The LIMSI SDR System for TREC-8

Jean-Luc Gauvain, Yannick de Kercadio, Lori Lamel and Gilles Adda

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gauvain,kercadio,lamel,gadda}@limsi.fr

ABSTRACT

In this paper we report on our TREC-8 SDR system, which combines an adapted version of the LIMSI 1998 Hub-4E transcription system for speech recognition with an IR system based on the Okapi term weighting function. Experimental results are given in terms of word error rate and average precision for both the SDR'98 and SDR'99 data sets. In addition to the Okapi approach, we also investiged a Markovian approach, which although not used in the TREC-8 evaluation, yields comparable results. The evaluation system obtained an average precision of 0.5411 on the reference transcriptions and of 0.5072 on the automatic transcriptions. The word error rate measured on a 10 hour subset is of 21.5%.

INTRODUCTION

There expansion of different media sources for information dissemination (radio, television, internet) has led to a need for automatic processing tools. Todays methods for audio segmentation, transcription and indexation are manual, with humans reading, listening and watching, annotating topics and selecting items of interest for the user. Even partial automation of some of these activities can allow more information sources to be processed and significantly reduce processing costs while eliminating tedious work. Some application areas that could benefit from automated transcription and indexing technology include the creation and access to digital multimedia libraries (disclosure of the information content and content-based indexation, such as are under exploration in the OLIVE [13] project), media monitoring services (selective dissemination of information based on automatic detection of topics of interest) as well as new emerging applications such as News on Demand (such as the Informedia [10] project) and Internet watch services. Such applications are feasible due to the large technological progress made over the last decade, benefiting from advances in micro-electronics which have facilitated the implementation of more complex models and algorithms.

Automatic speech recognition is a key technology for audio and video indexing, for data such as radio and television broadcasts. Most of the linguistic information is encoded in the audio channel of video data, which once transcribed can be accessed using text-based tools. This is in contrast to the image data for which no common description language is available.

In this paper we describe the LIMSI spoken document indexing and retrieval system developed for the TREC-8 SDR evaluation. This system combines a state-of-the-art speech recognizer [9] with an Okapi-based IR system. A Markovian-based IR system has also been developed and contrastive experimental results using this system are provided. All of our development work was carried out using the TREC-7 SDR data set (100 hours) and the associated set of 23 queries. This year's SDR task was quite more challenging than the SDR'98 track in that the audio data was increased to about 550 hours of broadcasts, which has strong implications on the transcription process. The next section describes the LIMSI speech transcription system and the modifications made for use in this evaluation, trying to find the best compromise between accuracy and speed. In the following section the two IR systems are presented, and experimental results for various configurations are provided.

TRANSCRIBING BROADCAST NEWS

At LIMSI we have been working on using statistical models to transcribe broadcast news data since 1996. Due to the availability of large audio and textual corpora via the Linguistic Data Consortium (LDC)[1], most of our work on broadcast news transcription has been carried out on American English. In the context of the EC LE OLIVE project [13], broadcast news transcription systems for French and German have recently been developed.

Radio and television broadcast shows are challenging to transcribe as they contain signal segments of various acoustic and linguistic natures. The signal may be of studio quality or may have been transmitted over a telephone or other noisy channel (i.e., corrupted by additive noise and nonlinear distortions), or can contain speech over music or pure

[1] http://www.ldc.upenn.edu

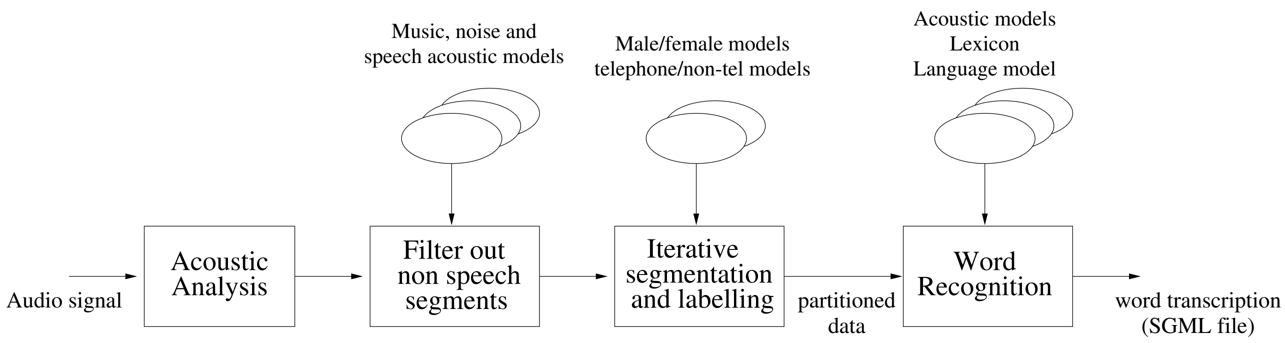

Figure 1: Overview of transcription system for audio stream.

music segments. Gradual transitions between segments occur when there is background music or noise with changing volume, and abrupt changes are common when there is switching between speakers in different locations. The speech is produced by a wide variety of speakers: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, new dialects, non-native speakers, etc. Speech from the same speaker may occur in different parts of the broadcast, and with different background noise conditions. The linguistic style ranges from prepared speech to spontaneous speech. Acoustic and language modeling must accurately account for this varied data.

Two principle types of problems are encountered in automatically transcribing broadcast news data: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. Problems associated with the acoustic signal properties are handled using appropriate signal analyses, by classifying the signal according to segment type and by training acoustic models for the different acoustic conditions. Noise compensation is also needed in order to achieve acceptable performance levels. Most broadcast news transcription systems make use of unsupervised acoustic model adaptation as opposed to noise cancelation, which allow adaptation without an explicit noise model. In order to address the variability observed in the linguistic properties, the differences in speaking styles need to analyzed with regard to lexical items, word and word sequence pronunciations, and frequencies and distribution of hesitations, filler words, and respiration noises. Once such an analysis is carried out. the variability needs to be accounted for in the acoustic and language models [3].

System overview

The LIMSI SDR'99 transcription system shown in Figure 1, is based on the LIMSI 1998 Hub-4E system which achieved an official word error of 13.6% in the Nov'98 ARPA evaluation. Prior to recognition the audio stream is first partitioned. Data partitioning serves to divide the continuous stream of acoustic data into homogenous segments, associating appropriate labels with each segment. The segmentation and labeling procedure [4] first detects and rejects non-speech segments, and then applies an iterative maximum likelihood segmentation/clustering procedure to the speech segments. The result of the partitioning process is a set of speech segments with cluster, gender and telephone-band/wideband labels. The speech recognizer uses continuous density HMMs with Gaussian mixture observation densities for acoustic modeling and 4-gram statistics for language modeling. The states of the context-dependent phone models are tied by means of a decision tree.

Audio Partitioner

The goal of partitioning is to divide the continuous audio stream into homogeneous acoustic segments, to remove non-speech segments and to assign bandwidth and gender labels to each segment. The audio partitioning procedure, introduced for the Nov'97 evaluation [4, 5] and used in the LIMSI Nov'98 Hub-4E system [9], is as follows:

1. First, the non-speech segments are detected (and rejected) using Gaussian mixture models (GMMs). Four GMMs, each with 64 Gaussians serve to detect speech, pure-music and other (background). All test segments labeled as music or silence are removed prior to further processing.

2. An iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. Given the sequence of cepstral vectors, the algorithm tries to maximize an objective function which is a penalized log-likelihood. Alternate Viterbi reestimation and agglomerative clustering yields a sequence of estimates with non decreasing values of the objective function. The algorithm stops when no further merges are possible. The cluster size is constrained to ensure that each cluster corresponds to at least 10s of speech. This procedure is controlled by 3 parameters: the minimum cluster size (10s), the maximum log-likelihood loss for a merge, and the segment boundary penalty.

When no more merges are possible, the segment boundaries are refined (within a 1s interval) using the last set of GMMs and an additional relative energy-based boundary penalty. This is done to locate the segment boundaries at silence portions, so as to avoid cutting words.

3. Speaker-independent GMMs corresponding to wideband and telephone speech (each with 64 Gaussians) are then used to label the segment bandwidths. This is followed by segment-based gender identification, using 2 sets of GMMs with 64 Gaussians (one for each bandwidth). The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

Speech Recognizer

As usual [3, 6, 4, 9], the acoustic feature vector contains 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.5kHz band for telephone data) every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalisation. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficents and the log energy, along with the first and second order derivatives. Each phone model is a tied-state left-to-right CD-HMM with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent but position-dependent. The tied states are obtained by means of a decision tree.

The decoding procedure of the LIMSI Nov'98 Hub-4E system has been changed in order to reduce the computation time required to process the 550 hours of BN data for the SDR'99 evaluation. Word recognition is performed in three passes:

1. *Word graph generation:* An initial hypothesis and word graph are generated using a small bigram-backoff language model and gender-specific sets of position-dependent cross-word triphones.

2. *3-gram decoding with acoustic model adaptation:* Unsupervised acoustic model adaptation is performed for each segment cluster using the MLLR technique [14] using the initial hypotheses. Each segment is decoded with a trigram language model, the adapted acoustic models, and the word graph.

3. *4-gram decoding with acoustic model adaptation:* The final hypotheses are generated using a 4-gram language model with acoustic model adaptation using the hypotheses of pass 2.

Acoustic model training

We used the acoustic models of the LIMSI Nov'98 Hub-4E system. These models were trained on about 150 hours of broadcast data (only the official Hub-4E training data from 1995, 1996, and 1997). The acoustic models are position-dependent triphones with about 11500 tied states (366K Gaussians), obtained using a divisive decision tree based clustering algorithm. Two sets of gender-dependent acoustic models were built using MAP [8] adaptation of SI seed models for each of wideband and telephone band speech. A portion of the Hub-4E training data was also used to build the Gaussian mixture models for partitioning (speech, music and noise models) and for gender and bandwidth identification. About 2 hours of pure music portions taken from the acoustic training data were used to estimate the music GMM.

Language model training

The language models of the LIMSI Nov'98 Hub-4E system were used. The language models are fixed and were obtained by interpolation of backoff n-gram language models trained on different data sets. To build the n-gram LM, four models trained on the following sources were interpolated:

1. BN transcriptions from LDC (years 92-95) and from PSMedia (years 96 and 97 (the period 15/10/96 - 14/11/96 was excluded): 203 M words

2. NAB newspaper texts and AP Wordstream texts prior to September 1995: 202 M words

3. NAB newspaper texts and AP Wordstream texts from July 1996 to August 1997 (the period 15/10/96 - 14/11/96 was excluded) : 141 M words

4. Transcriptions of the acoustic data, BN data (including the 1995 MarketPlace data): 1.6M words

The interpolation coefficients of these four LMs were chosen so as to minimize the perplexity on the Nov'96 and Nov'97 evaluation test sets. A backoff 4-gram LM is then derived from this interpolation by merging the four component LMs [20]. Bigram and trigram LMs were build in a similar manner for use in the first two decoding steps.

All words occuring a minimum of 15 times in the broadcast news texts (63,954 words) or at least twice in the acoustic training data (23,234 mots) were included in the recognition vocabulary, resulting in a 65,122 word list. The lexical coverage is 99.5% on the Hub-4E Nov'97 eval test set and 99.1% on the Hub-4E Nov'96 eval test set.

The BN texts from PSmedia (also used for query expansion in our IR system) were processed using a modified version of a perl script from BBN made available by LDC. The

BN training texts were cleaned in order to be homogeneous with the previous texts. These texts were processed so as to treat some frequent word sequences as compound words, and to treat the most frequent acronyms in the training texts as whole words instead of as sequences of independent letters.

Lexicon

Pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The 65k vocabulary contains 65,122 words including 72,734 phone transcriptions. Frequent inflected forms have been verified to provide more systematic pronunciations. As done in the past, compound words for about 300 frequent word sequences subject to reduced pronunciations were included in the lexicon as well as the representation of frequent acronyms as words.

Transcription results

Table 1 reports the word recognition results on the eval test sets from the last three years. All of our system development was carried out using the Hub-4E eval96 and the SDR'98 data set. For the SDR'98 data set we built a system respecting the rules from last year's SDR evaluation. Since the SDR'98 test data is part of the standard Hub-4E training data, acoustic models were trained on only about 80 hours of acoustic data as opposed to 150h. Similarly language models were trained using only those texts predating the test epoch (Jan'98).

The word transcription error is seen to be on the order of 20% on the broadcast data. The better results for the Hub-4E Nov'97 (h4-97) and Nov'98 (h4-98) test sets are due to prior selection of the test data to include a higher proportion of prepared speech. The word error of the SDR'99 is about 15% higher than the LIMSI Nov'98 Hub-4E system. The difference in performance of the SDR'98 and SDR'99 systems can be attributed to the difference in training data.

	Test set (Word Error)				
	h4-96	*h4-97*	*h4-98*	*sdr98*	*sdr99*
System	*1.8 h*	*3 h*	*3 h*	*100 h*	*10 h*
Hub4'98	19.8	13.9	13.6	-	-
SDR	22.6	16.5	16.0	24.4*	21.5

Table 1: Summary of BN transcription word error rates on the 3 last DARPA evaluation test sets (h4-96, h4-97, h4-98) and the SDR'98 and '99 test sets using the LIMSI HUB4'98 system and the LIMSI SDR'99 system (about 15xRT). *Results on the SDR'98 test set were obtained with a system trained on about half the amount of acoustic data and less LM texts, in accordence with the SDR'98 evaluation condition.

INFORMATION RETRIEVAL

Our SDR'99 IR system has been designed following the Okapi approach [18]. In order for the same IR system to be applied to different text data types (automatic transcriptions, closed captions, additional texts from newspapers or newswires), all of the documents are preprocessed in a homogeneous manner. This preprocessing or tokenization, described below, is the same as what is done to prepare text sources for training the speech recognizer language models [7], and attempts to transform them to be closer to the observed American speaking style. There is no stop list, that is to say no words are discarded during the pre-processing stage. The index terms are obtained after translation using a lexicon of stems. Query expansion is obtained via Blind Relevance Feedback (BRF) using both the SDR'99 audio data collection and a parallel text corpus of broadcast news transcripts.

All development was carried out using exclusively the SDR'98 evaluation data, consisting of about 2800 documents with the associated 23 queries. Two approaches for IR were explored, the first based on the Okapi term weighting function and the second using a Markovian one [11, 15]. Due to the limited amount of development data and our limited experience with IR systems, we chose to submit the Okapi-based system for the evaluation even though comparable results were being obtained with the Markovian approach. Some comparative results for the two approaches are given at the end of this section.

The parameter values were chosen to simultaneously optimize performance on automatic recognizer transcripts and the provided manual reference transcriptions. Better IR performance can be obtained if the parameters for the two transcription types are optimized independently, but this would result in two different IR systems. It is also worth noting that the reference transcripts of the SDR'98 data are detailed manual transcriptions, whereas for the SDR'99 data these are closed captions. The different transcript types made us uncertain as to the reliability of our development work.

Tokenization

The tokenizer transforms the texts to a unified format. The basic operations include translating numbers and sums into words, removing all the punctuation signs, removing case distinctions and detecting acronyms and spelled names such as *K.G.B.* However removing all punctuation markers implies that certain hyphenated words such as *anti-communist*, *non-profit* are rewritten as *anti communist* and *non profit*. While this offers advantages for speech recognition, it can lead to IR errors. To avoid IR problems due to this transformation, the output of the tokenizer (and recognizer) is checked for common prefixes, in order to rewrite the sequence of words *anti communist* as a single word. The prefixes that are handled include *anti*, *co*, *bi*, *counter*. A rewrite lexicon containing compound words formed with these prefixes and a limited number of named entities (such as *Los-Angeles*, *Saint-Tropez*) is used to transform the texts. Similarly all numbers less than one hundred are treated as a single entity (such as *twenty-seven*).

Stemming

In order to reduce the number of lexical items for a given word sense, each word is translated into its stem (as defined in [2, 16]) or, more generally, into a form that is chosen as being representative of its semantic family. The stemming lexicon (using the UMass 'porterized' lexicon) [2] contains about 32000 entries and was constructed using Porter's algorithm on the most frequent words in the collection, and then manually corrected.

The IR term list was limited to 45k entries (after stemming) for implementation reasons. For the SDR'99 audio data collection, this filtering only affected the R1 condition where the least frequent terms were removed.

Baseline search

The score of a document d for a query is given by the Okapi-BM25 formula[17]. It is the sum over all the terms t in the query of the following weights:

$$cw_{t,d} = qtf_t \frac{(K+1) \times tf_{t,d}}{K \times (1 - b + b \times L_d) + tf_{t,d}} \log \frac{N}{N_t} \qquad (1)$$

where $tf_{t,d}$ is the number of occurrences of term t in document d (i.e. term frequency in document), N_t is the number of documents containing term t at least once, N is the total number of documents in the collection, L_d is the length of document d divided by the average length of the documents in the collection, and qtf_t the number of occurrences of term t in the query.

The parameter values of the Okapi formula were chosen in an attempt to maximize the average precision on the SDR'98 data set. The resulting values were thus a compromise between the optimal configuration for the R1 and S1 conditions, in order to be able to use the same values for both conditions. The S1 transcripts were obtained with a speech recognizer trained on 75 hours of acoustic data and language model training texts predating the test period. The recognition word error rate on this data (using the NIST SDR'98 scoring procedure) was 24.4% (cf. Table 1). The parameters were fixed for all the evaluation conditions at: b=0.86; and K=1.2 for the baseline run without query expansion, and K=1.1 with query expansion.

Query expansion

The text of the query may or may not include the index terms associated with relevant documents. One way to cope with this problem is to use query expansion based on terms present in retrieved documents on the same (Blind Relevance Feedback) or other (Parallel Blind Relevance Feedback) data collections [19]. We have experimented with both approaches, and our submitted system incorporated both BRF and PBRF using 6 months of commercially available broadcast news transcripts for the period of June-December 1997 [1]. This corpus contains 50 000 stories and 49.5 M words.

For a given query, the terms found in the top B documents from the baseline search are ranked by their offer weight (ow_t), and the top T terms are added to the query. As proposed in [18] the following formula for ow_t was used:

$$ow_t = r_t \log \frac{(r_t + 0.5)(N - N_t - B + r_t + 0.5)}{(N_t - r_t + 0.5)(B - r_t + 0.5)} \qquad (2)$$

where r_t is the number of documents (among the B documents) containing the term t.

Since only the T terms with best offer weights are kept, we filtered the terms using a stop list of 144 common words, in order to increase the likehood that these terms are relevant.

data	*base*	*brf*	*pbrf*	*brf+pbrf*
R1	0.4689	0.5597	0.5609	0.5803
S1	0.4594	0.5329	0.5442	0.5636

Table 2: Development IR results on the SDR'98 data set (b=0.86, K=1.1, B=15, T=5) for the baseline system, and with 3 configurations for query expansion.

Four experimental configurations are reported in Table 2 for the SDR'98 development data: baseline search (*base*), query expansion using BRF (*brf*), query expansion with parallel BRF (*pbrf*) and query expansion using both BRF and PBRF (*brf+pbrf*). For BRF and PBRF, the terms are added to the query with a weight of 1. For BRF+PBRF, the terms from each source are added with a weight of 0.5. The parameter values used for these experiments are the result of our development work. We felt that it was safest to add only a few terms, assuming that only a small number of documents were relevant. Therefore the development experiments compared performance for relatively small values of B and T, with the best performance being obtained with $B = 15$ and $T = 5$. The results reported in Table 2 clearly demonstrate the interest of using both BRF and PBRF expansion techniques with consistent and comparable improvements over the baseline for the two conditions (R1 and S1). As has been previously reported by other sites, there is only a slight performance degradation in going from the R1 condition to the S1 condition, even with a transcription word error of 24%.

Evaluation Results

The parameter setting optimized on the SDR'98 data set (cf. Table 2) were used for all our submissions on the SDR'99 data set. Table 3 summarizes the results of the LIMSI IR system for the R1, S1, and cross-recognizer conditions. In addition to the official numbers obtained with query expansion using both BRF and PBRF, the results for the 3 other configurations (no query expansion, query expansion with BRF and query expansion with PBRF) are also provided.

data	*base*	*brf*	*pbrf*	*brf+pbrf*
R1	0.4711	0.5330	0.5126	0.5411
S1	0.4327	0.4978	0.4848	0.5072
B1	0.4180	0.4787	0.4702	0.4828
B2	0.4212	0.4786	0.4748	0.4839
HTK	0.4436	0.5163	0.4933	0.5176
ATT	0.4178	0.4956	0.4621	0.4925
SHEF	0.4041	0.4659	0.4593	0.4787
CMU	0.2732	0.2980	0.3368	0.3234

Table 3: LIMSI official IR results on the SDR'99 data set (b=0.86, K=1.1, B=15, T=5).

The highest average precision is obtained on the manual transcriptions (R1: 0.5411), but as already observed on our development results the performance degradation using speech recognizer outputs is fairly modest (2% and 3% for the HTK and LIMSI automatic transcriptions). Comparing Tables 2 and 3, it can be observed that the gain using PBRF for query expansion is smaller on the SDR'99 data set than it was on the SDR'98 data set. This is may be linked to the choice of the epoch for the PBRF corpus or to a suboptimal tuning of the BRF parameters.

ADDITIONAL RESULTS

In this section some post-evaluation experiments with the Okapi-based system are reported. We also report here some of the development experiments comparing the Okapi and Markovian approaches.

Adjusting System Parameters

Having no experience with IR system tuning before this evaluation, we found it rather difficult to properly set the Okapi parameters (K and b) and the query expansion parameters (B and T), so as to maximize the average precision for both the R1 and S1 conditions on the SDR'98 test set with the associated 23 queries.

Extensive experiments were carried out to investigate the IR performance for a range of parameter values. Figures 2 through 4 show the effect of the Okapi parameters (b and K) on the average precision for SDR'98-R1, SDR'99-R1 and SDR'99-S1 respectively, using a baseline system without query expansion. The iso-data lines of the resulting surfaces are shown, along with their projections on the base plane which highlights the location of the extrema.

Figures 5 through 7 show the effect of the BRF parameters (B and T) on the average precision for SDR'98-R1, SDR'99-R1 and SDR'99-S1 respectively, using the system with query expansion based on both BRF and PBRF.

It is clear from these plots that the best parameter settings for the SDR'99 data set cannot be easily predicted from the SDR'98 results. However it was clearly possible to choose better BRF parameter values than those resulting from our development work. In particular too few terms are kept (i.e. the T value was really underestimated). New results using T=10 (which corresponds to the best results on the SDR'98-S1 data) are given in Table 4 (label *cw* for the Okapi term weighting).

Markovian term weigthing

As a natural extension of our work on speech recognition relying on Markovian assumptions for both acoustic and language modeling, we investigated a term weighting function based on a simple query/document model in place of the Okapi formula. A comparable approach has been previously employed with success [11, 15]. Assuming a unigram model, the following term weighting is used:

$$mw_{t,d} = qtf_t \times \log(\alpha \Pr(t|d) + (1-\alpha)\Pr(t)). \quad (3)$$

Table 4 gives the results for both Okapi (*cw*) and Markovian (*mw*) term weightings on the SDR'99 data set with the following parameter settings: b=0.86, K=1.1, B=15, T=10, α=0.5. In both cases query expansion relies on the term offer weight defined above. It can be seen that very comparable results can be achieved using the two term weighting schemes.

data	*meth.*	*base*	*brf*	*pbrf*	*brf+pbrf*
98-R1	*cw*	0.4689	0.5648	0.5591	0.5786
	mw	0.4695	0.5936	0.5574	0.5889
98-S1	*cw*	0.4594	0.5118	0.5621	0.5761
	mw	0.4558	0.5121	0.5884	0.5745
99-R1	*cw*	0.4711	0.5318	0.5147	0.5487
	mw	0.4691	0.5354	0.5098	0.5430
99-S1	*cw*	0.4327	0.5239	0.4919	0.5350
	mw	0.4412	0.5302	0.4943	0.5398

Table 4: Comparison of IR results on the SDR'98 and SDR'99 data sets using both Okapi and Markovian term weightings (b=0.86, K=1.1, B=15, T=10, α=0.5). R1: reference transcript. S1: automatic speech transcription.

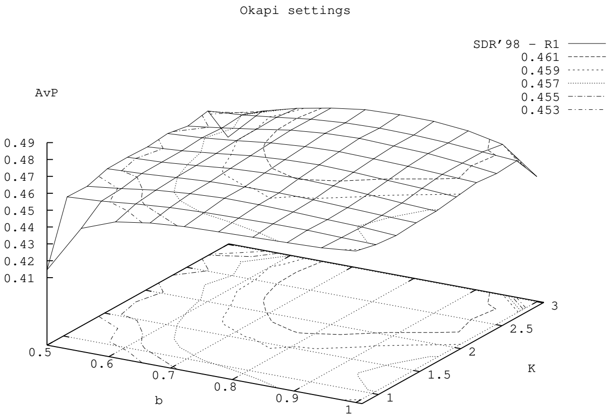

Figure 2: Plot of average precision vs Okapi parameters b and K for the baseline system (no query expansion), SDR'98 - R1. (Best AveP is 0.4836 for b=0.80 and K =2.5)

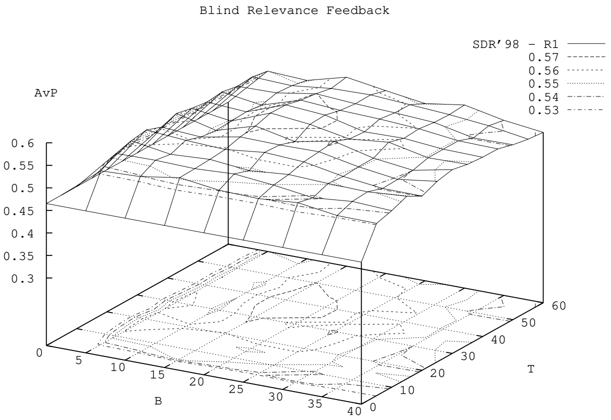

Figure 5: Plot of average precision vs BRF parameters B and T for BRF+PBRF query expansion, SDR98 - R1. (Best AveP is 0.5835 for B=15 and T=35).

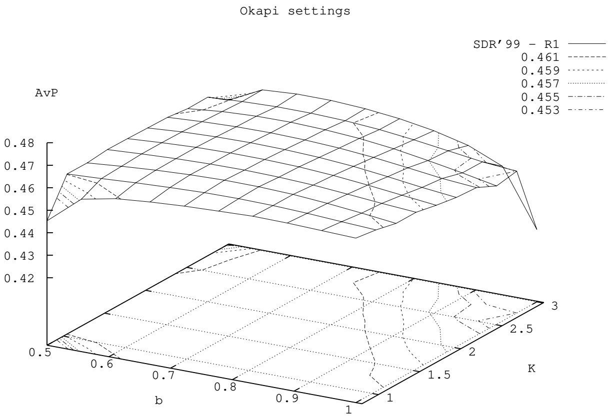

Figure 3: Plot of average precision vs Okapi parameters b and K for the baseline system (no query expansion), SDR'99 - R1. (Best AveP is 0.4736 for b=0.75 and K =1.3).

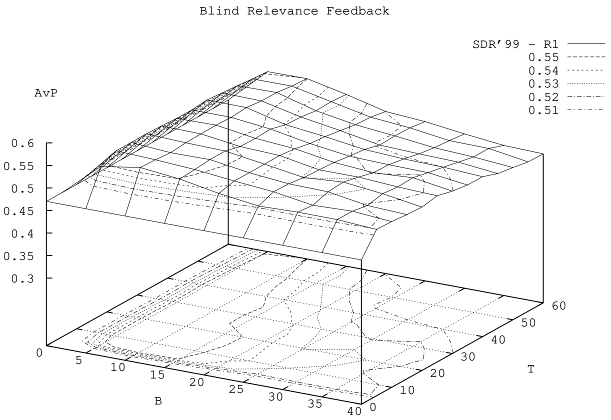

Figure 6: Plot of average precision vs BRF parameters B and T for BRF+PBRF query expansion, SDR99 - R1. (Best AveP is 0.5615 for B=5 and T=40).

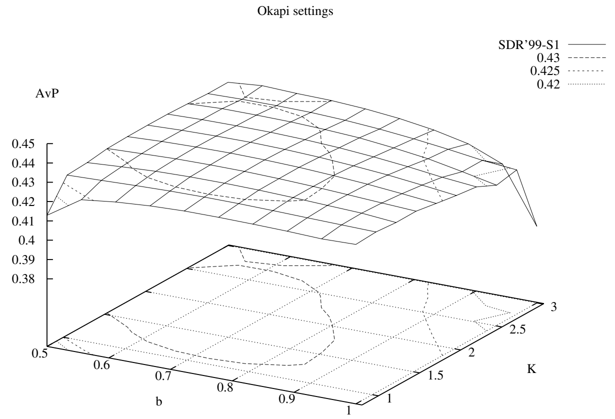

Figure 4: Plot of average precision vs Okapi parameters b and K for the baseline system (no query expansion), SDR'99 - S1. (Best AveP is 0.4401 for b=0.65 and K =2.1).

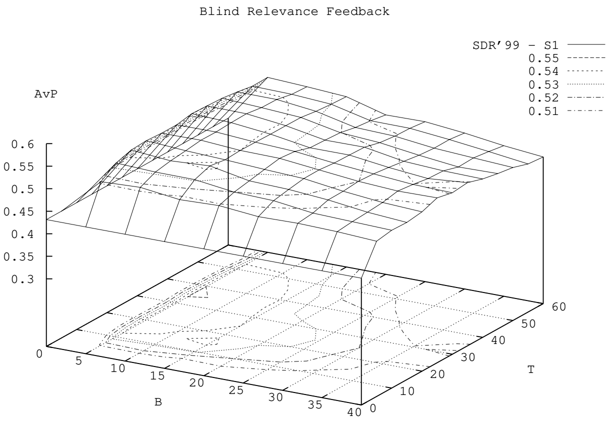

Figure 7: Plot of average precision vs BRF parameters B and T for BRF+PBRF query expansion SDR99 - S1. (Best AveP is 0.5515 for B=5 and T=35).

SUMMARY & DISCUSSION

In this paper we have presented our complete SDR'99 system, and highlighted our development work. This system was built by combining an adapted version of the LIMSI 1998 Hub-4E transcription system for speech recognition with an IR system based on the Okapi term weighting function. The transcription system achieved a word error of 21.5% measured on a 10h subset of the SDR'99 data set. Using the parameter settings optimized on the SDR'98 data set, average precision of 0.5636 and 0.5072 respectively were obtained on the SDR'98 and SDR'99 data sets using the transcriptions produced by the LIMSI recognizer. These values are quite close the the average precisions obtained on manual transcripts, indicating that the transcription quality is not the limiting factor on IR performance. Our post-evaluation experiments indicate that (unfortunately) the evaluation settings for the BRF were suboptimal.

ACKNOWLEDGMENTS

This work has been partially financed by the European Commission and the French Ministry of Defense. The authors gratefully acknowledge the participation of Michèle Jardino, Remi Lejeune and Patrick Paroubek to this work.

REFERENCES

[1] http://www.thomson.com/psmedia/bnews.html

[2] ftp://ciir-ftp.cs.umass.edu/pub/stemming/

[3] J.L. Gauvain, G. Adda, L. Lamel and M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *ARPA Speech Recognition Workshop*, Chantilly, VA, pp. 56-63, February 1997.

[4] J.L. Gauvain, L. Lamel and G. Adda, "The LIMSI 1997 Hub-4E Transcription System," *DARPA Broadcast News Transcription & Understanding Workshop*, Landsdowne, VA, pp. 75-79, February 1998.

[5] J.L. Gauvain, L. Lamel and G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. ICSLP'98*, **5**, pp. 1335-1338, Sydney, December 1998.

[6] J.L. Gauvain, L. Lamel, G. Adda and M. Adda-Decker, "Transcription of Broadcast News", *Proc. ESCA EuroSpeech'97*, Rhodes, pp. 907-910, September 1997.

[7] J.L. Gauvain, L. Lamel, M. Adda-Decker, "The LIMSI Nov93 WSJ System." *Proc. ARPA Spoken Language Technology Workshop*, Princeton, NJ, pp. 125-128, March 1994.

[8] J.L. Gauvain and C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, **2**(2), pp. 291-298, April 1994.

[9] J.L. Gauvain, L. Lamel, G. Adda and M. Jardino, "The LIMSI 1998 HUB-4E Transcription System," *Proc. DARPA Broadcast News Workshop*, Herdnon, VA, pp. 99-104, February 1999.

[10] A.G. Hauptmann, M. Witbrock and M. Christel, "News-on-Demand-'An Application of Informedia Technology'," *Digital Libraries Magazine*, September 1995.

[11] D. Hiemstra and K. Wessel, "Twenty-One at TREC-7: Ad-hoc and Cross-language track," *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, NIST, Gaithersburg, MD, November 1998.

[12] S. E. Johnson, P. Jourlin, G. L. Moore, K. Spärk Jones and P. C. Woodland, "Spoken document retrieval for TREC-7 at Cambridge University", *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, NIST, Gaithersburg, MD, November 1998.

[13] F. de Jong, J.L. Gauvain, J. deb Hartog and K. Netter, "OLIVE: Speech Based Video Retrieval," *Proc. CBMI'99*, Toulouse, October 1999.

[14] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2), pp. 171-185, 1995.

[15] D. Miller, T. Leek and R. Schwartz, "Using Hidden Markov Models for Information Retrieval," *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, NIST, Gaithersburg, MD, November 1998.

[16] M. F. Porter, "An algorithm for suffix stripping", *Program*, **14**, pp. 130–137, 1980.

[17] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu and M. Gatford, "Okapi at TREC-3", *NIST Special Publication 500-226: Overview of the Third Text REtrieval Conference (TREC-3)*, NIST, Gaithersburg, MD, November 1994.

[18] K. Spärk Jones, S. Walker and S. E. Robertson, "A probabilistic model of information retrieval: development and status," *a Technical Report of the Computer Laboratory, University of Cambridge, U.K.*, 1998.

[19] S. Walker and R. de Vere, "Improving subject retrieval in online catalogues: 2. Relevance feedback and query expansion," *British Library Research Paper 72*, British Library, London, U.K., 1990.

[20] P.C. Woodland, T. Neieler and E. Whittaker, "Language Modeling in the HTK Hub5 LVCSR," presented at the 1998 Hub5E Workshop, September 1998.