End to end evaluation of the ISL View4You broadcast news transcription system

T. Kemp*, M. Weber, A. Waibel
ISL Interactive systems laboratories, University of Karlsruhe,
Am Fasanengarten 5, 76129 Karlsruhe, Germany

kemp@sony.de, weber@ira.uka.de, ahw@cs.cmu.edu

Abstract

In this paper, we introduce the Interactive Systems Laboratories video indexing and retrieval system 'View4You'. The main components of the system, namely the segmenter, the speech recognizer and the information retrieval engine, are described in detail.
In the View4You system, public television newscasts are recorded on a daily basis. The newscasts are automatically segmented and an index is created for each of the segments by means of automatic speech recognition. The user can query the system in natural language. The system returns a list of segments which is sorted by relevance with respect to the user query. By selecting a segment, the user can watch the corresponding part of the news show on his or her computer screen.
Several end to end evaluations on real world data, using questions from naive users, are described. By substituting each of the components of the system with a perfect (manually simulated) one, the effect of the components' imperfection on the end to end result can be determined. We show, that the information retrieval component has the largest impact on the system performance, followed by the segmentation. The quality of the speech recognizer, as long as its error rate is below approximately 25%, is shown to have only a relatively small importance.

1 Introduction

More and more information is produced and stored in the form of video data which is opaque to textual queries. The manual indexing of video material, however, is tedious, time-consuming and expensive. To facilitate searchable databases of video data, an automatic way of creating an index is highly desirable. The goal of the View4You system is to provide a framework for research aimed in this direction, by creating a limited, but operational prototype system that allows the optimization and evaluation of both the relevant components and the system as a whole.

The video material for our prototype system is taken from public TV newscasts. The main advantage of this approach is the easy access to large amounts of videos. Additionally, news shows offer a high variety of acoustical conditions and topics, which makes their indexing and retrieval a scientifically challenging task.

Figure 1 shows a block diagram of the View4You prototype system. Public newscasts are recorded daily using standard consumer electronics equipment. The newscasts are then automatically segmented into chunks of approximately 10 to 90 seconds length. For each of the segments, a speech recognizer generates a hypothesis of the segment's audio. The segment boundaries, the hypothesis of the speech recognizer and the video data for the segment are stored in the multimedia database. From the internet, newspaper articles are collected and added to the database.

*Now with SONY International (Europe) GmbH, 70736 Fellbach

User queries can be issued in natural language, and can be entered either by keyboard or through a speech interface. They are processed by the query server, which performs a search in the multimedia database and returns the found segments sorted by relevance (similarity) with regard to the query. In our frontend, the retrieved video segments are presented as thumbnail pictures of the beginning of the video segment, and found newspaper texts are represented by a graphical symbol. By clicking on the picture or the symbol, the video is played on the screen. Newspaper texts are displayed in a text window.

Much research has been focused on each of the components of the View4You system, mostly on the automatic speech recognition of broadcast news data (Kubala et al, 1997)(Gauvain et al, 1997)(Woodland et al, 1997)(Wegmann et al, 1999)(Chen et al, 1999)(Beyerlein et al, 1999). However, we are aware of only one other complete video indexing system (by CMU's Informedia group (Hauptmann and Wactlar, 1997)), that allows the actual retrieval of automatically transcribed broadcast news videos. Also, several systems have been presented that handle other types of spoken documents (e.g. (Young et al, 1997)(Choi et al, 1998)).

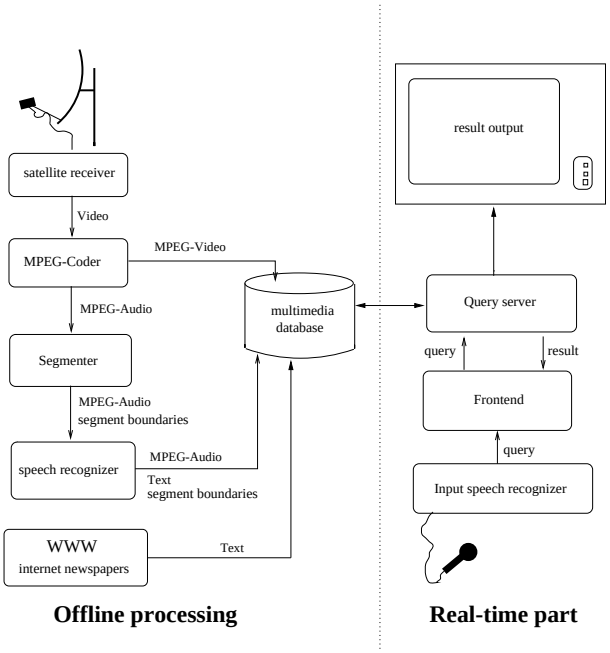

Figure 1: System overview

In the following paragraphs, we describe the database, the segmenter, the speech recognizer, and the retrieval engine of the View4You system.

2 The user interface

Figure 2 shows the user interface of the prototype system. On the right side of the screen, the segments that have been found to the query 'Informationen über Albanien und Italien bitte' ('Please give informations about Albania and Italy') are shown, sorted by decreasing relevance. The search result with the highest relevance score has been selected by the user and is currently being played (on the left of the screen).

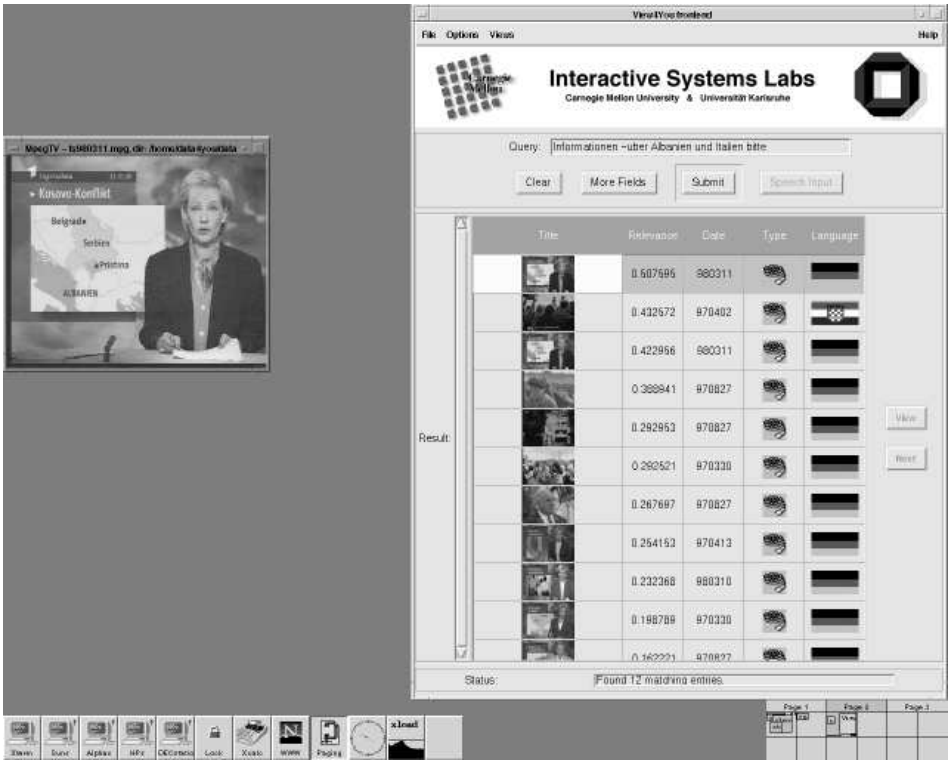

Figure 2: User interface

3 The broadcast news database

3.1 The news program

Our German data is collected from a German TV news show called 'tagesschau' (which can be translated as 'view of the day'). The 'tagesschau' is by far the most prominent newscast on German TV: between 8 p.m. and 8:15 p.m, which is its scheduled time since 1961, every third German TV set is regularly switched to the 'tagesschau'. Events of all types are reported, including sports and business news. The look is very traditional: the anchor speaker reads the news from a sheet of paper in front of him or her, and there are neither jingles separating the different reports nor background music of any type. The anchor speaker segments are intermixed with reports from different events. The audio for the reports originates either from the location of the shot itself, or is supplemented by a studio off speaker. In the latter case, the original audio is still played, but with decreased volume.

3.2 Data collection

The TV news shows are received from the television satellite ASTRA-1b and are stored as MPEG-compressed files, using the MPEG-1 compression algorithms. The total data rate of the MPEG stream is set to 1.2 MBit/s. The audio data is compressed using MPEG audio layer 2 compression at a data rate of 192 kbit/s and a sampling rate of 44.1 kHz. In a recognition experiment, we compared this with compressing the audio at a sampling rate of 32 kHz while maintaining a data rate of 192 kbit/s, which should result in better quality in the relevant frequency range from 0 to 16 kHz. However, there was no significant difference in the word error rate, so that we decided to continue working with the 44.1 kHz sampling rate, thereby ensuring compatibility to the data format used in CMU's Informedia project (Wactlar et al, 1999).
The recorded audio signal is sampled down to a 16 kHz/16 bit PCM format which is used throughout all experiments. The video part of the signal is stored 'as is' to allow video output

in the prototype system. It is not used for the creation of the database index.

There are large differences in style between the US news shows used by the ARPA broadcast news evaluations (NIST, 1998) and the 'tagesschau' newscast. We tried to segment the 'tagesschau' using the same so-called F-conditions used by ARPA, but found that three out of 7 different F-conditions (F1, F5 and FX) are virtually nonexistent in the 'tagesschau'. Most of the data would be categorized into one of two other F-conditions. Therefore, we decided to use only two classes, anchor-speaker and non-anchor, where 'anchor-speaker' can be identified with ARPA's F0 condition, and 'non-anchor' means everything else (and would mostly be tagged F4 or F2).

In the 'tagesschau' recordings, roughly 60% of the data is speech from male speakers. About 40% of the data is anchor speaker speech, i.e. high-bandwidth high-quality recordings without additive background noise.

4 The automatic Segmenter

There are at least three reasons why segmentation is required in a broadcast news indexing system like View4You. First, speech recognition technology usually requires segments of relatively short length as input. *Any* segmentation for which the maximum segment length does not exceed the capability of the speech recognizer, and that avoids cutting within words, satisfies this requirement.

Second, as speakers tend to repeat within a given news show, speaker adaptation schemes can be used to improve ASR performance. This is usually done by an initial segmenter run, followed by a clustering step that tries to group segments from one speaker together. The speaker adaptation for a given segment is carried out using all segments of the corresponding cluster. A segmentation that is used for speaker adaptation needs to have a high *segment purity*, e.g. one segment should contain only one single speaker and acoustic condition. A speaker turn can be segmented into two or more segments without harm, since over-segmentation is unproblematic due to the clustering step.

The third reason for segmentation in a BN retrieval system is user friendliness. For a given query topic, it is rarely an appropriate answer to return one complete, unsegmented news show and let the user decide which part of it is of interest. Ideally, the information system presents only the relevant parts. For this, however, the starting point and length of each story must be determined by the segmenter. Since it is disturbing to have either too short segments, where part of the information remains hidden, or too long segments where irrelevant information is displayed, the segment boundary must match the story boundary as exactly as possible.

Recently, several groups have investigated the problem of segmenting broadcast news in the context of ARPA's hub-4 broadcast news transcription and understanding evaluations ((Woodland et al, 1998) (Polymenakos et al, 1998)(Chen et al, 1998)(Wegmann et al, 1998)(Gauvain et al, 1998)(Sankar et al, 1998)(Siegler et al, 1997)(Wactlar et al, 1996)). The goal of the segmentation in the ARPA-supported experiments was mainly to provide a basis for speech recognition and speaker adaptation, not to find the true story boundaries as required by a retrieval system. For the View4You segmentation, however, it is necessary to find the true story boundaries, and therefore the evaluation in this paper measures how well different segmenting approaches can find the story boundaries. The main difference to the more common segment purity or word error rate based segmenter evaluations is, that oversegmentation matter in our case, but does not matter (much) if segment purity or WER is measured.

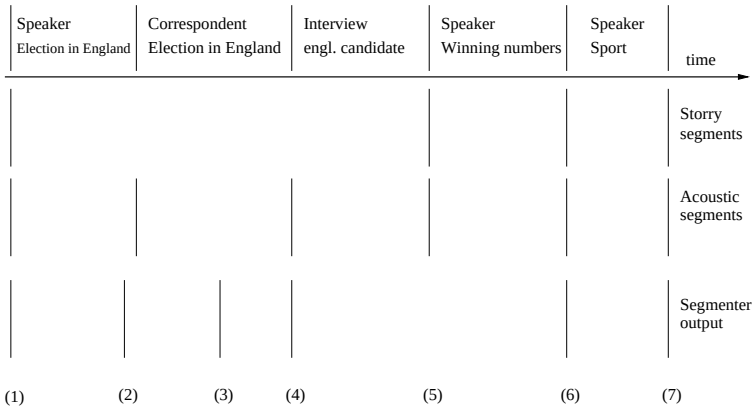

Figure 3: Description of the segmenting problem

The different approaches which have been used in the ARPA evaluations can be categorized into three classes (Chen et al, 1998):

- Model-based segmentation. Different models, e.g. Gaussian mixture models, are constructed for a fixed set of acoustic classes, such as anchor speaker, music etc, using the training corpus. The incoming audio stream can be classified by maximum likelihood selection. Segment boundaries are assumed where a change in the acoustic class occurs.
- Metric-based segmentation. The audio stream is segmented at maxima of the distances between neighbouring windows placed in evenly spaced time intervals.
- Energy-based segmentation. Silence in the input audio stream is detected either by a decoder or directly by measuring and thresholding the audio energy. The segments are then generated by cutting the input at silence locations.

All segmenting approaches are reported to work reasonably well for speech recognition and speaker adaptation. However, no evaluation has yet been carried out to examine how well the different algorithms work when applied to the problem of finding the true story boundaries.

4.1 Evaluation metrics

The result of a segmentation can contain two possible types of error. Type-I-errors occur if a true segment boundary has not been spotted by the segmenter (deletion). Type-II-errors occur if a found segment boundary does not correspond to a segment boundary in the reference (false alarm, or segment insertion). The information retrieval community uses two closely related numbers, precision (PRC) and recall (RCL). Precision and recall can be expressed by Type-I-error rate and Type-II-error rate, and vice versa. They are defined as

$$\text{RCL} = \frac{\text{number of correctly found boundaries}}{\text{total number of boundaries}}$$
$$\text{PRC} = \frac{\text{number of correctly found boundaries}}{\text{number of hypothesized boundaries}}$$

Most segmentation algorithms can be made to work at different operating points. Each operating point corresponds to a (PRC,RCL) pair. As the relative cost of a missed boundary versus the cost of a false alarm depends on the application, a segmenter is fully characterized by a plot of Precision over Recall for all possible operating points. Such a plot is referred to as

'receiver operator characteristic'.

Sometimes it is desirable to have one single number for the performance of an algorithm instead of two. In such cases, the F-measure F is frequently used (Rijsbergen, 1979). It can be parameterized to put higher weight to either PRC or RCL. The neutral parametrization, where Precision and Recall are weighted equally, is used throughout this work. F is defined as

$$F = \frac{2 * PRC * RCL}{PRC + RCL} \tag{1}$$

Like RCL and PRC, it is bounded between 0 and 1.

The correct position of a segment boundary is not exactly defined. In most cases, two segments are separated by a short period of silence. Any segment boundary within the silence period should be regarded as correct. Therefore, a tolerance Δt is defined. If a segment boundary is hypothesized within the time interval $t_0 - \Delta t < t < t_0 + \Delta t$ of the reference boundary t_0, it is judged correct. For our experiments, we chose $\Delta t = 1.5$ s.

4.2 The View4You standard segmenter

In our system, we have evaluated all three approaches (model-based, distance based, and energy based). We found, that a model-based approach performed best at medium level of recall. This approach has also the advantage of being computationally very efficient. Therefore, we chose a model-based segmenter for the View4You system.

In model-based segmentation (Woodland et al, 1998)(Sankar et al, 1998), a set of models for different acoustic classes is defined and trained prior to segmentation. The incoming audio stream is classified using the models, usually imposing additional minimum class length constraints. Boundaries between the classes are used as segment boundaries. Model-based segmentation assumes knowledge about the type of the audio that is to be segmented.

In our model-based segmenter, a speech recognizer was used with a four-word dictionary ('Anchor', 'Field', 'Music' and 'Silence'). The corresponding HMM states used diagonal variance gaussian mixture models (GMMs) as emission probabilities. The GMMs were trained on two hours of manually labelled audio. The audio data used for training was disjunct from the four newscasts used for testing. The number of mixture components per class was chosen according to table 1. By duplicating HMM states, a minimum word duration as shown in table 1 was enforced. No state transition probabilities and no language model were used.

In the acoustic preprocessing, 16 mel-spectral parameters were computed every 50 msec, using a 16 msec window. Although this parameterization does not make use of two thirds of the signal, it performed equally well as compared to a frameshift of 10 msec, but requires only one third of the computing time. Mel-cepstral parameters led to performance degradation and were therefore replaced by the mel-spectral parameters. (Woodland et al, 1998) proposed to perform MLLR adaptation on the segments resulting from the first run, and then re-run the segmenter with the adapted models. However, using this method did not improve the performance on our data.

The 'word' boundaries in the hypothesis of the recognizer were used as segment boundaries. Different operating points could be achieved by changing the value of the word insertion penalty during the search: a high word insertion penalty led to fewer words in the hypothesis and hence fewer segment boundaries.

The results for the operating point yielding the highest F-measure are shown in table 2.

class	number of mixtures	minimum length
anchor speaker	128	5 sec
field speech	128	5 sec
music	32	2.5 sec
silence	2	0.2 sec

Table 1: Parameters for the model-based segmenter

Algorithm	RCL	PRC	F-measure
model-based	0.56	0.70	0.62

Table 2: Performance of the model-based segmenter

4.3 Alternative segmenting approach

The model-based segmenter is used as the standard segmenter in our system. In (Kemp et al, 2000), we present a new hybrid segmentation strategy, and compare the results of the new segmenter to those of the three standard approaches. The hybrid segmentation outperforms the other methods, however at the cost of a very high computational load. The performance of the hybrid segmenting algorithm is given in table 3.

Algorithm	RCL	PRC	F-measure
hybrid	0.67	0.94	0.79

Table 3: Performance of the model-based segmenter

Figure 4 summarizes the result of the evaluation of the four segmenting approaches. For more details, see (Kemp et al, 2000).

5 The View4You speech recognizer

In this section, we give a detailed description of the View4You speech recognizer which is used to generate the index for the video database.

5.1 Preprocessing

In the preprocessing stage, 13 cepstral parameters are computed from 30 melscale filter bank coefficients using a 16 ms (256 samples) Hamming window. The frame shift is chosen as 10 ms. A simple energy-based speech detection is performed, and cepstral mean subtraction is applied where the cepstral mean is computed on the speech part of the signal only. The 13-component vector is merged with its delta and delta-delta coefficients into a 39-component intermediate feature vector. This intermediate feature vector is then LDA-transformed (Fukunaga, 1990) to the final 16-component feature vector. In all experiments described, we made use of vocal tract length normalization (Zhan & Westphal, 1997), which warps the power spectrum to a reference vocal tract length before the computation of melscale filter bank coefficients takes place.

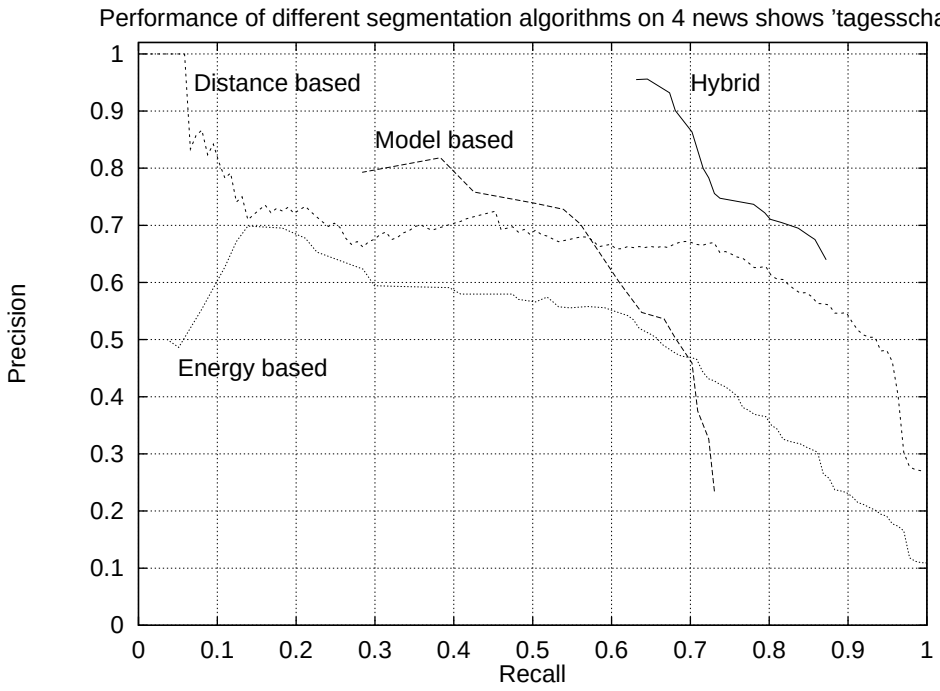

Figure 4: Result of different segmenting algorithms

5.2 Dictionary and vocabulary

The recognition dictionary contains the 60788 most frequent words from the language model training corpus (see table 4). For all but the most frequent 10000 words, dictionary variants were discarded. The resulting recognition dictionary has 61685 entries. Words are represented as sequences of phones. The View4You system uses 44 different base phones (including silence), which are derived from the SAMPA phoneset.
Broadcast news, and especially the parts that are not read by the anchor speaker, is partly spontaneous. To capture some of the effects of spontaneous speech, like e.g. hesitations, five specialized noise models were added to the phone set: two hesitation models, one generic model for breathing noise, one model for other noise originating from the human vocal tract, and one model for all other noises. The noise phones and the silence phone are assumed to be independent of the phonetic context and were therefore not subject to the decision-tree clustering process described below. Each phone except the silence phone is further divided into three sub-phonetic units, which correspond roughly to the onset, the static part, and the end of each phone. The underlying subphone HMM is a simple left-to-right three state HMM with self-loops but without skips. The silence phone is a one-state model with self-loop. All transition probabilities are equal, and no training of the transition probabilities took place.

5.3 Acoustic and language models

The system uses context-dependent HMM acoustic models for each of the sub-phonetic units. To achieve optimal state tying, a decision tree is constructed (Lee, 1988) for each of the sub-phonetic units using the training data and a set of linguistically oriented questions with a maximum context width of ± 1 phone. Therefore, each leaf of the decision tree represents a different set of triphone contexts of the underlying sub-phonetic unit. The acoustic model for each leaf is a mixture of 30 gaussians with diagonal covariances.

The language model of the View4You recognizer is a standard Kneser-Ney backoff (Kneser & Ney, 1995) trigram language model built on the data described in table 4. The test set perplexity of the language model is 381. The OOV rate on the same test set is 4.43%. The language model

contains an UNKNOWN token which was also used in the computation of the perplexity.

Corpus	Number of sentences	Number of words
tagesschau	10.442	113.328
Germany Live	244.681	3.821.552
Bayern 5	150.689	2.119.375
FAZ	2.100.603	39.668.914
SZ 1995	1.839.732	28.582.448
SZ 1996	1.862.219	28.531.746
sum	6.208.366	102.837.363

Table 4: Text material used for the statistical 3-gram model

5.4 Decoder

In evaluation mode, the decoder works in 7 passes. In the first pass, a tree structured vocabulary without tree copies selects probable words for each starting point. This pass uses only delayed bigrams and trigrams. Using the hypothesis from this pass, the vocal tract length normalization parameter α is estimated, and the first pass is repeated with the normalized input features.

The third pass uses a flat, linear structured vocabulary allowing full bigrams and a better trigram approximation. In the fourth pass, the resulting back pointer table from the third pass is pruned and converted into a word lattice which can be rescored using the full trigram language model. The first best hypothesis from the rescored lattice is the intermediate output of the system. The *gamma* (Kemp & Schaaf, 1997) confidence measure is applied to the intermediate output, and all words with an a-posteriori likelihood of more than 0.5 are used to estimate three MLLR (Legetter & Woodland 1995) transformation matrices. With the adapted models, steps (2) to (4) are repeated. The output from the final lattice rescoring pass with both VTLN and MLLR adaptation is taken as the final output of the system.

5.5 Effect of background noise on recognition performance

To evaluate the effect of the different types of noise found in the data, the word error rate for each of the noise conditions was measured. For this, one single recognizer was trained on 15.5 hours of manually transcribed broadcast news speech, i.e. no optimization with respect to noise or channel condition was performed. The recognition results, which include both VTLN and MLLR adaptation, are summarized in table 5.

category	word error rate
anchor speaker	11.9%
music	20.1%
street noise	23.0%
conference noise	27.9%
single 2nd speaker	30.5%
other noise	29.5%

Table 5: Dependency of error rate from background noise

The highest word error rates are observed if there is one or more other speakers in the

background. This condition occurs typically when a public speech or announcement, given by a non-German speaker, is reported. The original - non-German - speech is lowered in volume while the German translation is played in the foreground. This combination of two speakers turns out to be particularly harmful to recognition performance. Background music has relatively low impact on word error rate, probably due to low confusability with actual speech. Conference noise, although consisting basically of the noise of many voices in the background under reverberating conditions, harms less than a single second speaker. This surprising result can be attributed to the fact that in our data this background noise level is typically lower than the volume of the 'single second speaker' condition.

Street noise (demonstrations, riots, whistles, singing and the like) is the most common background noise condition in our data. It is, however, significantly easier to recognize than background speakers.

6 The information retrieval (IR) engine

We chose the Okapi similarity measure (Beaulieu et al, 1997) for our experiments. This measure has been evaluated thoroughly in the context of NIST's TREC information retrieval contests (TREC), and has been found to be very powerful. The Okapi measure can be parameterized to meet the special requirements of a given task. We use a parameterization that has been found to be very good for short queries (Wilkinson et al, 1995):

$$d(q,d) = \sum_{t \in Q \wedge t \in d} \left(\frac{f_{d,t}}{f_{d,t} + \sqrt{\frac{f_d}{E(f_d)}}} \right) \log \left(\frac{N - f_t}{f_t} \right)$$
$$= \text{Okapi}(k_1 = 1, k_2 = 0, k_3 = 0, b = 1, r = 0, R = 0)$$

where $E(.)$ denotes the expected value, N is the number of documents in the collection, f_t is the number of documents containing term t, $f_{d,t}$ is the frequency of term t in document d, and f_d is the number of terms in document d, which is an approximation to the document length. A *term* in this context is the same as a word, however, the 500 most frequent words ('I', 'other' and the like) are excluded. Morphological stemming is applied to both the query and the database records. The database engine computes the distance between a query and each article in the database and returns the articles sorted in decreasing order of similarity to the query.

7 End-to-end evaluation

The View4You system is a fully operational prototype of a video indexing machine. As only very few such systems exist world-wide, it is particularly interesting to evaluate the system as a whole, as opposed to evaluating each of its components. In such an analysis, the imperfections of each component - segmenter, speech recognizer, and Information Retrieval engine - can be assessed and the impact on the end to end performance can be evaluated. In this section, we describe the results of our end to end evaluation experiments.

7.1 Evaluation methodology

For the evaluation, a database was automatically created by the View4You system by processing a set of TV broadcasts. Both steps of the database creation process - segmentation and speech recognition - were carried out without any human intervention.

A set of questions in natural language was defined (see below).

To generate the reference, the database was segmented manually in two steps. In the first step, the newscasts were segmented into topic stories, where a segment boundary was inserted

only if the (semantical) *topic* of the segment changed. In the second step, the so defined segments were further segmented with respect to their acoustic background. A segment boundary was introduced whenever the acoustic background changed significantly, e.g. from anchor speaker to field speech, or from conference noise to battlefield noise (see figure 3).
Each of the so defined segments was then manually judged relevant or irrelevant with respect to each of the test queries.

During the evaluation, each of the queries was presented to the system and the segments that were returned by the system were compared with the list of the segments that were judged relevant for this query. However, the interpretation of the results is not trivial. Most information retrieval algorithms assign a *relevance* score to each of the segments, and return a list of segments sorted by this relevance score. This list can sometimes comprise the whole database. The evaluation result therefore depends on the number of resulting segments which are taken into account during the evaluation. This number, however, cannot be chosen a priori, but depends on the - generally unknown - number of 'true hits' in the database.

Due to this problem, there is no generally accepted single way to represent the results of an information retrieval evaluation. In the text retrieval literature (see, e.g., (TREC)), the results are usually presented in one or more of the following ways:

- A plot of Precision (PRC) over Recall (RCL)
- average Precision (avePRC), and
- R-Precision.

Average precision is defined as the average of the eleven values of PRC at Recall 0, 0.1, 0.2, ..., 1.0. If a given Recall value cannot be accurately achieved, the corresponding precision is determined by interpolation. There is some freedom of choice for the PRC at Recall zero, which can be defined more or less arbitrarily. In our evaluation, we chose the PRC value computed at the first found item.
In our parameterization of the Okapi algorithm, many segments are assigned a relevance score of zero. The maximum possible value RCL_{max}, therefore, is obtained if all segments that have a nonzero relevance score are included into the evaluation. When computing average precision, the values for PRC are set to zero for all RCL values higher than RCL_{max}.
R-Precision is defined as the PRC value which is obtained when the number of segments evaluated is set to the number of 'true hits' in the database (with respect to the current query). At this operating point, PRC equals RCL and, from [1], also F.

7.2 Evaluation with different segmentation

All evaluation measures for information retrieval performance depend on the underlying segmentation. If the segment boundaries are different between the reference segmentation and the segmentation in the database under evaluation, a retrieved segment can be *partially* overlapping with a true segment. The question is whether such a segment should be judged 'correct' or 'in error'. If it is judged 'correct' merely if it fully covers a correct segment, the (degenerate) case of no segmentation will usually provide the best result, which is clearly undesirable.
To overcome this problem, we computed Precision and Recall based on time rather than on segment basis. This is comparable to having a 'segmentation' into very short chunks of equal size, say one second each, together with a manual judgement whether each of these one-second chunks is relevant to the query or not. A result of RCL=0.7 does, therefore, not mean that 60% of all segments have been correctly retrieved, but that 60% of the audiovisual material that is available in the database has been correctly retrieved. This kind of evaluation has the additional

advantage, that long segments, which tend to carry more information than short ones, are implicitely weighed higher when PRC and RCL are computed. Evaluation by time also penalizes under-segmentation. If a segment is too long and spans both over the real segment and over parts of the neighbouring segments, the PRC will degrade in a time-based evaluation paradigm - in contrast to the classical segment-based paradigm. Since the user does not want to skim through irrelevant material, it seems reasonable to reflect such an effect in the result.

All evaluations described in this paper were therefore carried out as follows.
Each of the segments from the reference was temporally aligned and transformed into a time range. For each of the retrieved segments, the time covered by this segment was computed and compared to the reference. Both the overlap - where the retrieved segment was judged relevant - and the non-relevant portions of the retrieved segments are accumulated in separate variables. This process is repeated for each segment under consideration. At the end, PRC and RCL are computed from the total time spent in correctly retrieved portions of video, the total time of relevant material in the database (taken from the reference), and the total time spent in portions of the video that were not relevant wrt the query.

7.3 Averaging over the queries

Both PRC and RCL are well defined for a single query. For a set of queries, however, the results can be averaged in one of two different ways. In the *user-oriented view*, the values of PRC and RCL - one of each per query - are averaged. The result can be interpreted as the expected values for PRC and RCL for the next query. In the *system view*, the relevant time that was correctly found (and, separately, the irrelevant time presented to the user) are added up over all queries. PRC and RCL are computed from this global sum. This view can be interpreted as the expected value for PRC and RCL for each second of the system's output. If the amount of data which is available in the database has the same length for all queries, the user-oriented and the system view are identical. Otherwise, the system view puts more emphasis towards queries where more material is available in the database.
The choice for one of the views is dependent on the application. In our research, we chose the user-oriented view to avoid a bias towards or against any of our queries.

Question set

The set of questions that was used in the evaluation was determined in the following way. Several naive users were asked to query the system, and their questions were used in the evaluation. The questions (translated from German) are summarized in table 6.

1	Are there reports about Jerusalem?
2	Will Helmut Kohl run for chancellor again?
3	I want to know the winning numbers of the Lottery!
4	Is there anything about Benyamin Netanyahu?
5	I am interested in anything that recently happened in Africa!
6	What is the situation in Albany?
7	What are the results of the National Soccer League?
8	Are there any reports about refugees?
9	I'd like to see reports about President Herzogs visit to Japan!
10	Is there anything new in the Mykonos trial?

Table 6: Questions used in the evaluation

Test database

For the evaluation experiments, 10 news shows broadcast between April 3, 1997 and April 12, 1997, were used. This test database contained 174 different stories (when segmenting on the semantical level), and 365 segments when segmentation according to background noise condition was performed. 62 stories or 163 segments were judged relevant with respect to one or more of the queries.

The test database is relatively small. Therefore, in the first set of experiments, we did not compute average precision and R-precision, but carried out the evaluation based on all segments that were assigned a nonzero relevance by the system. This operating point is the one that would be chosen by a user who puts his emphasis on recall.

7.4 Result with manual segmentation and error-free speech recognition

In a first experiment, we used the manually generated story boundaries as segmentation, and simulated a 100% speech recognition accuracy by using the (manual) transcriptions. In this situation, all errors that are being made by the system originate from imperfections of the information retrieval engine. In a second experiment, we further segmented the semantical stories into portions of homogeneous noise conditions, e.g., into one part from the anchor speaker and one other part from an interviewer in the field. This segmentation was performed by hand, simulating a perfect acoustic segmenter. Again, the transcriptions were used to simulate a perfect speech recognizer. The results are summarized in table 7.

Segmentation	PRC	RCL	F
manual (stories)	0.72	0.79	0.75
manual (acoustic segments)	0.84	0.46	0.60

Table 7: Result with different manual segmentation using transcriptions

Even with perfect segmentation and perfect speech recognition, the performance of the system is not good. This can be attributed to two main factors.

Reasons for low recall (RCL) The main reason for low recall is the lack of background knowledge. A segment about a flood in Uganda is not considered relevant with respect to the keyword 'Africa', because the system does not know that Uganda is situated in Africa. Similarly, the system does not see a relationship between the Soccer league and a report about the 'game between Karlsruher SC and Hertha BSC Berlin'. To overcome this problem, a thesaurus can be used.

Reasons for low precision (PRC) The main reason for low precision is the lack of semantical analysis of the user query. The information retrieval algorithm treats each word in the query as an independent, implicitely OR'ed term. For some queries, however, this is inadequate. Most of the degradation in precision in the query about, e.g., President Herzogs visit in Japan, is due to found reports about President Herzogs visits elsewhere, or reports about Japan, or even reports about other Presidents.
The easiest way to overcome this problem would be to put the load onto the user by requiring him to enter a boolean expression instead of a natural language query (e.g. "President Herzog" AND "Japan" AND "visit").

7.5 Result with manual segmentation and speech recognition

We evaluated the system using the same manual segmentation into acoustically defined segments as before, but used a speech recognizer to generate the transcripts. The word error

rate of the speech recognizer, determined on an independent test set, was 22.7%. The result is shown in table 8.

Segmentation	SR error rate	PRC	RCL	F
manual (acoustic segments)	0% (transcriptions)	0.84	0.46	0.60
manual (acoustic segments)	22.7%	0.81	0.43	0.56

Table 8: Result with manual segmentation and automatic speech recognition

Although every fifth word in the resulting hypothesis is wrong, the end to end performance is not seriously affected. This finding is in accordance with results published by the Informedia group at CMU (see, e.g., (Wactlar et al, 1996)).

7.6 Result with automatic segmentation and speech recognition

In another set of experiments, the impact of the segmentation on the end to end performance was evaluated. For this, two different segmentation algorithms - the model based algorithm and the hybrid algorithm - were applied to the input data. In both cases, automatic speech recognition was used to obtain the transcriptions. The result is given in table 9.

Segmentation	PRC	RCL	F
model-based	0.78	0.34	0.48
hybrid	0.68	0.44	0.54

Table 9: Result with different automatic segmenters and automatic speech recognition

The quality of the segmentation strongly affects the end to end performance of the system.

Effect of the operating point of the segmenter The segmenter can be parameterized to find most of the true boundaries at a relatively high level of false boundary hypotheses, or to find only a smaller percentage at a low level of false alarms. It is interesting to see how this parameterization affects the end to end performance. In table 10, the results are given for three different parameterizations of the segmenter.

Segmenter PRC	Segmenter RCL	Segmenter F	end to end PRC	end to end RCL	end to end F
0.687	0.851	0.760	0.68	0.44	0.54
0.722	0.738	0.730	0.58	0.48	0.53
0.947	0.674	0.787	0.57	0.54	0.55

Table 10: End to end results results with different operating points of the hybrid segmenter

Obviously, an increase in the F-measure of the segmenter translates into an increase of the end to end performance of the system as a whole.

7.7 Effect of the speech recognition performance

In order to shed some additional light on the effect of the speech recognition performance on the end to end result, we repeated the evaluation using two different speech recognition systems. The first one was the best system we had available at this time, running at an overall error rate of 21.5% on the independent test set. The other one was taken from an early experimental stage and performed significantly worse (at 33.7% word error rate) The results with the three different speech recognizers are given in table 11.

Segmentation	SR error rate	PRC	RCL	F
model-based	21.5%	0.78	0.37	0.50
model-based	23.7%	0.78	0.34	0.48
model-based	33.7%	0.77	0.29	0.42

Table 11: Results using different speech recognition systems

7.8 Summary

Table 12 summarizes the results.

Segmentation	SR error rate	PRC	RCL	F
manual (semantic stories)	0% (transcriptions)	0.72	0.79	0.75
manual (acoustic segments)	0% (transcriptions)	0.84	0.46	0.60
manual (acoustic segments)	23.7%	0.81	0.43	0.56
hybrid	23.7%	0.68	0.44	0.54
model-based	23.7%	0.78	0.34	0.48
model-based	21.5%	0.78	0.37	0.50
model-based	33.7%	0.77	0.29	0.42

Table 12: Summary of results

From this, we draw the following conclusions.

- The highest impact on the F-score (25% absolute) is due to the information retrieval component of the system.
- Segmentation is important: switching from semantic stories to acoustically defined segments degrades the performance by 15% absolute. Using a suboptimal segmentation algorithm degrades the performance by another 8% absolute.
- A speech recognizer with an error rate of 23.7% causes a degradation of only about 4%. However, at 33.7% word error rate, the loss is roughly 10%.

7.9 Increasing the size of the database

It is of considerable interest to study the effect of a larger database on these results. For this, we manually segmented, transcribed, and judged (with respect to the queries) another 55 news show recordings, broadcast between Summer 1997 and Summer 1998. We then re-ran the experiments on this larger testset of approximately 15 hours of speech. As a result of the increased size of the database, we could no longer use the whole list that was returned by the system in the evaluation. The results are therefore given in terms of R-precision. It should, however, be kept in mind that all results are time-based rather than segment-based to allow comparisons between different segmentations.

The new results are summarized in table 13.

Segmentation	SR error rate	R-precision (equals F)
manual (acoustic segments)	0% (transcriptions)	0.45
manual (acoustic segments)	22.7%	0.43
hybrid	22.7%	0.39
model-based	22.7%	0.30

Table 13: Summary of results on 65 news shows

As expected, the performance on the larger database is worse than on the small one. Much of this effect can be attributed to the nature of the questions, which were aimed to a one-time event. In such cases, the addition of more material to the database will necessarily degrade performance, as only new opportunities for errors are added.

8 Conclusion and Discussion

The end to end performance of our View4You video indexing and retrieval system is relatively low (F = 0.4 on a large database), despite our use of state-of-the-art algorithms for information retrieval, speech recognition, and segmentation. We can attribute the lion's share of the errors to the information retrieval part of the system, followed by the segmentation into scene cuts rather than semantically oriented topics. The errors of the speech recognizer do not play a dominant role if the word error rate is below 25 - 30%. With a good segmentation algorithm, the influence of errors of the segmenter (which does not always identify the scene cuts correctly) can also be made small.

We expect further improvements of the system from analyzing the natural language query into boolean expressions, and from the use of a thesaurus. Currently, therefore, our research is aimed into these directions.

9 Acknowledgements

This work was carried out at the Interactive Systems Labs, Karlsruhe. The authors would like to thank all members of the Interactive Systems Labs for helpful discussions and support. The views and conclusions contained in this document are those of the authors.

References

M.M. Beaulieu, M. Gatford, X. Huang, S.E. Robertson, S. Walker, P. Williams (1997), *Okapi at TREC-5*, Proc. of the 5th Text Retrieval Conference, NIST, Gaithersburg, MD

P. Beyerlein, X.L. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, M. Pitz, A. Sixtus (1999), *The Philips/RWTH System for Transcription of Broadcast News*, Proc. ICASSP 99

J. Billa, K. Ma, M. Siu, G. Zavaliagkos (1997), *Acoustic modeling work at BBN*, in Proc. of the Hub-5 Conversational Speech Recognition workshop, NIST, Linthicum Heights, Maryland

S.S. Chen, P.S. Gopalakrishnan (1998), *Speaker, environment and channel change detection and clustering via the bayesian information criterion*, in Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 127 ff, Landsdowne Conference Resort, Landsdowne, VA

S.S. Chen, E.M. Eide, M.J.F. Gales, R.A. Gopinath, D. Kanevsky, P. Olsen (1999), *Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News*, Proc. ICASSP 99

J. Choi, D. Hindle, J. Hirschberg, I. Magrin-Chagnolleau, C. Nakatani, F. Pereira, A. Singhai, S. Whittaker (1998), *SCAN-Speech Content Based Audio Navigator: A System Overview*, Proc. ICSLP 98

K. Fukunaga (1990): *Introduction to statistical pattern recognition*, Academic Press Inc., San Diego, CA 92101, ISBN 0-12-269851-7, San Diego, 1990

J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker (1997), *Transcribing Broadcast News Shows*, Proc. ICASSP 97, pp. 715 ff

J.-L. Gauvain, L. Lamel, G. Adda (1998), *The LIMSI 1997 Hub-4E Transcription System*, DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA

Alexander G. Hauptmann and Howard D. Wactlar (1997), *Indexing and Search of Multimodal Information*, Proc. ICASSP 97, pp. 195 ff

Kai-Fu Lee (1988), *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*, Ph. D. thesis, CMU-CS-88-148, Carnegie Mellon University, Pittsburgh, PA 15213

T. Kemp, T. Schaaf (1997), *Estimating confidence using word lattices*, in Proc. EUROSPEECH-97, Vol 2, pp. 827 ff, Rhodes, Greece

T. Kemp, M. Schmidt, M. Westphal, A. Waibel (2000), *Strategies for automatic segmentation of audio data*, in Proc. ICASSP 2000, Istanbul, Turkey

R. Kneser, H. Ney (1995), *Improved Backing-Off for M-Gram Language Modelling*, in Proc. ICASSP 95, Detroit

F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, R. Schwartz (1997), *Brodcast News Transcription*, Proc. ICASSP 97, pp. 203 ff

C.J. Legetter, P.C. Woodland (1995): *Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models*, Computer Speech and Language **9** (1995), 171-185

National Institute of Standards (NIST), *Proceedings of the DARPA Broadcast News transcription and understanding workshop*, Lansdowne, VA, February 8-11, 1998

L. Polymenakos, P. Olsen, D. Kanvesky, R. Gopinath, P. Gopalakrishnan, S. Chen (1998), *Transcription of broadcast news - some recent improvements to IBM's LVCSR system*, Proc. ICASSP 1998, pp. 901 ff, Seattle, Washington

C.J. Van Rijsbergen (1979), *Information Retrieval*, London, Butterworth, pp. 174 ff.

A. Sankar, F. Weng, Z. Rivlin, A. Stolcke, R. Gadde (1998), *The development of SRI's 1997 broadcast news transcription system*, DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA

M. Siegler, U. Jain, B. Ray, R. Stern (1997), *Automatic segmentation, classification and clustering of broadcast news audio*, Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 97ff, The Westfields Conference Center, Chantilly, VA

http://www-nlpir.nist.gov/TREC/

H. Wactlar, A. Hauptmann, M. Witbrock (1996): *Informedia: news-on-demand experiments in speech recognition*, Proc. of ARPA SLT workshop 1996

H. Wactlar, M. Christel, Y. Gong, A. Hauptmann (1999), *Lessons Learned from Building a Terabyte Digital Video Library*, **Computer**, February 1999, pp. 66-73

S. Wegmann, F. Scattone, I. Carp, L. Gillick, R. Roth, J. Yamron (1998), *Dragon System's 1997 broadcast news transcription system*, DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA

S. Wegmann, P. Zhan, L. Gillick (1999), *Progress In Broadcast News Transcription at Dragon Systems*, Proc. ICASSP 99

R. Wilkinson, J. Zobel, R. Sacks-Davis (1995): *Similarity measures for short queries*, in Proc. of TREC-4, NIST

P.C. Woodland, M.J.F. Gales, D. Pye, S.J. Young (1997), *Broadcast News Transcription using HTK*, Proc. ICASSP 97, pp. 719 ff

P.C.Woodland, T. Hain, S. Johnson, T. Niesler, A. Tuerk, S. Young (1998), *Experiments in broadcast news transcription*, Proc. ICASSP 1998, pp. 909 ff, Seattle, Washington

S.J.Young, M.G. Brown, J.T. Foote, G.J.F. Jones, K. Sparck Jones (1997), *Acoustic Indexing for Multimedia Retrieval and Browsing*, Proc. ICASSP 97, pp. 199 ff

P. Zhan, M. Westphal (1997), *Speaker normalization based on frequency warping*, in Proc. ICASSP-97, Munich