THE THISL SDR SYSTEM AT TREC-9

Steve Renals and Dave Abberley[*]

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
email: s.renals@dcs.shef.ac.uk; dca@softsound.com

ABSTRACT

This paper describes our participation in the TREC-9 Spoken Document Retrieval (SDR) track. The THISL SDR system consists of a realtime version of a hybrid connectionist/HMM large vocabulary speech recognition system and a probabilistic text retrieval system. This paper describes the configuration of the speech recognition and text retrieval systems, including segmentation and query expansion. We report our results for development tests using the TREC-8 queries, and for the TREC-9 evaluation.

1. INTRODUCTION

The TREC-9 Spoken Document Retrieval (SDR) track followed on from the TREC-8 track, using the same audio collection: 902 shows (502 hours) of US broadcast news material covering the period February–June 1998. The collection contained 21 754 individual news items, totalling 389 hours of news material. The basic task was to retrieve the set of stories relevant to each of 50 topics.

There were three principal dimensions of variation to be investigated in this year's evaluation:

Story Boundaries The main task assumed unknown story boundaries. Each episode was treated as a continuous audio stream and it was the task of the SDR system to find the location (time) of the relevant news stories. The known story boundary condition, in which stories are segmented manually and irrelevant material such as adverts are removed, was used as a contrast.

Query Length Previous SDR tracks used *short* (sentence length) queries. In TREC-9, a *terse* query was also provided for each topic, which typically contained 2–3 words, to reflect queries submitted to web search engines.

Cross-Recognizer Effects In addition to the baseline recognizer and reference (subtitle) transcripts, we also used the transcripts produced by other evaluation participants (Cambridge University and LIMSI). This illuminates the effect of speech recognizer word error rate on SDR system performance.

Much of the paper describes experiments on the development test set, using the TREC-8 SDR queries. Since that evaluation included short queries only, we generated terse queries ourselves: our TREC-8 terse queries are thus not comparable with similar queries that have been generated by other groups. In our development experiments we took average precision on the short queries as our primary metric.

The paper is structured as follows. Section 2 describes the speech recognition component of the system, which is based on the ABBOT hybrid connectionist/HMM large vocabulary speech recognizer, running in real-time mode. Section 3 outlines the text retrieval system that we have used, together with a discussion of the algorithms employed for query expansion and segmentation. Section 4 presents the results we obtained on the TREC-9 SDR track and further discussion of some of the issues raised ends the paper, along with some conclusions.

2. SPEECH RECOGNITION

2.1. Abbot

ABBOT (Robinson et al., 1996) is a connectionist/HMM system (Bourlard and Morgan, 1994) which estimates posterior phone probabilities given the acoustic data at each frame. This discriminative approach differs from that used by most recognizers in that it does not include a generative model of the data. That is, the joint probability of the acoustics and word sequence is not estimated; instead an estimate of the posterior probability of the word sequence given the acoustic data is provided. This may be interpreted as a probabilistic finite state acceptor model (Hennebert et al., 1997).

A recurrent network (RNN) trained as a phone classifier (Robinson, 1994) is used as the principal posterior probability estimator. This approach is attractive since fewer parameters are required for the connectionist model (the posterior distribution is typically less complex than the likelihood) and connectionist architectures make very few assumptions on the form of the distribution. Additionally,

[*]Now at: SoftSound, Cambridge CB4 0WS, UK

this approach enables the use of posterior probability based pruning in decoding (Renals and Hochberg, 1999).

We produced two sets of transcriptions of the audio data for the TREC-9 evaluation, referred to as S1 and S2. Both systems used the same language model (LM) and search components. The S1 system was configured to run in realtime, while the S2 system used a richer acoustic model.

2.2. S1 Acoustic Model

The S1 acoustic model comprised two RNNs each of which estimated 54 context-independent posterior phone probabilities for each frame of acoustic data. Both networks were trained using a sequence of 12th order perceptual linear prediction features (Hermansky, 1990) (plus log energy). One network estimated the phone probabilities for the current frame conditioned on the past sequence of acoustic features. The second network was trained using a frame sequence that was reversed in time, and thus estimated the phone probabilities conditioned on the future. The two estimated probability streams were averaged in the log domain to produce a final set of probability estimates. The models were trained using the 104 hours of Broadcast News training data released in 1997 (the first half of the complete broadcast news training set).

2.3. S2 Acoustic Model

The acoustic model for the S2 system was obtained by log domain merging of the probability estimates produced by the RNNs used in the S1 system with those produced by an acoustic model using modulation-filtered spectrogram features. This is essentially the system used by the SPRACH group in the 1998 broadcast news evaluation (Cook et al., 1999; Robinson et al., 2001).

The modulation-filtered spectrogram (MSG) was developed by Kingsbury et al. (1998) as a feature representation that is robust to the signal variations caused by reverberation and noise. The robustness is obtained by emphasising modulation in the speech spectral structure occurring at rates of 16Hz or less (as measured with a critical-band-like resolution) and adapting to slowly-varying components of the speech signal (a form of automatic gain control). MSG feature processing involves first calculating an auditory-like spectrum, then filtering the amplitude in each frequency band by two parallel banks of filters, one low-pass below 16Hz, and the second bandpass between 2Hz and 16Hz. Each channel is then passed, in series, through two feedback Automatic Gain Control units with time constants of 160ms and 640ms. The resulting spectra are used as features; orthogonalization (e.g. via the discrete cosine transform) provides no benefit for these features in our experience with connectionist models. However, we do increase the robustness of the system to environmental conditions by normalizing the statistics of every feature channel to zero mean and unit variance over each segment, or over entire recordings if no segmentation is performed.

The MSG acoustic model used an MLP containing 8000 hidden units trained on the full 200 hours broadcast news training set, with the training data downsampled to 4 kHz bandwidth. Experiment has previously indicated that although the word error rate of the bandlimited MSG-based system is higher than that of the PLP-based S1 system, the errors are different and the overall performance may be improved by merging the two.

2.4. Language Modelling and Search

The same backed-off trigram LM was used by both the S1 and S2 systems (Robinson et al., 2001). Approximately 450 million words of text data were used to generate the model, comprising: the Broadcast News acoustic training transcripts (1.6M words); the 1996 Broadcast News LM text data (150M words); and the 1998 North American News text data (LA Times/Washington Post (12M words), Associated Press World Service (100M words), NY Times (190M words)). The models were trained using version 2 of the CMU-Cambridge SLM Toolkit (Clarkson and Rosenfeld, 1997) using Witten-Bell discounting. We used a lexicon containing 65 432 words, including every word in the broadcast news training data. The dictionary was constructed using phone decision tree smoothed acoustic alignments. The LM and lexicon were constructed from material pre-dating the acoustic data and were fixed throughout the evaluation.

For both systems we used a large vocabulary stack decoder CHRONOS (Robinson and Christie, 1998).

2.5. Results

Table 1 gives the word error rate estimates obtained using the S1 and S2 systems. These estimates were obtained using a 10 hour sample of the test corpus defined by NIST.

3. TEXT RETRIEVAL

3.1. Basic Text Retrieval System

We used a standard probabilistic system using a short stop list of 132 words (with an additional stop list of 78 words

System	Sub.	Del.	Ins.	WER
S1	22.0	6.1	3.9	32.0
S2	20.0	5.4	3.8	29.2

Table 1: Word error rates (WER) for the S1 and S2 speech recognition systems, estimated using a 10 hour subset of the corpus.

when processing a query), the Porter stemming algorithm and term weighting similar to that used in the Okapi system. Specifically, following Robertson and Spärck Jones (1997), we used the following function $CW(t,d)$ to compute the combined relevance weight between a term t and a document d:

$$CW(t,d) = \frac{CFW(t) * TF(t,d) * (K+1)}{K((1-b)+b*NDL(d)) + TF(t,d)}. \quad (1)$$

$TF(t,d)$ is the frequency of term t in document d, $NDL(d)$ is the normalized document length of d:

$$NDL(d) = \frac{DL(d)}{\overline{DL}}, \quad (2)$$

where $DL(d)$ is the length of document d (ie the number of unstopped terms in d). $CFW(t)$ is the collection frequency weight of term t and is defined as:

$$CFW(t) = \log\left(\frac{N}{N(t)}\right), \quad (3)$$

where N is the number of documents in the collection and $N(t)$ is the number of documents containing term t. The parameters b and K in (1) control the effect of document length and term frequency.

3.2. Segmentation

Since the core task of the SDR track involves the situation where story boundaries are unknown, segmentation of the audio stream assumes some importance. Unlike some other broadcast news speech recognition systems (eg, Odell et al. (1999)), we do not perform any acoustic segmentation in the recognition phase (the audio stream is decoded directly); anyway, there is no good correlation between segments obtained purely from low-level audio features and story segments required for information retrieval. Although other approaches, such as those investigated in the TDT programme, are of some interest, we have no evidence of their suitability for spoken document retrieval.

Thus we have retained the simple approach used last year, based on overlapping rectangular windows of the audio stream[1]. At query time, those relevant segments which overlap are merged. Previously for this type of automatic segmentation, we have used (1) with $b = 0$, since each segment is the same length. However with short segments (30s) this can result in a large number of identical scores, with no good way of breaking the tie. Since the segments do not contain identical numbers of terms — and since we need a tie-breaker — we have used a small non-zero value for b (typically 0.1).

[1] Also used successfully with an SDR system for a 3 000 hour archive of BBC news broadcasts.

The procedure for merging was as follows. The ranked list of (presumed) relevant segments was processed in best first order. Segments that could be potentially merged with the current segment must: (1) come from the same episode; (2) overlap in time; and (3) be within a rank Δ^r of the current segment. If these conditions are met then the two segments are merged. If the scores of the two segments are within a factor m of each other, and the ranks are within $\Delta^f \leq \Delta^r$ then we assume an *equal* merge; otherwise the higher ranked segment *dominates* the other. In an equal merge, the score of the merged segment is set to be the maximum score of the two segments increased by a factor s, and the reference time is set to be the mid-point of the segment. For a dominating merge, the score and reference time of the merged segment are set to be the same as for the highest scoring component segment. The merging process is iterated until convergence, with parameters Δ^r and Δ^f halved on each iteration.

The overall segmentation procedure is summarized as follows:

1. Entire news episode decoded into a stream of text
2. For indexing, the text stream is split into documents using a fixed length rectangular window with a frame length of t_ℓ and a frame shift of t_s
3. At retrieval time a list of $5R$ segments are retrieved, and the above merging process is carried out. We conducted a number of development experiments to obtain values for the segment merging parameters: $\Delta^r = 1600$, $\Delta^f = 200$, $m = 0.95$ and $s = 1.005$
4. The top R merged segments are then returned.

Previously we have used $t_\ell = 30s$ and $t_s = 15s$. We conducted a variety of experiments looking at the effect of varying the frame shift, with a constant frame length ($t_\ell = 30s$) — our hypothesis was that the possible cost of redundant segments of decreasing the frame shift might be offset by the segment merging algorithm. The results (table 2) indicated a frame shift of $t_s = 9s$ to be a good tradeoff between average precision and index size.

This merging scheme was developed using the TREC-8 development set. Given the several heuristically set parameters, there is a distinct possibility of over-tuning. An alternative approach (Johnson et al., 2000) merged all segments originating within 4 or 5 minutes of each other from a single episode. While this approach may well prove to be robust in actual usage, we believed it may be counter-productive for the SDR track since different relevant documents are sometimes located within less than 4 minutes of each other (owing to adverts, etc.)

	Short Queries		Terse Queries	
Shift/s	AveP	R-P	AveP	R-P
6	0.526	0.524	0.486	0.490
9	0.526	0.518	0.477	0.485
12	0.518	0.508	0.487	0.476
15	0.510	0.507	0.470	0.477
20	0.498	0.492	0.459	0.467

Table 2: Varying segmentation frame shift (t_s), affect on average precision and R-precision, for development test on TREC-8 queries, with $t_\ell = 30s$.

3.3. Query Expansion

Following experiments on TREC-7 and TREC-8 data (Abberley et al., 1999; Renals et al., 2000) we have applied a query expansion approach whereby the relevance of potential expansion terms to original query terms is obtained by a product of term frequencies weighted by collection frequency weights. Specifically, the query expansion weight $QEW(Q,e)$ for a potential expansion term e and a query Q, across a set of nr (pseudo) relevant documents is defined as:

$$QEW(Q,e) = CFW(e) \sum_{t \in Q} CFW(t) \sum_{i=1}^{nr} TF(e,d_i) \cdot TF(t,d_i). \quad (4)$$

$QEW(Q,e)$ is used to rank the expansion terms, and the top nt are chosen to expand Q. nr is chosen such that only those documents with a relevance score of greater than $rf \cdot W$ ($rf < 1$) are used. The expanded query terms are weighted by $(nt - rank + 1)/nt$, with terms in the existing query given an additional weight of 1.

In TREC-7 and TREC-8 we obtained significant benefits from query expansion using a parallel corpus largely consisting of newspaper and newswire text from the same period as the target broadcast news corpus. In TREC-9 we constructed a parallel corpus from the following sources:

- TREC-7 SDR reference transcripts (North American broadcast news, covering parts of June 1997 – January 1998): *c.*0.7M words
- TREC-7 SDR LM text data (LA Times and Washington Post, September 1997 – April 1998): *c.*14M words.
- TREC-8/9 SDR newswire LM text data (New York Times and AP Newswire, January 1998 – June 1998): *c.*30M words (AP), *c.*17M words (NYT).

This gave a total of 135 774 documents with an average document length of 321 words and a standard deviation of 303 words. When carrying out parallel corpus query expansion experiments on TREC-8, we found that 50% of the documents used for QE were from the AP newswire, 36% from the LA Times/Washington Post corpus, 12% from the New York Times and 2% from the TREC-7 reference transcripts.

	Short Queries		Terse Queries	
QE	AveP	R-P	AveP	R-P
None	0.336	0.356	0.351	0.376
Self Only	0.436	0.446	0.432	0.438
Parallel Only	0.499	0.504	0.462	0.476
Self+Parallel	0.490	0.504	0.464	0.478
Self then Parallel	0.490	0.489	0.493	0.492
Parallel then Self	0.526	0.518	0.477	0.485

Table 3: Query expansion using target and parallel corpora, with TREC-8 queries. Self+Parallel indicates that query expansion occurs on a corpus made up of the union of the target and parallel corpora; Self then Parallel indicates that QE is first performed on the target corpus, to produce an expanded query which is then expanded a second time using the parallel corpus. Parallel then Self uses the parallel corpus first, then the target corpus.

In addition to the parallel corpus query expansion, we also experimented with query expansion using blind feedback on the main (target) corpus (also using (4)). Table 3 shows the results of query expansion purely using the target corpus, purely using the parallel corpus and various configurations using both (parallel then self, self then parallel, self and parallel simultaneously). Using our primary metric of average precision with short queries, it appears that expanding the query first on the parallel corpus, then on the target corpus is best. However, this result does not hold for terse queries. So far we have not investigated this effect further. Using a parallel corpus augmented with a copy of the target corpus produced similar results to the parallel corpus alone, as virtually all the documents used for query expansion came from the parallel corpus — probably a side-effect of mixing short 30s segments with whole stories.

4. RESULTS

In the TREC-9 SDR track we performed experiments on the main unknown story boundary (SU) condition and the contrast known story boundary (SK) condition. The same transcriptions were used in each case. Although different text retrieval parameters were used for the SU and SK conditions, the parameters were not dependent on the form of the queries (short or terse). In all cases a query expansion approach of first expanding on the parallel corpus, then on the target corpus was adopted. Table 4 summarizes the parameter settings that we used.

As well as transcriptions produced by our own recogniz-

Parameter	SU	SK
Basic Text Retrieval		
b	0.1	0.7
K	1.0	1.0
Parallel QE		
b^{PQE}	0.7	0.7
K^{PQE}	1.0	1.0
$nrmax^{PQE}$	10	10
nt^{PQE}	20	20
rf^{PQE}	0.75	0.75
Self QE		
b^{SQE}	0.1	0.7
K^{SQE}	1.0	1.0
$nrmax^{SQE}$	40	10
nt^{SQE}	10	10
rf^{PQE}	0.75	0.75
Segment Merging		
Δ^r	1600	–
Δ^f	200	–
m	0.95	–
s	1.005	–

Table 4: Parameters used for TREC-9 Evaluation Runs. b is the length parameter and K the discounting parameter in the weighting function. The additional query expansion parameters are *nt* (number of terms to add), *nrmax* (maximum number of pseudo-relevant documents) and rf (multiple of best relevance score that a document must be greater than to be used in QE). The segment merging parameters (described in section 3.2) control the ranking distance threshold below which merging may occur (Δ^r), the ranking difference (Δ^f) and score multiple (m) to determine whether a merge is equal or dominating, and the factor to increase the score by in the case of an equal merge (s).

ers (S1 and S2), we also used the following transcriptions:

1. Reference transcriptions prepared from closed captions (R1);
2. Baseline speech recognizer transcription prepared by NIST (B1);
3. Speech recognition transcriptions prepared by LIMSI (LIMSI1 and LIMSI2);
4. Speech recognition transcriptions prepared by Cambridge University (CUHTK).

Results for the SU case in the TREC-9 SDR track are presented in table 5. The average precisions are, in all cases, 20-25% lower (relative) than for TREC-8. This suggests that the TREC-9 queries may have been more difficult in some way, or that the system was over-tuned to the TREC-8 queries. Secondly, we see that the performance on the terse queries is similar to that on short queries. Note that we optimised our system using short queries on TREC-8.

Transcriptions		Short Queries		Terse Queries	
ID	WER	AveP	R-P	AveP	R-P
R1	10.3	0.409	0.419	0.418	0.425
S1	32.0	0.392	0.399	0.392	0.396
S2	29.2	0.399	0.410	0.393	0.401
B1	26.7	0.387	0.401	0.384	0.398
CUHTK	20.5	0.373	0.388	0.373	0.387
LIMSI1	21.5	0.377	0.405	0.386	0.391
LIMSI2	21.2	0.395	0.407	0.397	0.421

Table 5: TREC-9 SDR track evaluation results for story boundary unknown (SU) condition.

Transcriptions		Short Queries		Terse Queries	
ID	WER	AveP	R-P	AveP	R-P
R1	10.3	0.509	0.489	0.492	0.477
S1	32.0	0.464	0.441	0.475	0.463
S2	29.2	0.465	0.435	0.478	0.463
B1	26.7	0.462	0.447	0.469	0.451

Table 6: TREC-9 SDR track evaluation results for story boundary known (SK) condition.

Finally, following the trend of previous evaluations, the link between word error rate and text retrieval accuracy is very weak. Indeed, out of all the speech recognition transcriptions, the highest average precision on short queries is achieved using S2 (with a WER of 29%).

For contrast, results for the SK case in the TREC-9 SDR track are presented in table 6. These results follow the same form as the SU results, indicating that the low average precisions (compared with TREC-8) are not due to the segmentation/merging procedure. The relative gap between SK and SU average precision is 10–20%.

5. DISCUSSION

5.1. Terse and Short Queries

The topics for which we get a substantially better performance from short queries compared with terse queries fall into two basic types: those where the terse query is little more than an abbreviation — eg, “I L O” (130), “N B A” (165) – and those where the terse query is expressed using different words or incompletely compared with the short query (136,155,156,171). The first case might be improved by better acronym processing, the second ought to be dealt with by query expansion. More analysis is required.

5.2. Query Expansion

Previously we used parallel corpus QE only, having found that query expansion on the target corpus did not give a reliable improvement. From our experiments on TREC-8, it seems that first expanding on the parallel corpus, with the resultant expanded query being expanded again using the target corpus gives a reliable improvement. An interesting factor to be investigated is that query expansion seems to have different behaviour on terse and short queries.

5.3. Non-lexical Information

Although we were able to compute various types of non-lexical information (eg, named entities, speaker changes, sentence boundaries) we chose not to use such information in this evaluation. In the case of named entities, this was because we did not have a principled way of using them. In the case of richer boundary information, we did not feel that this would be rewarded under the evaluation metrics in use. For example, in discussions with broadcast archive users of our system, it has been apparent that returning clips that begin and end at natural boundaries would enhance their appreciation of the system; the single reference time method of denoting segments does not give any credit for accurate begin/end points.

5.4. Standard QE Corpus

A great deal of effort has gone into standardizing the acoustic model and LM training data for speech recognition, to enable better evaluation of the underlying models and algorithms. It would be of interest to increase this standardization, by specifying a baseline query expansion corpus, to be used in a contrast run (at least).

6. CONCLUSION

Our major conclusions are as follows:

- There is only a weak link between speech recognition accuracy and spoken document retrieval precision and recall;
- Query expansion using both a parallel text corpus and the target corpus is reliable and extremely effective;
- Simple fixed segmentation, followed by query-time segment merging is reliable, causing a degradation of 10–20% compared with the hand-segmented case.

ACKNOWLEDGEMENTS

Dan Ellis and Tony Robinson worked on the system for TREC-8 SDR and the work described here uses the fruits of their labour. We thank Tony Robinson and SoftSound for use of the CHRONOS decoder. This work was supported by ESPRIT project THISL (EP23495) and EPSRC project SToBS (GR/M36717).

REFERENCES

Abberley, D., S. Renals, G. Cook, and T. Robinson (1999). Retrieval of broadcast news documents with the THISL system. In *Proc. Seventh Text Retrieval Conference (TREC–7)*, pp. 181–190.

Bourlard, H. and N. Morgan (1994). *Connectionist Speech Recognition—A Hybrid Approach*. Kluwer Academic.

Clarkson, P. and R. Rosenfeld (1997). Statistical language modeling using the CMU-Cambridge toolkit. In *Proc. Eurospeech*, pp. 2707–2710.

Cook, G., K. Al-Ghoneim, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams (1999). The SPRACH system for the transcription of broadcast news. In *Proc. DARPA Broadcast News Workshop*, pp. 161–166.

Hennebert, J., C. Ris, H. Bourlard, S. Renals, and N. Morgan (1997). Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems. In *Proc. Eurospeech*, Rhodes, pp. 1951–1954.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Amer. 87*, 1738–1752.

Johnson, S. E., P. Jourlin, G. L. Moore, K. Spärck Jones, and P. C. Woodland (2000). Audio indexing and retrieval of complete broadcast news shows. In *Proc. RIAO 2000, Content Based Mulitmedia Information Access*, pp. 1163–1177.

Kingsbury, B. E. D., N. Morgan, and S. Greenberg (1998). Robust speech recognition using the modulation spectrogram. *Speech Communication 25*, 117–132.

Odell, J. J., P. C. Woodland, and T. Hain (1999). The CUHTK-Entropic 10xRT broadcast news transcription system. In *Proc. DARPA Broadcast News Workshop*, pp. 271–275.

Renals, S., D. Abberley, D. Kirby, and T. Robinson (2000). Indexing and retrieval of broadcast news. *Speech Communication 32*, 5–20.

Renals, S. and M. Hochberg (1999). Start-synchronous search for large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Processing 7*, 542–553.

Robertson, S. E. and K. Spärck Jones (1997). Simple proven approaches to text retrieval. Technical Report TR356, Cambridge University Computer Laboratory.

Robinson, A. J. (1994). The application of recurrent nets to phone probability estimation. *IEEE Trans. on Neural Networks 5*, 298–305.

Robinson, A. J., G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams (2001). Connectionist speech recognition of broadcast news. *Speech Communication*. In press.

Robinson, T. and J. Christie (1998). Time-first search for large vocabulary speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 829–832.

Robinson, T., M. Hochberg, and S. Renals (1996). The use of recurrent networks in continuous speech recognition. In C.-H. Lee, K. K. Paliwal, and F. K. Soong (Eds.), *Automatic Speech and Speaker Recognition – Advanced Topics*, pp. 233–258. Kluwer Academic Publishers.