TRANSCRIPTION AND INDEXATION OF BROADCAST DATA

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

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

ABSTRACT

In this paper we report on recent research on transcribing and indexing broadcast news data for information retrieval purposes. The system described here combines an adapted version of the LIMSI 1998 Hub-4E transcription system for speech recognition with text-based IR methods. Experimental results are reported in terms of recognition word error rate and mean average precision for both the TREC SDR98 (100h) and SDR99 (600h) data sets. With query expansion using commercial transcripts, comparable mean average precisions are obtained on manual reference transcriptions and automatic transcriptions with a word error rate of 21.5% measured on a 10 hour data subset.

INTRODUCTION

With the increasing number of different media sources for information dissemination, there is a rapidly growing need for fast automatic processing of the audio data stream. 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. Automation of some of these activities can allow more information sources to be processed and significantly reduce processing costs while eliminating tedious work. Some existing applications that could greatly benefit from new technology are 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 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 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.

In this paper we describe the LIMSI spoken document indexing and retrieval system used in the TREC-8 SDR evaluation. This system combines a state-of-the-art speech recognizer [9] with a text-based IR system. Our development work made use of the TREC-7 SDR data set (100h) and the associated set of 23 queries. The SDR99 data set was substantially more challenging than the SDR98 data, in that the audio data was increased to about 600 hours of broadcasts, which has strong implications on the transcription process.

BROADCAST NEWS TRANSCRIPTION

Two principle types of problems are encountered in automatically transcribing broadcast news (BN) data: those related 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 BN transcription systems make use of unsupervised acoustic model adaptation as opposed to noise cancelation, which allow adaptation without an explicit noise model. The observed linguistic variability is explicitly accounted for in the acoustic and language models [4].

The transcription system shown in Figure 1, is based on the LIMSI 1998 Hub4-E system which achieved an official word error of 13.6% in the Nov98 ARPA evaluation. Prior to recognition the audio stream is partitioned, which serves to divide the continuous stream of acoustic data into homogeneous segments, associating appropriate labels with each segment. The segmentation and labeling procedure [5, 6] 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 of context dependent phones and 4-gram statistics for language modeling. The states of the context-dependent phone models are tied by means of a decision tree.

The decoding procedure used in the SDR99 evaluation was slightly modified from that of the LIMSI Nov98 Hub4E system, in order to reduce the computation time required to process the 600 hours of BN data. Word recognition is carried out in three steps: initial hypothesis generation with a

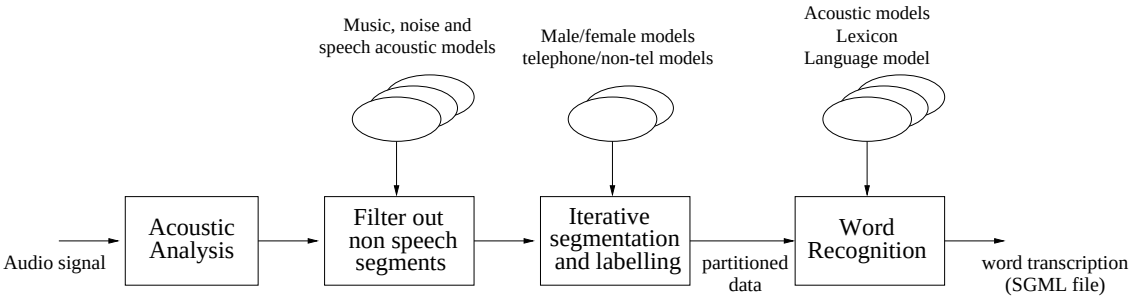

Figure 1: Overview of transcription system for audio stream.

bigram LM; word graph generation with a trigram LM; final hypothesis generation using a 4-gram LM. The initial hypothesis are used in unsupervised cluster-based acoustic model adaptation (MLLR technique [11]) prior to word graph generation. This step, which aims to reduce the mismatch between the models and the data, is crucial for generating accurate word graphs.

The acoustic models and language models used in the last decoding step are those of the LIMSI Nov98 Hub4E system. The acoustic models, trained on about 150 hours of broadcast data, are position-dependent triphones with about 11500 tied states (350K Gaussians). The state-tying is obtained via 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 Hub4 training data was also used to build the Gaussian mixture models for partitioning (speech, music and noise models) and for gender and bandwidth identification.

Fixed language models were obtained by interpolation of backoff n-gram language models trained on 3 different data sets: 203 M words of BN transcriptions; 343 M words of NAB newspaper texts and AP Wordstream texts; 1.6 M words corresponding to the transcriptions of the acoustic acoustic training data. The interpolation coefficients of these LMs were chosen so as to minimize the perplexity on the Nov96 and Nov97 evaluation test sets. The recognition word list contains 65122 words, and has a lexical coverage of 99.5% and 99.1% on the Hub4-Nov97 and Nov96 eval test sets. The 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. 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.

Table 1 gives the word recognition results five test sets. System development was carried out using the Hub4-eval96 and the SDR98 data set. For the SDR98 data set a system respecting the rules from the SDR98 evaluation was built.[1] The word transcription error is seen to be on the order of 20% on the broadcast data. The better results for the h4-97 and 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 SDR99 system running at about 15xRT is about 15% higher than the LIMSI Nov98 Hub4 system. The difference in performance of the SDR98 and SDR99 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 SDR98 and '99 test sets using the LIMSI HUB4'98 system and the LIMSI SDR99 system (about 15xRT). *Results on the SDR98 test set were obtained with a system trained on about half the amount of acoustic data and less LM texts, in accordance with the SDR98 evaluation condition.

Even though it is usually assumed that processing time is not a major issue since computer processing power increase continuously, it is also known that the amount of data appearing on information channels is increasing at a close rate. Therefore processing time is an important factor in making a speech transcription system viable for audio and video indexing. Processing time constraints evidently significantly change the way we select our models. For each operating point, the right balance between model complexity and search pruning level must be found. Figure 2 plots the word error rate as a function of processing time for 3 sets of acoustic models, which taken together minimize the word error rate over a wide range of processing times (from 0.5xRT to 25xRT). These results on a representative portion of the Hub4-98 data set are obtained with a 3-gram language model and without acoustic model adaptation using a Compaq XP1000 machine. The 350k model set (350k Gaussians,

[1] Since the SDR98 test data is part of the standard Hub4 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.

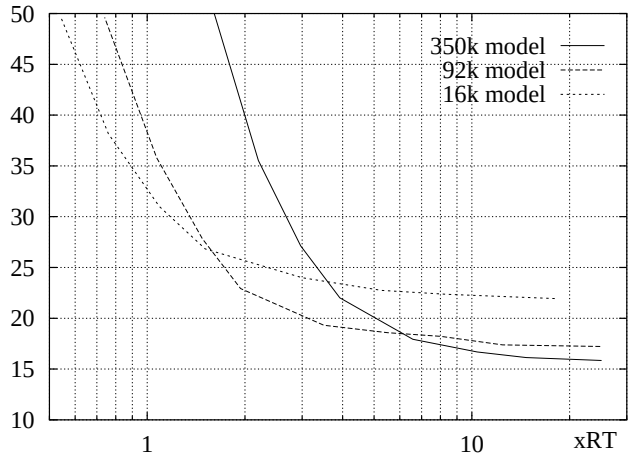

Figure 2: Word error vs processing time for three model sets with 350k, 92k and 16k Gaussians.

11k tied states, 30k phone contexts) is currently our best model set, and provides the best performance/speed ratio for processing times over 7xRT. The 92k model set (92k Gaussians, 6k tied states, 5k phone contexts) performs better in the range 2xRT to 6xRT, whereas a much smaller model set (16k Gaussians) is needed to reach or go under real-time.

INFORMATION RETRIEVAL

Two approaches for IR were explored, the first based on the Okapi term weighting function [15] and the second using a Markovian one [10, 12]. All development was carried out using the SDR98 test data (100h), consisting of about 2800 documents with the associated 23 queries. The SDR99 test data (600h) consists of 21750 documents with an associated set of 50 queries. It should be noted that the reference transcripts of the SDR98 data are detailed manual transcriptions, whereas for the SDR99 data these are closed captions.

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, is the same as the text source preparation for training the speech recognizer language models [7], and attempts to transform them to be closer to the observed American speaking style. The basic operations include translating numbers and sums into words, removing all the punctuation symbols, removing case distinctions and detecting acronyms and spelled names. However removing all punctuations 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 a sequence of words such as *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*) is used to transform the texts. Similarly all numbers less than one hundred are treated as a single entity (such as *twenty-seven*).

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, 13]) 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 score of a document d for a query is given by the Okapi-BM25 formula[14]. It is the sum over all the terms t in the query of:

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

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. As a natural extension of our work on speech recognition, we also investigated a Markovian 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 [10, 12]. Assuming a unigram model, the following term weighting was used:

$$mw_{t,d} = qtf_t * \log(\alpha \Pr(t|d) + (1 - \alpha) \Pr(t)). \quad (2)$$

The values of K, b (for $cw_{t,d}$) and α (for $mw_{t,d}$) were chosen in an attempt to maximize the average precision on the SDR98 data set. The resulting values were 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 parameters were fixed for all the evaluation conditions as: b=0.86, K=1.2 and α=0.3 with no query expansion; and K=1.1 and α=0.5 with 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 [16]. We combined the two approches in our system. For PBRF we used 6 months of commercially available broadcast news transcripts for jun-dec 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 [15], and the top T terms are added to the query. Since only the T terms with best offer weights are kept, the terms are filtered using a stop

list of 144 common words, in order to increase the likelihood that the resulting terms are relevant.

Table 2 gives the results for both *cw* and *mw* term weightings for the SDR98 and SDR99 data set. Four experimental configurations are reported: 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 results clearly demonstrate the interest of using both BRF and PBRF expansion techniques, as consistent improvements are obtained over the baseline system for the two conditions (R1 and S1). PBRF is particularly effective for the S1 condition (the recognizer transcripts) whereas BRF is much more efficient for the R1 condition (the manual transcripts). Overall comparable results are obtained for both conditions even though the recognizer transcripts have a 21.5% word error rate. The LIMSI official results for the SDR99 evaluation which were obtained using the Okapi term weighting and a (unfortunately) quite suboptimal query expansion tuning (T=5 instead of T=10) are 0.5411 (R1) and 0.5072 (S1). In the same conditions the results on the SDR98 data are 0.5803 (R1) and 0.5636 (S1).

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 2: Comparison of IR results on the SDR98 and SDR99 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.

SUMMARY & DISCUSSION

In this paper we have presented our recent research in transcribing and indexing television and radio broadcasts for information retrieval. These are necessary processing steps to enable automated processing of the vast amounts of audio and video data produced on a daily basis.

On unrestricted broadcast news shows the word error rates is about 20%. A complete indexing system has been built by applying standard text IR techniques on the output of our BN speech recognizer. Average precisions of 0.57 and 0.54 respectively were obtained on the SDR98 and SDR99 test sets using the transcriptions produced by the LIMSI recognizer. These values are quite close to the average precisions obtained on manual transcripts (0.58 and 0.55), indicating that the transcription quality is not the limiting factor on IR performance for current IR techniques.

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] F. de Jong, J.L. Gauvain, J. den Hartog, K. Netter, "OLIVE: Speech Based Video Retrieval," *Proc. CBMI'99*, Oct. 1999.

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

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

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

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

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

[9] J.L. Gauvain, L. Lamel, G. Adda, M. Jardino, "The LIMSI 1998 HUB-4E Transcription System," *DARPA Broadcast News Workshop*, Feb. 1999.

[10] D. Hiemstra, K. Wessel, "Twenty-One at TREC-7: Ad-hoc and Cross-language track," *Proc. TREC-7*, 1998.

[11] C.J. Leggetter, 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.

[12] D. Miller, T. Leek, R. Schwartz, "Using Hidden Markov Models for Information Retrieval", *Proc. TREC-7*, 1998.

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

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

[15] K. Spärk Jones, S. Walker, 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.

[16] S. Walker, 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.