# THE THISL SPOKEN DOCUMENT RETRIEVAL SYSTEM

*Dave Abberley (1), Steve Renals (1), Gary Cook (2) and Tony Robinson (2,3)*

(1) Department of Computer Science, University of Sheffield, UK
(2) Department of Engineering, University of Cambridge, UK
(3) SoftSound, UK

## 1. INTRODUCTION

The THISL spoken document retrieval system is based on the ABBOT Large Vocabulary Continuous Speech Recognition (LVCSR) system developed by Cambridge University, Sheffield University and SoftSound, and uses PRISE (NIST) for indexing and retrieval. We participated in full SDR mode.

Our approach was to transcribe the spoken documents at the word level using ABBOT, indexing the resulting text transcriptions using PRISE. The LVCSR system uses a recurrent network-based acoustic model (with no adaptation to different conditions) trained on the 50 hour Broadcast News training set, a 65,000 word vocabulary and a trigram language model derived from Broadcast News text. Words in queries which were out-of-vocabulary (OOV) were word spotted at query time (utilizing the posterior phone probabilities output by the acoustic model), added to the transcriptions of the relevant documents and the collection was then re-indexed. We generated pronunciations at run-time for OOV words using the Festival TTS system (University of Edinburgh).

Our key aims in this evaluation were to produce a complete system for the SDR task, to investigate the effect of a word error rate of 30-50% on retrieval performance and to investigate the integration of LVCSR and word spotting in a retrieval task. To achieve this we performed four basic experiments indexing on: transcribed text; IBM (baseline recognizer) SRT files; ABBOT SRT files; and ABBOT SRT files combined with word spotting of OOV words in the query.

This evaluation provided a stress test for our LVCSR system. In particular we developed our decoding algorithm and software to operate in a more "online mode". The result of this was the ability to decode arbitrarily long passages without segmentation into "utterances". When indexing, acoustic model computation required around 3.5× real time on a Sun Ultra 1/170, and lexical search required around 2.5× real time. At query time the word spotting component ran in about 0.25× real time per document per query.

This work was supported by ESPRIT Long Term Research Projects SPRACH (20077) and THISL (23495).

## 2. SYSTEM ARCHITECTURE

The outline of the basic THISL system is illustrated in figure 1. The ABBOT LVCSR system was used to provide approximate transcriptions of the audio documents so that the task could be treated as one of text retrieval. Since the current ABBOT system uses a finite vocabulary of around 65 000 words, a query-time wordspotter was incorporated to allow words that were OOV with respect to the LVCSR system to be retrieved.

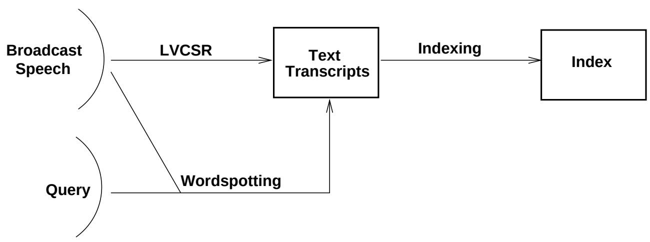


Figure 1: The indexing portion of the THISL Spoken Document Retrieval system used in TREC-6.

## 3. THE ABBOT LVCSR SYSTEM

ABBOT is a hybrid connectionist/HMM system [1] that differs from traditional HMMs in that the posterior probability of each phone given the acoustic data is directly estimated at each frame, rather than the likelihood of a phone (or state) model generating the data. This posterior probability estimation is achieved by using a connectionist network trained as a phone classifier. In the ABBOT system, a recurrent network [2] is used as the acoustic model (figure 2). Direct estimation of the posterior probability distribution using a connectionist network 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, this approach allows for an efficient search algorithm that uses a posterior probability-based pruning (section 3.3) [3] and is able to provide useful acoustic confidence measures [4].

Since the likelihood is required in the decoding process, the posterior is converted to a scaled likelihood, $L(x;q)$. This may be computed by dividing the posterior probability estimate of phone (or HMM state) $q$ given the data $x$, by the class prior $P(q)$ estimated as the relative frequency in the training data:

$$L(x;q) = \frac{P(q|x)}{P(q)} = \frac{p(x|q)}{p(x)} \; . \tag{1}$$

The assumptions underlying this acoustic model are discussed in detail in [1, 5].

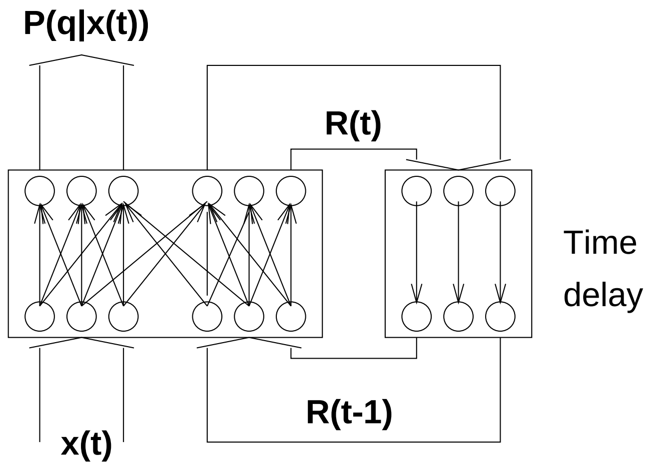


Figure 2: Recurrent network architecture used for acoustic modelling in the hybrid connectionist/HMM approach.

### 3.1. Acoustic Model

The acoustic model used in the THISL system consisted of two recurrent networks with 53 context-independent phone classes (plus silence). One network estimated the phone posterior probability distribution for each frame given a sequence of 12th order perceptual linear prediction features [6]. The other network performed the same distribution estimation with features presented in reverse order (since recurrent networks are time-asymmetric) and the two probability estimates were averaged in the log domain.

The context independent probability estimates ($P(q|x)$) were combined with a context class posterior probability $P(c|q,x)$, where $c$ is an acoustic context class, to give the joint posterior probability of context class and phone class, $P(q,c|x) = P(q|x)P(c|q,x)$ [7, 8]. The context classes were estimated using a decision tree algorithm and the context class posterior was estimated using a single layer network for each phone class. A total of 604 context-dependent phone models were used. This system is described in greater detail in [9].

The acoustic models were trained by a Viterbi training procedure using the 50 hours of Broadcast News acoustic training data from all focus conditions.

### 3.2. Language Model

The system used a 65,532 word vocabulary prepared by selecting the 80,000 most frequent words from the broadcast news text data and removing misspellings, processing errors, etc. A backed-off trigram language model was built from the Broadcast News text data (132 million words), resulting in test set perplexities typically in the range 200–300.

### 3.3. Search

The TREC/SDR evaluation was a stress test of our recognition system, since it involved performing LVCSR over the broadcast archive (around 35 hours of speech), with some "segments" of speech up to one hour long. We have extended the NOWAY start-synchronous decoder [10], to operate in an "online" mode, decoding arbitrarily long streams of speech without an additional CPU or memory burden.

NOWAY is based on a stack decoder framework and exploits the acoustic model posterior probability estimation in an effective pruning technique referred to as phone deactivation pruning [3]. This single pass algorithm is naturally factored into time synchronous state-level processing and time asynchronous word-level processing. This enables the search to be decoupled from the language model. Incremental output of the most probable final transcription is possible owing to the tree structuring of the search and the domination of language model equivalent paths.

In this evaluation, using posterior probability based phone deactivation pruning, the usual beam pruning and a unigram language model approximation at the state level we were able to decode the evaluation broadcast archive with an average of less than 1,500 model evaluations per frame (corresponding to a run time of less than 6× real time on a Sun Ultra 1/170).

## 4. INFORMATION RETRIEVAL ENGINE

Version 2.0 of the PRISE system [11] was used as the information retrieval engine for this evaluation. The system was used as supplied with no modifications. The standard PRISE stop list of 23 words and the SMART stemming algorithm were used.

## 5. RAPID WORD SPOTTING USING POSTERIOR PROBABILITIES

CSR systems can only recognize words which are contained in their lexicon. Although the ABBOT system used for these experiments had a 65k word vocabulary, approximately 1% of the words in the test set were out of vocabulary (OOV).

This raises a potential problem at the information retrieval stage: infrequent words are potentially important during retrieval but such words are most likely to be OOV and thus could have a deleterious effect on performance. To counteract this, a rapid word spotting module was added to the system to try and find any OOV query words.

The queries were scanned for OOV words. Any OOV words for which pronunciations did not exist were sent to an automatic pronunciation generator using the letter-to-sound rules in the Festival speech synthesis system [12].

The word spotting module used the context-independent posterior probability estimates from the recurrent network acoustic model, dynamically constructing word models for target words and using a set of looped phone garbage models. Any spotted words were added into the appropriate section of the speech recognition transcription. The transcriptions were then re-indexed and the standard retrieval procedure followed[1].

In the event, the only OOV word in the test queries was 'CIA' (ABBOT treats each letter of an abbreviation as a separate word and was thus expecting C. I. A.). Furthermore, no instances of it were found by the word spotting module (because it treated it as a word rather than a string of letters). Consequently, the word spotting module had no effect on system performance during this experiment.

## 6. EXPERIMENTS

### 6.1. Speech Recognition Performance

We applied the ABBOT system to the SDR test data, consisting of around 50 hours of Broadcast News, of which around 35 hours needed to be recognized. Table 1 shows the word error rate (WER) for this data set, broken down into the seven focus conditions.

We estimate the relative search error (introduced by pruning) to be around 15%. This was very much a baseline system which made no attempt to adapt to different focus conditions, or to segment out non-speech portions from the documents (e.g., musical interludes) to reduce the number of insertions.

[1]Obviously, this technique could not be used on a large corpus or in a practical system, but it does give an indication of the importance of OOV words

Table 1: ABBOT Performance at the Broadcast News Focus Conditions

| Focus | Description | WER |
|---|---|---|
| F0 | Baseline Broadcast Speech | 24.9% |
| F1 | Spontaneous Broadcast Speech | 43.2% |
| F2 | Speech / Telephone Channels | 50.8% |
| F3 | Speech / Background Music | 49.4% |
| F4 | Speech / Degraded Acoustic Conditions | 35.5% |
| F5 | Speech / Non-Native Speakers | 36.3% |
| FX | All other speech (combinations) | 55.7% |
| - | Overall | 40.1% |

### 6.2. IR Performance

We compared the performance of the system using the supplied transcript, the supplied output of the baseline recognizer and the output of the ABBOT recognizer. These results are summarized in Table 2.

Table 2: TREC SDR Results for PRISE IR System

| Transcription | Mean Rank | Mean Reciprocal |
|---|---|---|
| Reference | 11.59 | 0.6236 |
| Baseline Recognizer | 30.43 | 0.5062 |
| ABBOT LVCSR | 27.82 | 0.5784 |

Due to a problem with some of the Baseline Recognizer transcriptions, two of the (87) broadcasts had to be excluded from the final analysis. Omitting these sections from the excluded broadcasts at the indexing stage (rather than removing them after the search stage) produced results that differed by less than 2% from our submitted results. Also some of the queries used a slightly different format to that expected by our system. Changing formats again resulted in a minimal change to the system performance.

We have analysed the IR performance with respect to the WER and the focus conditions. Figure 3 shows a scatter plot of retrieval rank versus WER for the baseline and ABBOT recognizers using PRISE for the 49 retrieved target sections. The plot suggests that there is a good chance of obtaining a low retrieval rank if the WER of the target section is less than about 40%.

Figure 4 graphs the mean reciprocal retrieval performance against the WER for both recognizers. Also plotted are the cumulative WER distributions for each recognizer. In this case the WER was used as a rejection threshold, and only those documents (and corresponding queries) with a WER below that threshold were considered. For the ABBOT system, about 65% of documents had a WER of 40% or less,

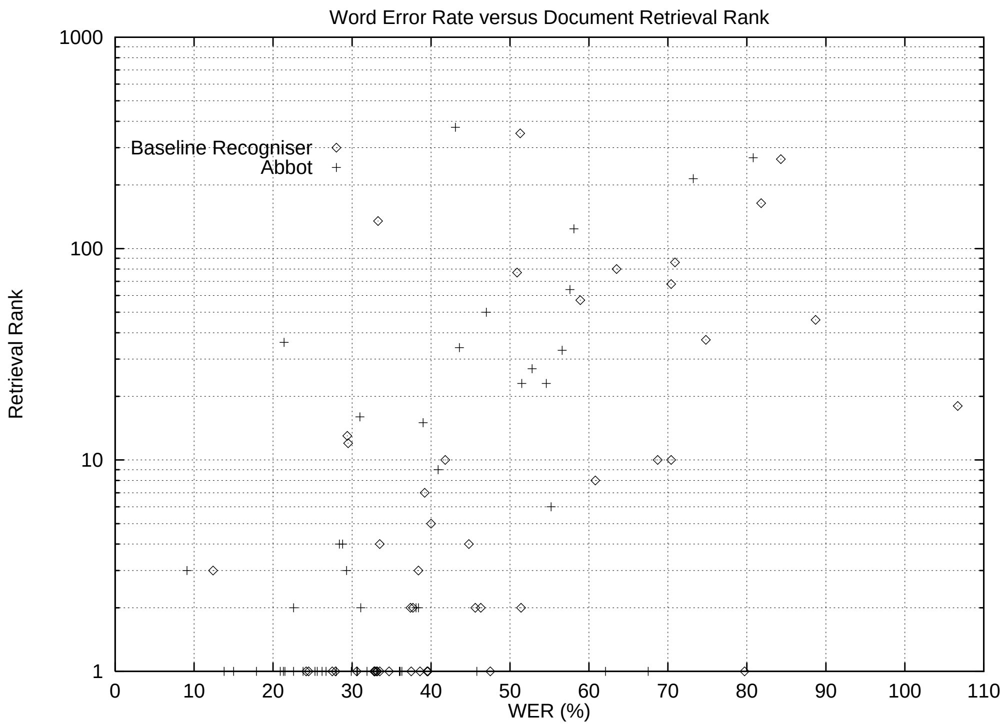


Figure 3: Document retrieval rank vs. WER.

and using those documents the mean reciprocal ranking for retrieval was around 0.75. The ROC curves reinforce the message of the scatter plot: that performance begins to fall sharply if the WER of the target document is over 40%.

Figure 5 graphs the mean reciprocal ranking against the WER for target sections containing speech largely from the F0 and FX focus conditions (twelve of each). It shows a similar picture to Figure 4: retrieval performance is good when WER is below 40%, above this figure it begins to deteriorate. Most of the F0 target sections had low WER resulting in an overall mean reciprocal figure of 0.7986 whereas some of the FX target sections had high WER contributing to an overall mean reciprocal figure of 0.6031.

## 7. CONCLUSION

Our principal goal in this evaluation was to develop a working spoken document retrieval system, and to apply our recognizer to tens of hours of broadcast speech data. We have succeeded in this objective. Future work will involve development of IR methodologies for spoken document retrieval (rather than treating the problem as text retrieval and using an "out-of-the-box" system) and to further improve the speech recognition component.

## 8. ACKNOWLEDGMENTS

Thanks to Paul Over, John Garofolo and Ellen Voorhees of NIST for help and advice with the PRISE system and the TREC evaluation. Thanks also to Alan Black of the University of Edinburgh for assistance with his Festival system.

## 9. REFERENCES

[1] H. Bourlard and N. Morgan, *Connectionist Speech Recognition—A Hybrid Approach*. Kluwer Academic, 1994.

[2] A. J. Robinson, "The application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, pp. 298–305, 1994.

[3] S. Renals, "Phone deactivation pruning in large vocabulary continuous speech recognition," *IEEE Signal Processing Lett.*, vol. 3, pp. 4–6, 1996.

[4] G. Williams and S. Renals, "Confidence measures for hybrid HMM/ANN speech recognition," in *Proc. Europ. Conf. Speech Communication and Technology*, (Rhodes, Greece), pp. 1955–1958, 1997.

[5] J. Hennebert, C. Ris, H. Bourlard, S. Renals, and N. Morgan, "Estimation of global posteriors and

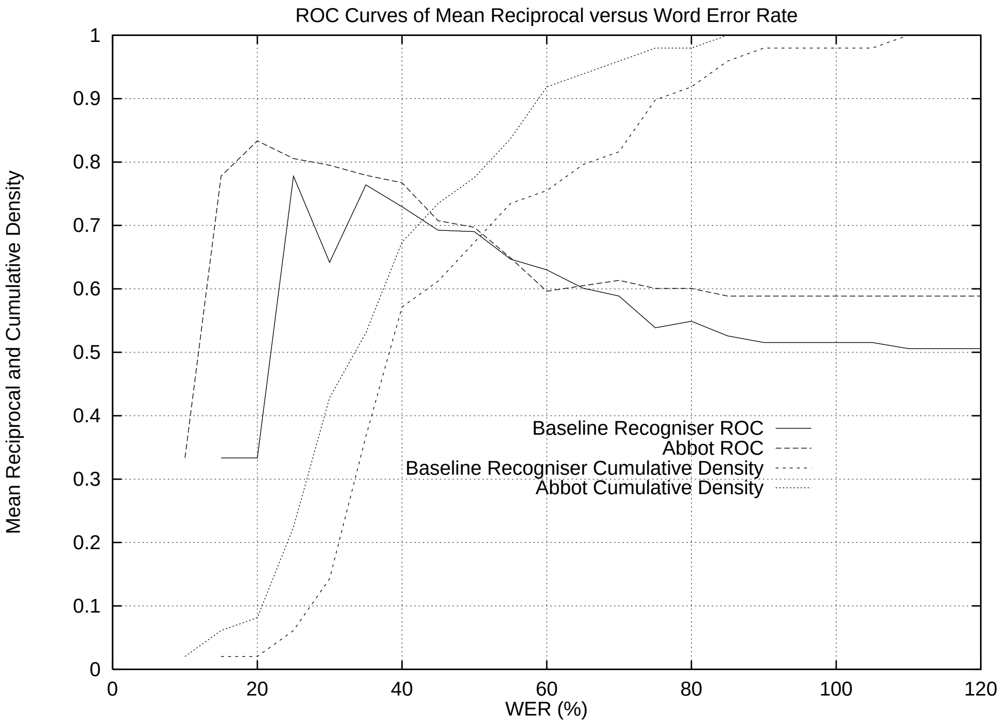


Figure 4: Mean reciprocal retrieval performance vs. WER.

forward-backward training of hybrid HMM/ANN systems," in *Proc. Europ. Conf. Speech Communication and Technology*, (Rhodes, Greece), pp. 1951–1954, 1997.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[7] H. Bourlard, N. Morgan, C. Wooters, and S. Renals, "CDNN: A context dependent neural network for continuous speech recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, (San Francisco), pp. 349–352, 1992.

[8] D. J. Kershaw, M. M. Hochberg, and A. J. Robinson, "Context-dependent classes in a hybrid recurrent network-HMM speech recognition system," in *Advances in Neural Information Processing Systems*, vol. 8, MIT Press, 1996.

[9] G. D. Cook, D. J. Kershaw, J. D. M. Christie, C. W. Seymour, and S. R. Waterhouse, "Transcription of broadcast television and radio news: The 1996 ABBOT system," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, (Munich), pp. 723–726, 1997.

[10] S. Renals and M. Hochberg, "Efficient search using posterior phone probability estimates," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, (Detroit), pp. 596–599, 1995.

[11] D. Harman, "User-friendly systems instead of user-friendly front-ends," *Journal of the American Society for Information Science*, vol. 43, pp. 164–174, 1992.

[12] A. Black and P. Taylor, "Festival speech synthesis system: system documentation (1.1.1)," Tech. Rep. HCRC/TR-83 (http://www.cstr.ed.ac.uk/projects/festival/manual-1.1.1/festival-1.1.1.ps.gz), Human Communication Research Centre, University of Edinburgh, 1997.

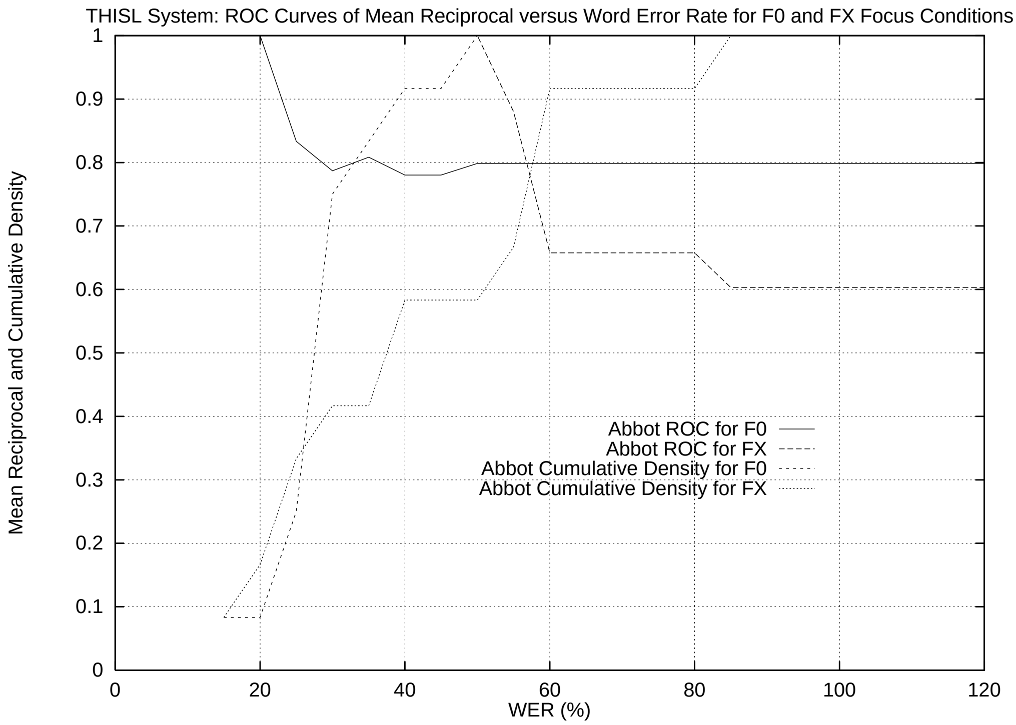


Figure 5: THISL system: mean reciprocal retrieval performance vs. WER for target documents at F0 and FX focus conditions.