

# ISSUES IN SPEECH-BASED RETRIEVAL OF VIDEO

*H. J. Nock, G. Iyengar, C. Neti*

IBM T. J. Watson Research Center, 1101 Kitchawan Road, Rte 134/PO Box 218,  
Yorktown Heights, NY 10598. USA.

## ABSTRACT

*This paper discusses issues arising when applying the IBM Audio-Indexing System to retrieval of video. Issues discussed include the relationship between speech transcription accuracy and retrieval performance, query processing schemes and the critical problem of mapping between cues in speech and the relevant video shots. The temporal relationship between the occurrence of cues in speech transcripts and relevant shots is quantified and then simple schemes for performing this mapping are described and evaluated. Experiments demonstrate the promise of more sophisticated schemes involving up-front video ranking and one possible implementation is discussed. Techniques are evaluated using the TREC-2002 Video Track queries and corpus, comprising a total of 68.45 hours of video.*

## 1. INTRODUCTION

Massive quantities of multimedia information are available in libraries and via the internet. Efficient techniques for managing and searching such information, particularly digital video, are the area of very active research at present. Reflecting the interest in this general area, a Video Retrieval Track was started at TREC in 2001 with a high-level goal of “the investigation of content-based retrieval from digital video”. This year’s TREC-2002 video track involved three main tasks, of which participants must complete at least one: shot boundary determination, feature extraction (ie. semantic feature or concept detection, including faces, speech, monologues) and search. IBM participated in all tasks [AAG<sup>+</sup>02], exploring several diverse methods for video analysis, indexing and retrieval including statistical modelling for feature extraction, pure content-based retrieval, statistical model-based retrieval, speech-based retrieval and combinations of all of these.

This paper will focus specifically on the TREC search task and approaches to the search task which are based on indexing through speech audio. Speech-based video search - particularly as formulated within the TREC Video Track - has some similarities with, but also some differences to, audio-indexing as investigated at TREC in the past. The paper discusses analysis and insights gained when using our

standard audio-indexing framework for TREC-2002 Video Retrieval. The resulting speech-based retrieval system is one important baseline for research into the video retrieval problem, since (as was reflected in the TREC-2002 results) speech-based indexing to some extent reflects the best that is currently achievable<sup>1</sup>. The system now provides a baseline for our current work, which is investigating integration of speech, non-speech audio, videotext and video image cues in order to improve retrieval beyond that achievable using speech alone.

The structure of the paper is as follows. Section 2 describes the TREC-2002 Video Track corpus and search task, motivating later sections. Section 3 outlines our experimental setup, including the IBM audio-indexing framework, data set usage and evaluation metric. Sections 4 and 5 quantitatively discuss the relationship between transcript accuracy and retrieval performance and the utility of techniques for query processing. Section 6 investigates the relationship between the occurrence of relevant video shots and associated cues in the speech soundtrack; it then considers schemes for mapping from speech cues to the relevant video shots. The paper ends with brief conclusions and possible future work.

## 2. TREC-2002 CORPUS AND SEARCH TASK

The TREC-2002 Video Track corpus (“VT02”) comprises 68.45 hours of MPEG-1/VCD from the Internet Archive and the Open Video Project. The data is old, in the main from the period 1950 through 1970. The data set has some advantages for TREC, not least of which is that the content is freely available, but the reader should bear in mind that the unusual content does have implications for the techniques used and the conclusions drawn. The corpus includes some silent videos for which speech-based retrieval is not possible. Audio and colour quality is very variable, so audio and video detectors for modern data must be retrained or

---

<sup>1</sup>We do not intend to suggest that speech-based retrieval reflects the state-of-the-art for *all* types of query. Some concepts such as indoor/outdoor, faces, music, etc can be performed well using purely video or non-speech audio information, as illustrated by the TREC-2002 feature extraction task, though our recent work has found that speech cues (where available) can also be helpful for detecting these traditionally “video-based” concepts [NIL<sup>+</sup>03].

returned. The production styles are also very different than would be seen in today’s fast moving TV shows or news programmes. The VT02 corpus is partitioned into three parts: search test collection (“ST”, 40.12 hours, 176 videos including 19 silent videos), feature development collection (“FD”, 23.26 hours, 96 videos including 12 silent videos) and feature test collection (“FT”, 5.07 hours, 23 videos including 2 silent videos).

The TREC-2002 search task is stated as, “Given a multimedia statement of information need and the common shot boundary reference for the search test collection, return a ranked list of 100 shots from the standard set which best satisfy the need”. Twenty-five such statements of information need, comprising text and one or more image or video (with soundtrack) examples were provided by NIST just prior to the evaluation. Examples include:

- *Find me overhead views of cities - downtown and suburbs. The viewpoint should be higher than the highest building visible;*
- *Find me shots with a map (sketch or graphic) of the continental US (Figure 1);*
- *Find shots of Price Tower, designed by Frank Lloyd Wright and built in Bartlesville, Oklahoma (Figure 2);*
- *Find shots of one or more women standing in long dresses. Dress should be one piece and extend below the knees. The entire dress from top to end of dress below knees should be visible at some point.*



Figure 1: Example Image for “continental US map” Query

NIST did not require query processing to be done fully automatically this year, so at most sites query formulation was performed by “expert users” who looked at the NIST queries, at VT02 data partitions other than that reserved for the search test task and at background information on the internet before formulating the final query to be submitted to their system.



Figure 2: Example Image for “Price Tower” Query

### 3. EXPERIMENTAL SETUP

**Document Indexing and Retrieval System:** The IBM Audio-Indexing framework is described in detail in [DFR99]. It combines a large vocabulary automatic speech recognizer and a text-based information retrieval system. The document indexing phase is as follows. Given speech transcripts with time-stamped words for all of the non-silent videos in the corpus, the transcripts are tokenized into sentence-like units, tagged with part-of-speech and then each word is decomposed into its component morph based on the tag. The morph-ed transcripts are divided into “documents” using a 100 word sliding window with window shifts of 50 words; in our current implementation, we “break” windows across stretches of music which are longer than 3 seconds, taking this to be indicative of scene or subject changes, though later empirical results show this to have little effect on overall retrieval performance. Word-level indexes are constructed for word unigrams and word bigrams. During the document matching phase (ie. search time), a textual statement of information need is processed using tokenization, tagging and morphological decomposition as used in the original transcript processing. Database documents are then ranked against the query using the IBM version [DFR99] of the OKAPI formula [RWJ<sup>+</sup>95]. A technique similar to Local Context Analysis (“LCA”, first presented in [XC96] and discussed in [FSMR99]) is applied to the top results from the first retrieval pass and statistics about word cooccurrence are used to expand the query with further weighted terms for use in a second retrieval pass.

The baseline system for our TREC-2002 work was constructed using only the first retrieval pass. No attempt was made to index the silent videos. The issues of transcript accuracy, query formulation and expansion, mapping between speech cues and the NIST-required ranking of relevant shots (as opposed to intervals of speech indicated by “documents”) will be addressed in detail below<sup>2</sup>.

<sup>2</sup>We note in passing that our final TREC-2002 submission was based upon an additively weighted fusion of shot-level rankings from three audio-indexing systems and this combination outperforms the single audio-

**Data Sets and Usage:** The scenario for all experiments in this paper is the “no interaction”, “held-out search data” scenario:

- “no interaction”: between the point at which a formulated query is submitted and results are returned, there is no manual interaction with the system (the contrast being an interactive run involving eg. relevance feedback). However, as noted earlier, the conversion between NIST statement of information need and the final query as submitted to the system is done by an expert user;
- “held-out search data”: no information from the search test collection that is not automatically extracted can be used to tune the system parameters prior to retrieval (the contrast being a true “static library” model, in which the system builder can exploit any properties of the library data in developing their models and only the queries are an unknown).

The Feature Development set (“FD”) was used for system development. Since no VT02 ground truth was made available pre-evaluation, we constructed a (potentially incomplete) shot-level relevance assessment by pooling the results from three (quite complementary) audio-indexing systems that were built to index FD. All FD results are scored against this *partial* ground truth; ST results presented are scored against the more complete (but still pooled) ST relevance assessments provided post-evaluation by NIST.

**Evaluation Metric** NIST assessed video retrieval performance using non-interpolated Mean Average Precision (MAP) at the shot level over top 100 returned shots. A two-stage procedure is used to calculate MAP. First, a per-query Average Precision (AP) is calculated by computing precision after every relevant retrieved shot and then averaging these precisions over total number of relevant shots in the collection. Then, the per-query averages are combined (averaged) across all queries to yield Mean AP (MAP).

#### 4. EFFECT OF TRANSCRIPT ACCURACY

A series of increasingly accurate speech transcriptions were produced for the VT02 corpus using different IBM automatic speech recognition (ASR) systems. The first set of transcriptions were produced using an IBM real-time transcription system tuned for Broadcast News. Later transcriptions were produced using an off-line, multiple pass transcription system similar to [CEG<sup>+</sup>02]. A preliminary stage

indexing system discussed in this paper. However, the issues discussed in the paper were pertinent to all three component systems and improvements to the component systems in these areas should translate into improved “fusion system” performance. For reference, the individual component systems perform in the MAP range 0.10–0.11 and the “fusion system” at 0.15; details of system fusion scheme are in [AAG<sup>+</sup>02].

removed purely silent videos based on the output of a Gaussian Mixture Model (GMM)-based detector. The next stage identified speech regions of the soundtrack as follows. Per-video segmentation of the non-silent videos into segments which are well modelled by a single Gaussian is performed using the Bayesian Information Criterion (“BIC”; see [CG98]). The result is a division of each video into short segments suitable for decoding. GMM-based silence, music and speech classification is run over each of the resulting segments and pure music or silence segments are removed. Then, a decoding of all segments using an IBM 10×Real-time Broadcast News transcription system is used to provide more accurate silence locations. These are used to adjust the boundaries of the speech segments obtained from the GMM classification. The first full transcription pass for the retained speech regions uses a set of Broadcast News acoustic models adapted to VT02 data using a small set of word-level transcribed videos from FD and supervised Maximum Likelihood Linear Regression (MLLR, [Leg95, GW96]) with global mean and precision transforms. These models are combined with an interpolated trigram language model to produce an initial transcript. This transcript is used to refine the acoustic models for a second transcription pass as follows. BIC-based speech-only segments are clustered into “speaker- and environment-similar” clusters [CG98]; the second decoding pass uses unsupervised MLLR adaptation of the (already supervised adapted) acoustic models to each of these clusters, again using global mean and precision transforms. The final result is a transcript with time-stamped words. The word error rate (WER) of the final transcripts used throughout the paper is estimated at 38.7% on a held out set of six videos from ST which were manually transcribed. This compares favorably to 42.7% for the best of the publically-released transcriptions on the same set and represents a 36% improvement over transcriptions produced by the ASR engine used in IBM’s TREC-2001 submission.

The Spoken Document Retrieval community reports that below a certain ASR WER level (30-40%, depending upon task and site), gains in ASR-based retrieval arising from reductions in WER begin to tail off. Table 1 shows the results of indexing our transcripts of different accuracies and shows a trend in which all ASR improvements translate into large improvements in MAP; no tailing off of MAP improvements has yet appeared on the VT02 corpus. Ground truth manual transcriptions are not available for the search test set, so MAP for retrieval on perfect transcripts is not currently known. Note that we find a consistent relationship between transcript % Correct and MAP rather than % Accuracy and MAP. This is because some of the intermediate recognition systems have good transcription performance within speech regions but do not remove music particularly well; this leads to a low overall accuracy reflecting the large number of insertions, but is rather misleading because the

insertions are mostly of words which are not relevant to indexing eg. multiple occurrences of the hesitation [MMMM] in retained music regions.

Transcript % Correct	FeatureDevelopment Set (FD) MAP
52.8	0.09
59.7	0.13
67.9	0.17
72.8	0.21

Table 1: Relationship between speech recognition performance and retrieval Mean Average Precision

## 5. QUERY PROCESSING

The queries used by IBM for the TREC-2002 submission were constructed by manually expanding the NIST textual statements of information need with pertinent words from the audio soundtrack of the NIST-supplied example shots and then using retrieval on FD to suggest further possible query terms. This *ad-hoc* query formulation scheme will be referred to as “Manually Expand on FD”. Post-evaluation experiments investigated whether this had improved performance beyond that achievable using simple prefix-stripped versions of the NIST textual statements of information need<sup>3</sup>. This scheme is termed “Automatic Prefix-Strip”. Table 2 shows manual query tweaking did improve performance on the retrieval training data FD, as expected, but the ultimate gain on held-out ST data was minimal. This bodes well for full automation of the query processing step in future years and the result is not entirely surprising. Firstly, analysis later in the paper will show helpful speech cues often occur outside relevant shots; thus, adding words taken from the sample shots that NIST provided with the queries is not guaranteed to add useful query terms. Secondly, whilst FD shares some similarity with ST in terms of broad video topics, the specifics of videos are very different: thus, the query terms added based on FD are also often not useful.

The IBM Audio-Indexing System has had some success using Local Context Analysis (“LCA”, [XC96]) in a two-phase retrieval process: after retrieval using the original query, LCA analyses the top returned documents to automatically expand the query with (weighted) pertinent terms

<sup>3</sup>Standard prefixes stripped from the queries include ‘Find me shots containing’, ‘Find me pictures of’ etc.

Query Processing	FeatureDevelopment	SearchTest
Automatic Prefix-Strip	0.19	0.10
Manually Expand on FD	0.23	0.11

Table 2: Results: Query Expansion Schemes

for use in a second retrieval pass. We found LCA did not perform well overall on the VT02 data, hurting more queries than it helped, and we believe there are two explanations. Firstly, many queries have only a small number of relevant shots (six queries have less than ten relevant items) and the documents spanning those shots are often not ranked highly in the first pass. Secondly, even when documents spanning the relevant shots are highly ranked, the words contained are not particularly consistent. For example, the query “images of George Washington” finds documents involving the president but containing few other common cues. Both of these problems mean that LCA does poorly at selecting relevant terms for query expansion. Since this result suggests that knowledge-based query expansion may be more appropriate for this task, later experiments investigated automatic query expansion using WordNet [Fel98] but again performance degraded slightly.

## 6. RELATING SPEECH CUES TO SHOTS

NIST defines the basic unit of indexing and retrieval as a video shot, so a successful TREC-2002 video retrieval system must be able to map between cues to the occurrence of a relevant object in the speech soundtrack and the actual shot which contains the relevant object. We first present quantitative analysis of this relationship in Section 6.1. The mapping between speech cues and relevant items is a key problem in successfully using speech to retrieve relevant shots in video. Whilst it might be expected that successful solutions will come from integrating speech-based retrieval with content- and model-based retrieval, success was achieved at this year’s TREC-2002 using the simple speech-only solutions described in Section 6.2.

### 6.1. Quantifying Relationship

A widely mentioned, though rarely quantified, belief in the speech-based video retrieval community is that relevant items tend to occur a little later than any associated cues in the speech soundtrack; this belief is based upon a rule-of-thumb of video production which states that speech (and music) should in general precede the associated visual content.

For the following analysis a speech cue is defined as the occurrence of a morph-ed, query term match in the speech soundtrack. The FD shot-level ground truth is then used in combination with the best ASR transcripts to investigate the temporal relationship between occurrence of speech cues and the associated relevant shots<sup>4</sup>.

<sup>4</sup>Perhaps the ideal version of this experiment would involve a human listening to every video containing a relevant shot(s) and manually marking every term which they judge as indicative of the concept of interest. This is very labour intensive and we believe that the conclusions drawn would be similar to those from our experiment. Firstly, whilst our analysis is done using transcripts which are not perfect, it is rare that a transcript

Time interval relative to shot center (seconds)	Number of ground truth items with cue in interval
(-60,-20)	86
(-20,-10)	61
(-10,-3)	68
(-3,-1)	24
(-1,0)	32
(0,1)	27
(1,3)	48
(3,10)	79
(10,20)	62
(20,60)	105

Table 3: Temporal Occurrence of Cues for Relevant Shots

There are 202 ground truth items in our FD ground truth, of which 21 have cues occurring only after the center of the relevant shot but within the same video and 14 have cues occurring only before the center of the relevant shot but within the same video. Thus the majority of ground truth items have cues occurring both before and after the center of that shot. The average shot length in the NIST-supplied shot boundary reference is 7 seconds. Table 3 gives more indication of the temporal dispersion of cues and shows that cues can occur much before and after the center of a relevant shot.

These results show that the relationship between speech cues and relevant items is not as straightforward as basic video production theory and other authors might suggest. At least for VT02, helpful word cues occur both before and after relevant shots and often more than our average document length of 23 seconds from the center of the relevant shot. Whilst this may be attributable in part to the era of the VT02 corpus, we have (qualitatively) observed similar trends on more modern data.

## 6.2. Simple Document-to-shot Mapping Schemes

The standard IBM audio-indexing system indexes documents, where the average length of a document is 100 words (on average, about 23 seconds) and is constructed by sliding a 100-word window across the transcript with a window shift of 50 words. Thus, the system provides a ranking of intervals of speech which are longer (on average) than the average shot length (7 seconds). In contrast, NIST requires a ranking of shots. Preliminary experiments investigated shorter document lengths more closely related to the length

error inserts a relevant speech cue although some relevant cues may be misrecognised. Secondly, whilst the set of speech cues that we consider is potentially incomplete, any occurrence of a word from this set is almost certainly a genuine speech cue. This means the quantitative results are in some sense approximate lower bounds on the true distribution of speech cues: thus, the conclusion that relevant cues occur both before and after the shot would still hold although the proportions of cues in each time span may change a little.

Mapping Scheme	FeatureDevelopment (FD)	SearchTest (ST)
	MAP	MAP
LOS	0.16	0.07
AOD	0.19	0.11
ACTM	0.19	0.09
AFLTM	0.26	0.09

Table 4: Results for Different Document-to-Shot Mappings

and boundaries of shots but this gave poor results, most likely because of the potentially long separation between the occurrence of speech cues and relevant shots. Use of 100-word documents centered on shots also degraded performance. Therefore we maintained use of 100 word documents defined in a sliding window fashion and focused instead upon schemes for a post-processing mapping between the interval of time indicated by retrieved “documents” and the relevant shots. Since only very limited ground truth data was available pre-evaluation this year (see discussion of **Data Sets and Usage**), the following schemes were investigated as simple starting points:

- longest shot overlapping document (“LOS”);
- all shots overlapping document (“AOD”);
- all shots in document containing term match (“ACTM”);
- all shots between times of first and last term match in document (“AFLTM”).

The results are shown in Table 4, which illustrates the importance of the document-to-shot mapping in determining final performance. Schemes AOD and AFLTM give the best performance. This may be a consequence of the TREC data; subjectively, the film style is such that consecutive shots often present alternative views of the same event or scene. However, now that more ground truth is available, future work should investigate use of a statistical model in performing the document-to-shot mapping.

## 6.3. Up-front Video Ranking

Subsection 6.1 shows that speech cues can occur at time scales longer than a single “100 word document”. However, preliminary experiments using documents longer than 100 words with similar post-processing document-to-shot mapping schemes did not significantly change performance. One possible alternative is to use multimodal cues to provide an up-front ranking or hard subsetting of videos into “relevant” and “irrelevant” sets; this list can then be used to filter or otherwise modify the shot ranking returned by the standard speech-based retrieval scheme. We present a “cheating”-type analysis that shows such an approach could lead to significant improvement in performance.

Document-To-Shot Mapping	Baseline Search Test (ST) MAP	Manually Subset Videos ("Cheat") Search Test (ST) MAP
LOS	0.07	0.10
AOD	0.11	0.15
ACTM	0.09	0.13

Table 5: Potential Gains From Video Subsetting

### 6.3.1. Potential Gains from Video Subsetting

We investigate the potential utility of an up-front video-ranking scheme by assuming that we have an oracle (ie. a perfect black box) that can magically separate the full search test set of videos into “relevant” and “irrelevant” videos. We then use the “relevant” list to subset the set of shots returned by the speech-based retrieval system described earlier. Whilst it is obvious that such an oracle will improve results, we are interested in the size of the potential gains because we hypothesise that the problem of video ranking (using eg. speech cues or video cues) may be simpler to solve in some fashion than the overall problem of shot ranking. Table 5 shows the oracle results, which suggest gains of up to 44% are possible.

### 6.3.2. Automatic Video Subsetting or Video Ranking

We outline one implementation of a scheme to automatically rank videos and then use this ranking to filter the results from the speech-based retrieval system. The scheme runs queries against both the document-level index (as discussed above), which ranks 100 word documents and assigns scores  $S$  to shots using the AOD document-to-shot mapping scheme, and then against a video-level index, which assigns OKAPI-based scores  $V$  to each video. We combine these scores to get an overall shot score using a product rule  $V^\alpha * S^\beta$ . Weights  $\alpha, \beta$  are optimised on FD prior to running the queries on held-out ST data. (The product form is chosen because, if we had a perfect video subsetting mechanism assigning video level scores of zero or one, this scheme with any non-zero  $\alpha$  will give the same result as the cheating experiment.) This scheme gives a 5% relative improvement in MAP. We attribute the limited gains in part to the weights estimated on FD being poor estimates for ST (in principle gains of 10% or more could be achieved with the “optimal” weights), but mainly due to the very similar use of speech-based information in the ranking procedures for both documents and videos in this initial experiment. Future work will replace the up-front speech-based video ranking with a multimodal approach which additionally incorporates image and non-speech audio information.

## 7. CONCLUSIONS AND FURTHER WORK

This paper has discussed issues arising when applying the IBM audio-indexing framework to the TREC-2002 Video Retrieval Track search task. These include the effects of improving transcript accuracy, query expansion and the challenging problem of relating speech cues to relevant shots. The resulting system forms an important baseline for our current video retrieval research, which is examining schemes for integrating multimodal (video, non-speech audio and speech) cues to improve retrieval performance over that which is achievable using speech alone.

## 8. ACKNOWLEDGEMENTS

We thank Martin Franz for supplying the basic spoken document retrieval software and the IBM TREC-2002 team for assistance and discussion (particularly Arnon Amir, Alejandro Jaimes, Haim Permuter, Larry Sansone, John Smith, Belle Tseng and Savitha Srinivasan).

## 9. REFERENCES

- [AAG<sup>+</sup>02] Bill Adams, Arnon Amir, Sugata Ghosal, Giridharan Iyengar, Alejandro Jaimes, Christian Lang, Ching-yung Lin, Apostol Natsev, Milind Naphade, Chalapathy Neti, Harriet J Nock, Haim H Permuter, Raghav Singh, John R Smith, Savitha Srinivasan, Belle L Tseng, Ashwin T Varadaraju, and Dongqing Zhang. Ibm research trec-2002 video retrieval system. In *Proc. TREC Workshop (Video Retrieval Track)*, 2002.
- [CEG<sup>+</sup>02] SS Chen, EM Eide, MJF Gales, RA Gopinath, D Kanevsky, and P Olsen. Automatic Transcription of Broadcast News. *Speech Communication*, May 2002.
- [CG98] Scott Shaobing Chen and PS Gopalakrishnan. Speaker, Environment and Channel Change Detection And Clustering via the Bayesian Information Criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, 1998.
- [DFR99] S Dharanipragada, M Franz, and S Roukos. Audio-Indexing for Broadcast News. In *Proceedings of the Sixth Text REtrieval Conference (TREC-7)*, page 157, 1999.
- [Fel98] C Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.
- [FSMR99] M Franz, J Scott McCarley, and S Roukos. Ad hoc and Multilingual Information Retrieval at IBM. In *Proceedings of the Sixth Text REtrieval Conference (TREC-7)*, page 157, 1999.
- [GW96] MJF Gales and PC Woodland. Mean and Variance Adaptation within the MLLR Framework. *Computer Speech and Language*, 1996.
- [Leg95] CJ Leggetter. *Improved Acoustic Modelling for HMMs using Linear Transformations*. PhD thesis, University of Cambridge, 1995.
- [NIL<sup>+</sup>03] HJ Nock, G Iyengar, C-Y Lin, MR Naphade, C Neti, JR Smith, and BL Tseng. User-Trainable Video Annotation Using Multimodal Cues. In *SIGIR (Submitted)*, 2003.
- [RWJ<sup>+</sup>95] SE Robertson, S Walker, S Jones, MM Hancock-Beaulieu, and M Gattford. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. NIST Special Publication 500-226, 1995.
- [XC96] J Xu and WB Croft. Query Expansion Using Local and Global Document Analysis. In *Proc 19th ACM SIGIR*, pages 4–11, Switzerland, 1996.