THE INTERACTIVE SYSTEMS LABS VIEW4YOU VIDEO INDEXING SYSTEM

Thomas Kemp *Petra Geutner* *Michael Schmidt* *Borislav Tomaz* *Manfred Weber*
Martin Westphal *Alex Waibel*

Interactive Systems Laboratories, ILKD
University of Karlsruhe
76128 Karlsruhe, Germany

ABSTRACT

The recognition of broadcast news is a challenging problem in speech recognition. To achieve the long-term goal of robust, real-time news transcription, several problems have to be overcome, e.g. the variety of acoustic conditions and the unlimited vocabulary. Recently, a number of sites have been working on content-addressable multi-media information sources. In the presented paper, we focus on extending this work towards a multi-lingual environment, where queries and multimedia documents may appear in multiple languages. In cooperation with the Informedia project at CMU [4], we attempt to provide cross-lingual access to German and Serbo-Croatian newscasts.

1. THE VIEW4YOU SYSTEM

In the View4You system, German and Serbocroatian public newscasts are recorded daily using standard consumer electronics equipment. 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.

In this work, we give an overview over the three main parts of the View4You system, namely the segmenter, the speech recognizer, and the information retrieval engine.

2. SYSTEM OVERVIEW

Figure 1 shows a block diagram of the View4You prototype system. The 15-minute newscast is first segmented into segments of approximately 10 to 90 seconds. 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. The newspaper articles can be used to adapt the vocabulary of the recognizer [1].
User queries are possible 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, found video segments are presented as thumbnail pictures of the beginning of the video segment, and found newspaper texts are presented as 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.

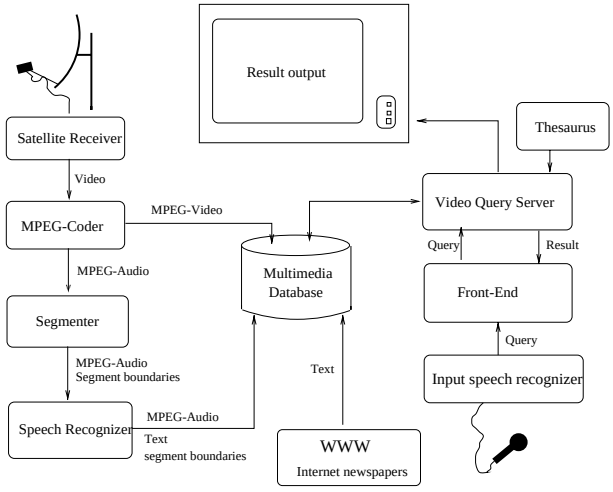

Figure 1. System overview

In the following paragraphs, we describe the training database, the segmenter, the speech recognizer, and the retrieval engine in detail.

3. THE BROADCAST NEWS DATABASE

Data processing

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 layer 2 audio 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 speech frequency range. 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.
The recorded audio signal is then sampled down to the final 16 kHz/16 bit PCM format which was used throughout all our 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 construction of the database, the segmenting, or the training of the speech recognizer.

Training data

For training the acoustic models of our recognizer, we used 8 news shows broadcast between November 25, 1996 and December 17, 1996. The training set is divided into 346 segments. Only 309 of these segments contain speech; the other segments mostly contain only music. There is a total of 16850 words in the training set, or 7153 seconds of

speech. Roughly 60% (4209 seconds) is from male speakers. Less than 50% of the data (3428 seconds) is anchor speaker speech. The distribution of the rest of the data into the different noise categories (determined on the test set) is shown in table 1. Note that one segment may belong to more than one noise category.

category	seconds	percent
street noise	1050	64%
conference-like noise	305	18%
single 2nd speaker	462	28%
music	257	15%
other noise	90	5%

Table 1. Distribution of background noise (test data)

There is only a limited number of different anchor speakers. In our training set, five of the eight shows are moderated by the same (female) anchor named Dagmar Berghoff. The remaining three broadcasts are spoken by three different male speakers.

The 4 test shows have been broadcast on 30/03/1997 (female unknown anchor), 13/04/1997 (female unknown anchor), 26/05/1997 (male unknown anchor) and 30/06/1997 (known male anchor).

3.1. Segmentation

The data is manually segmented into segments of uniform acoustic condition, i.e. a segment boundary is introduced if there is a change in the acoustic condition. Examples for such changes are scene cuts from anchor speaker to a field report or from one field report to another with different background noise. Additional boundaries were introduced for topic changes, even if there was no change in the acoustic condition. Therefore, it is impossible to detect all segment boundaries automatically without analyzing the meaning of the segment's content.

3.2. The OOV problem

The index into View4You's video database consists of the output of our speech recognizer. Therefore, only words that are in the vocabulary of the recognizer can be searched for. If a video contains keywords that are unknown to the recognizer, they cannot be found in the index, and the user can not retrieve the video by this keyword. OOV (out-of-vocabulary) words therefore pose a problem to the View4You system, and the vocabulary of the speech recognizer should be as large as possible to ensure low OOV rates. Currently, our speech recognizer is limited to a vocabulary of 64k words. On the North American Business News (NAB) corpus, a vocabulary of this size covers more than 99% of the text, and even with a 20k vocabulary, OOV rates on NAB do not exceed 3%.

We measured the OOV rate on German news shows for a 60k vocabulary which was derived from our language model corpus. The result is shown in picture 2.

As can be seen from the graph, the average OOV rates in German broadcast news are approximately 3 to 4 times higher than in the English NAB task. There are two reasons for this. First, as news topics have a high variety, the vocabulary of broadcast news shows generally exceeds the vocabulary of business news text, regardless of the language used [4]. Second, the German language allows the construction of compound nouns, like e.g. 'kindergarten' from 'kinder' (children) and 'garten' (garden). Words like these form an open set and can not be covered by a static vocabulary of any reasonable size.

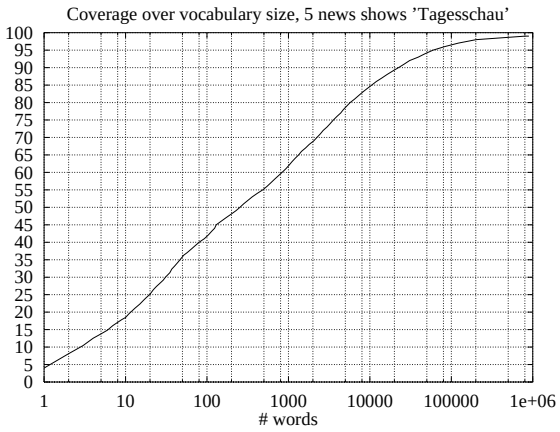

Figure 2. Coverage over vocabulary size for the test set

4. SEGMENTER

The goal of the segmenter is to cut the large 15 minute recording into smaller segments of uniform acoustic condition. The View4You segmenter uses a simple four-state HMM model. The four states are associated with the four major acoustic conditions: anchor speaker, field speech, music, and silence. The four states are fully interconnected with uniform transition probabilities. Each state is modeled as a mixture of gaussians with diagonal covariances. The acoustic features are 16 mel-filterbank coefficients per 16 ms frame. The frame shift was chosen to 50 ms. The silence state is modeled with 2 gaussians, the noise state with 32 and both the anchor speaker and the field speech state use 128 gaussians. The result in terms of precision (PRC) and recall (RCL) of found boundaries on the testset is summarized in table 2.

preprocessing	PRC	RCL
16 mel filterbank	33.5%	89.2%
16 mel cepstral	27.0%	56.2%

Table 2. Performance of the segmenter

5. THE VIEW4YOU SPEECH RECOGNIZER

In this section, we give a detailed description of the View4You speech recognizer.

Preprocessing

In the preprocessing stage, 13 cepstral parameters per 16 ms frame are computed from 30 melscale filter bank coefficients. The frame shift is chosen as 10 ms. A simple energy-based speech detection is performed, and cepstral mean subtraction is applied to the speech segments 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 to the final 16-component feature vector. In all experiments described, we made use of vocal tract length normalization [7], which warps the power spectrum to a reference vocal tract length before the computation of melscale filter bank coefficients takes place.

Dictionary and vocabulary

The recognition dictionary contains the 60788 most frequent words from the language model training corpus described below. For all but the most frequent 2000 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 made up of non-anchor speech, is partly spontaneous. To capture some of the effects of spontaneous speech, e.g. hesitations, five specialized noise models were added to the phone set: two hesitation models, one 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 beginning, middle, and end of each phone. The underlying phone 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.

Acoustic models

The system uses context-dependent HMM acoustic models for each of the sub-phonetic units. A decision tree is constructed 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. 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 gaussian pdf's with diagonal covariances. The number of gaussians in each mixture and the number of triphones (i.e., the number of leafs of the decision tree) can be varied in our recognizer. In the next paragraph, recognition results with different numbers for these parameters are shown.

Decoder

The decoder works in 3 passes. In the first pass, a tree structured vocabulary without tree copies selects probable words for each starting point. This pass uses only approximate bigrams and trigrams. The second pass uses a flat, linear structured vocabulary allowing full bigrams and a better trigram approximation. In the third pass, the resulting back pointer table from the second 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 final output of the system.

Language models

The language model of the View4You recognizer is a standard Kneser-Ney backoff trigram language model built on the concatenation of two corpora ('ONLINE-0' and 'FAZ'). The structure of the language model training corpus is summarized in table 3.

database	time covered	size (kWords)
ONLINE-0	20/06/96 to 28/02/97	6052
FAZ	1992-1994	39669
total	1992 to 28/02/1997	45721

Table 3. Corpora used for language modelling

ONLINE-0 contains parts of the training transcriptions, transcriptions of radio news and text data from internet newspapers. The FAZ corpus contains one year's worth of data from a major german newspaper.

Using specialized recognizers

As the test data can be divided into a non-disturbed, clean anchor speaker part and a noisy, channel-dependent part, we first decided to build two specialized recognizers for each of this two conditions. However, the error rate of the specialized recognizers was higher than that of a system trained on all data. This is probably due to the low amount of training material available. Therefore, we decided to use a single recognizer for all types of data.

To evaluate the effect of the different types of noise found in the data, we computed the word error rate for each of the noise conditions separately. The results are summarized in table 4. The highest word error rates are observed if there is one or more other speakers in the background.

category	word error rate
anchor speaker	25.1%
street noise	39.0%
conference noise	50.1%
single 2nd speaker	50.5%
music	41.6%
other noise	37.4%

Table 4. Dependency of error rate from background noise

Parameter allocation

In a set of experiments, we trained several systems that were clustered to different numbers of polyphones. Each polyphone was modeled as a mixture of 30 gaussians without any parameter tying. The resulting word error rates on the anchor speaker segments of one news show are shown in figure 3. A system with approximately 1000 polyphones (or about 15 data frames per gaussian) yielded the best results.

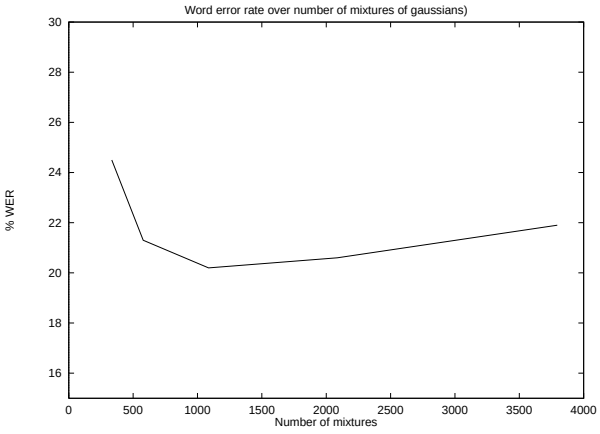

Figure 3. Error rate over number of polyphones

Recognition results

We ran our recognizer on a test set of two news shows, using vocal tract length normalization. The transcriptions were used to compute three MLLR adaptation matrices which were used to adapt the mean vectors of our system. With the adapted system, a final three-pass decoder run was performed. The results are shown in table 5.

Shortly before the printing of this paper, 11 additional transcribed newscasts (3 hrs of speech) and a text corpus with 150 million words became available. Utilizing this new material, significant improvements could be achieved. Preliminary results of our most recent system are shown in table 6.

show (date)	Anchor	non-anchor	total
30/03	20.2%	41.0%	30.6%
13/04	22.7%	44.5%	33.6%
total	21.5%	42.8%	32.2%

Table 5. View4You recognizer word error rates

show (date)	Anchor	non-anchor	total
30/03	16.2%	30.0%	23.1%
13/04	17.8%	33.0%	25.4%
total	17.0%	31.5%	24.3%

Table 6. Most recent system's word error rates

6. THE INFORMATION RETRIEVAL (IR) ENGINE

We built our information retrieval engine using the Okapi similarity measure [3]. This measure has been evaluated thorougly in the context of NIST's TREC information retrieval contests [2], and has been found to be especially powerful. The Okapi measure can be parameterized to the special requirements of a task. We use a parameterization that has been found to be very good for short queries [6]:

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

where 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. 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.

6.1. End-to-end evaluation

We measured the end-to-end performance of the View4You system using ten different queries (e.g. 'Are there any reports about Jerusalem?', 'I want to see reports about the visit of president Herzog in japan'). Including all errors introduced by the segmenter, the speech recognizer and the information retrieval system itself, the average precision of the retrieved segments was 0.48. Although this may seem low at first glance, a human user of the system can usually determine the appropriateness of a segment by the thumbnail picture used to represent it or by viewing just the first few seconds. Therefore, most of our test persons judged the system useful despite its pre-mature state.

7. THE USER INTERFACE

Picture 4 shows the user interface of the prototype system. On the right side, the segments that have been found to the query 'Informationen ueber Albanien und Italien bitte' ('Please give informations about Albania and Italy'), sorted by decreasing relevance. The search result with the highest relevance score is selected and is currently being played (in the left of the screen).

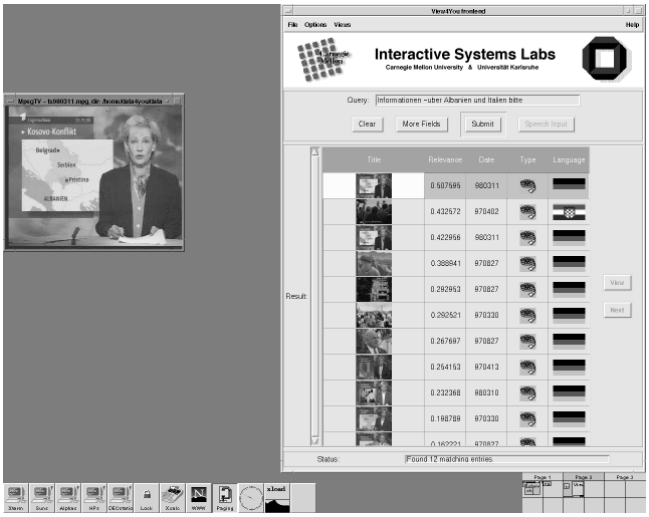

Figure 4. User interface

8. 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

[1] T. Kemp, A. Waibel, *Reducing the OOV rate in broadcast news speech recognition*, elsewhere in these proceedings

[2] http://www-nlpir.nist.gov/TREC/

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

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

[5] P. Placeway, J. Lafferty: *Cheating with imperfect transcripts*, in Proc. ICSLP 96, Philadelphia, September 1996

[6] R. Wilkinson, J. Zobel, R. Sacks-Davis: *Similarity measures for short queries*, in Proc. of TREC-4, NIST, November 1995

[7] P. Zhan, M. Westphal, *Speaker normalization based on frequency warping*, in Proc. ICASSP-97, Munich, April 1997