word recognition rate (WRR) was calculated simply by comparing words in the manual and ASR transcripts and calculating the proportion recognised correctly by the ASR system. Further we used the same technique to count the recognition rate for topical segments into which the meetings were divided as described in the next section. This resulted in a list of segments ranked by WRR. Since for this experiment we were not focusing on the influence of special types of errors, we simply ranked files according to the number of correctly recognized words in relation to the total number of words in the segment. Thus we were able to prepare a list of segments ranked by word recognition rate.

4.3 The AMI Corpus and Segmentation

The meetings in the AMI corpus are of approximately 30-minutes in length. The provided AMI collection already contains manually created topic segmentations of the transcript. Topics and subtopics form a hierarchical structure, and labels have been assigned by annotators choosing from a list of suggestions. This topic segmentation was made based on the manual transcripts, but is provided only for a subset of the meetings (139 out of the total of 173 provided).

Since we wished to use as many meetings as possible for our experiments, and our long-term goal is to use speech data for retrieval in cases where manual transcription and

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

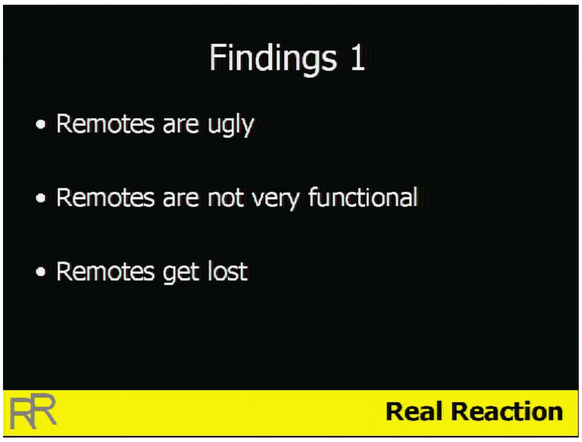
segmentation are not available, we decided to automatically segment the AMI meeting transcripts ourselves. Hsueh and Moore applied and compared two approaches to segment the AMI Corpus [6], an unsupervised algorithm based on lexical cohesion, and a supervised classification. However, since we do not need additional information about the topical hierarchy for indexing and retrieval purposes, we performed linear segmentation using Choi's C99 algorithm [3]. The C99 algorithm works with the fundamental unit of the sentence placing segment boundaries between the end of one sentence and the start of the next. The manual transcripts had punctuation and capitalisation included and could thus be processed directly by the C99 algorithm. However, the ASR transcripts do not include the sentence boundary punctuation marks required. In order to make these suitable for processing by the C99 algorithm, full stops were inserted before the first word after a period of silence, indicated by uppercase in the ASR transcripts.

In order to be able to examine the difference between ASR and manual transcription in another dimension, we applied the segmenting algorithm to both of them. The total number of segments for our 160 meetings testset was 3831 for the ASR transcripts, and noticeably less, 2678, for the manual ones. This yielded an average word count per segment of approximately 221 and 320 respectively for the two transcripts. Direct comparison of the contents and retrieval behaviour of the two segmented collections is not meaningful with different segmentation points and numbers of segments, thus we projected the segment borders of each source onto the other text with the use of the word timing information. This resulted in four different segment sets: ASR transcripts with automatic segmentation (referred to as `aut-aut`), manual transcripts with automatic segmentation (`man-man`), ASR transcripts with segment borders projected from `man-man` (`aut-man`), and manual transcripts with segment borders corresponding to the borders of `aut-aut` (`man-aut`). As outlined earlier the manual transcripts do not cover the whole region of the ASR transcripts since they do not include areas regarded as not relevant to the meetings by the transcribers, in order to allow for this, the additional words in the ASR transcript were placed in the adjoining manual segment.

We also generated simple segmentations of the manual and ASR transcripts based only on timing information. Segment boundaries were placed at regular intervals of 90 seconds, this time interval being chosen as roughly the average length of the automatic segmentation based on the transcribed contents. The boundary points were applied with flexibility to prevent words at the boundaries being split between segments. The different starting points of the two transcripts described previously mean that the time based segmentations are slightly different. We refer to these two segmentations as `aut-time` and `man-time`.

5. EXPERIMENTAL INVESTIGATION

In this section we describe our initial retrieval experiments carried out using our processed AMI collection. We performed segment indexing and retrieval using the `lemur`[3] `Indri` language model toolkit. We first describe the design of our search test collection and then report experimental results.



Findings 1

- Remotes are ugly
- Remotes are not very functional
- Remotes get lost

Real Reaction

Figure 1: A sample slide (`TS3003b.258.51__303.34`)

5.1 Search Test Collection

To create a search task as outlined earlier, we used slides taken from the PowerPoint presentations provided with the AMI corpus as search queries. The objective was for each selected slide to find the point(s) in a specific meeting where the slide was being projected. A key assumption in this being a meaningful retrieval experiment was that the slide was being discussed when it was projected. Using the slides and corresponding segment(s) in this way formed a search task which we are using as a starting point for our research programme examining meeting search.

Figure 1 shows an example slide provided with the AMI corpus. Slides are usually available in two different formats in the AMI dataset, as `ppt` files and also as screenshots generated when they were projected in a meeting. The `jpg` pictures carry additional information in their filenames: the ID of the meeting in which the screenshot was captured, and also a time range within the meeting indicating when exactly the given slide was projected (Eg. TS3003b.258.51__303.34 (Figure 1)). We used the latter piece of information when selecting the slides to be used as queries[4].

Since we are interested in the relationship between word recognition rate (WRR) and retrieval effectiveness for the meeting segments, we created three sets of queries based on ASR word recognition rate of the segments. For two query sets we selected slides whose timestamps referred to segments in the top and bottom of the list of segments ranked by WRR (`max` and `min` respectively). Working down the list from "worst" recognised segments (ignoring the very worst segments where no words were recognised correctly), we located 14 slides for use as queries in the `min` set down to a rank of 56 segments. The WRR of these segments was between 23–44%. Working from the"best" slides we located 24 query slides with WRR ranging from 90–99%. The third set of 25 query slides was chosen randomly (`random`). The WRR for the corresponding relevant random segments was between 64–89%. The average lengths of the queries were 17.2 words,

[3] `http://www.lemurproject.org/`

[4] We realise that since in fact the position in the meeting where the slide is displayed is known that the search task could be seen as unnecessary, for the purpose of our experiment we are of course assuming that this information is not known and that we wish to identify the specific region in the meeting where the topic described in the slide was under discussion.

26.4 words and 27.7 words respectively with ranges 5-30, 1-69 and 2-115 words.

In the retrieval experiments all segments spanning fully or partially over the time range in which a slide was displayed were considered as relevant documents.Appropriate versions of the `qrel` relevance files (as required for the automatic evaluation tool `trec_eval`[5]) corresponding to the different segment borders of the manual and ASR transcripts were generated. Due to the different segmentation boundaries, the number of relevant documents (i.e. segments) varies for the same query across the collections. Experimentally we treated this as a known-item search since we are seeking the single region when the slide was projected, although in some cases two or more segments are in fact be marked as relevant for the slide in the qrel file.

5.2 Experimental Results

Tables 1, 2, 3 and 4 show results for all the differently segmented versions of the corpus. Each table shows results for the three sets of queries (`min`, `max` and `random`). Results shown are Recall and Mean Reciprocal Rank (MRR) at ranked cutoff of 100 and 1000 documents. The MRR was calculated based on the first relevant segment found from the top of the ranked list. As described in the previous section, it can be noted that the number and location of relevant documents for each set of queries varies for each different segmentation of the transcripts.

By comparing the `aut-*` and `man-*` columns in Tables 1 amd 2, we can see that in general the manual transcripts perform better with respect to both Recall and MRR, although perhaps surprisingly there are instances where the ASR transcript performs better. Comparing results for the `min`, `max` and `random` query slides, it can be seen that in all cases the `min` queries perform worse for the ASR transcripts, while for the `max` queries the ASR performs better for the automated segmentations based on the manual transcripts, but not for those based on segments derived from the ASR transcripts. While direct comparison of Tables 3 and 4 cannot be made since the segmentation boundaries are slightly different, a general trend can be seen that in the `min` case retrieval effectiveness is much lower for ASR transcripts, while results are similar for `max` queries. These results are consistent with the findings in [11] that search items with better WRR are likely to be retrieved at higher ranks.

Comparing Recall results at ranks of 100 and 1000, it can be seen that many relevant items are found at ranks below 100, indicating that this retrieval task is actually rather challenging. While this task is a good starting point for our investigations, in order to better understand the retrieval behaviour, in further work we plan to perform manual relevance assessment to generate complete `qrel` files.

Initial analysis of the ranked lists reveals that the retrieval algorithm clearly appears to be favouring longer documents (segments) for this task. This is clearly seen through comparison of the average length of the top 100 and 1000 retrieved documents, the former being on average between 26-47% longer for the various test conditions. Examining individual queries where the ASR transcript performed better than the manual one did not show any consistent trend in terms of WRR or the length of the segments. There is though possibly a complex relationship between these factors and those for other segments that might be more highly ranked for manual transcripts, but where low WRR for ASR transcripts may be reducing their rank to the benefit of (possibly) better recognised relevant segments.

Comparison of the segmentation between the manual and auto transcripts showed that while the total number of segments for the manual segments is somewhat lower than those derived from the ASR transcripts, this does not mean that the manually derived segments are always longer. There is a general trend in this direction, however there are examples of situations where manual transcripts are divided into a number of segments where only a single one is generated from the ASR transcripts.

6. CONCLUSIONS AND FUTURE WORK

The search task described in this paper is an initial step in our plans to explore search of recorded meeting collections such as the AMI Corpus. These initial experiments have demonstrated that we are able to use slides taken from a meeting to locate positions where they appeared based on the transcription of the spoken content. The effectiveness of this search has been shown to be related to the speech recognition accuracy of the transcripts.

There are several immediate next steps in this work. The first is to construct proper segment-based relevance sets for the slide derived search topics. Slides can be re-used in multiple meetings and the same topic may be discussed without use of the slides. We also plan to extend our retrieval task to search topics collected from searchers, although this will require them to be instructed in the subject matter contained in the meeting dataset. Also differences in segment borders for the ASR and manual transcripts demand further analysis of what types of speech recognition errors result in variation in segmentation points and independently how segmentation points should best be chosen to optimise retrieval effectiveness. This raises questions such as what are the possible ways to adjust the ASR transcripts in order to improve the segmentation and retrieval performance, for example by augmenting them with related metadata, such as used in the Malach collection search task [8]

7. ACKNOWLEDGEMENTS

This work is funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2008 Grant No: 08/RFP/CMS1677.

8. REFERENCES

[1] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajiąc, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4):420–435, 2004.

[2] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.

[3] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33, 2000.

[5] http://trec.nist.gov/trec_eval/

		`min`, queries: 15 relevant: 56		`max`, queries: 24 relevant: 68		`random`, queries: 25 relevant: 36	
cutoff		`aut-aut`	`man-aut`	`aut-aut`	`man-aut`	`aut-aut`	`man-aut`
100	Recall	0.1848	0.2515	0.2618	0.2618	0.3600	0.4000
	MRR	0.0847	0.2156	0.3065	0.2874	0.2493	0.2382
1000	Recall	0.4141	0.4707	0.4451	0.4451	0.4600	0.4600
	MRR	0.0864	0.2162	0.3069	0.2881	0.2499	0.2348

Table 1: Results of retrieval on collection with ASR-based segmentation

		`min`, queries: 15 relevant: 49		`max`, queries: 24 relevant: 39		`random`, queries: 25 relevant: 42	
cutoff		`aut-man`	`man-man`	`aut-man`	`man-man`	`aut-man`	`man-man`
100	Recall	0.2778	0.3222	0.4583	0.5000	0.4400	0.4400
	MRR	0.1510	0.2840	0.2753	0.2753	0.2997	0.3094
1000	Recall	0.6206	0.6500	0.6042	0.6250	0.5200	0.5200
	MRR	0.1527	0.2847	0.2757	0.2758	0.2999	0.3097

Table 2: Results of retrieval on collection with segmentation based on manual transcripts

		`min`, queries: 15 relevant: 41	`max`, queries: 24 relevant: 55	`random`, queries: 25 relevant: 45
cutoff		`aut-time`	`aut-time`	`aut-time`
100	Recall	0.2815	0.3542	0.3533
	MRR	0.0817	0.1961	0.2796
1000	Recall	0.5781	0.4146	0.4233
	MRR	0.0840	0.1963	0.2865

Table 3: Results of retrieval on collection of ASR transcripts with time segmentation

		`min`, queries: 15 relevant: 41	`max`, queries: 24 relevant: 54	`random`, queries: 25 relevant: 42
cutoff		`man-time`	`man-time`	`man-time`
100	Recall	0.4259	0.3625	0.3733
	MRR	0.1469	0.2128	0.3005
1000	Recall	0.6078	0.4313	0.4467
	MRR	0.1493	0.2130	0.3009

Table 4: Results of retrieval on collection of manual transcripts with time segmentation

[4] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC spoken document retrieval track: A success story. In *Proceedings of RIAO 2000*, pages 1–20, 2000.

[5] J. S. Garofolo, E. M. Voorhees, C. G. P. Auzanne, and V. M. Stanford. Spoken document retrieval: 1998 evaluation and investigation of new metrics. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 1–7, 1999.

[6] P.-Y. Hsueh and J. D. Moore. Automatic topic segmentation and labeling in multiparty dialogue. In *Proceedings of the first IEEE/ACM workshop on Spoken Language Technology (SLT)*, 2006.

[7] M. Larson, M. Tsagkias, J. He, and M. De Rijke. Investigating the global semantic impact of speech recognition error on spoken content collections. In *Proceedings of ECIR 2009*, pages 755–760, 2009.

[8] P. Pecina, P. Hoffmannova, G. J. F. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF 2007 cross-language speech retrieval track. In *Proceedings of the CLEF 2007 Workshop*, pages 674–686, 2007.

[9] A. Popescu-Belis, P. Poller, J. Kilgour, E. Boertjes, J. Carletta, S. Castronovo, M. Fapso, A. Nanchen, T. Wilson, J. de Wit, and M. Yazdani. A multimedia retrieval system using speech input. In *Proceedings of ICMI-MLMI 2009 (11th International Conference on Multimodal Interfaces and 6th Workshop on Machine Learning for Multimodal Interaction)*, 2009.

[10] S. Renals, T. Hain, and H. Bourlard. Recognition and interpretation of meetings: The AMI and AMIDA projects. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*, 2007.

[11] M. Sanderson and X. M. Shou. Search of spoken documents retrieves well recognized transcripts. In *Proceedings of ECIR 2007*, pages 505–516, 2007.

[12] X. M. Shou, M. Sanderson, and N. Tuffs. The relationship of word error rate to document ranking. In *Proceedings of the AAAI Spring Symposium on Intelligent Multimedia Knowledge Manangement, Technical Report SS-03-08*, pages 28–33, 2003.