REAL-TIME RECOGNITION OF BROADCAST NEWS

Gary Cook *Tony Robinson* *James Christie*

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.

ABSTRACT

Although the performance of state-of-the-art automatic speech recognition systems on the challenging task of broadcast news transcription has improved considerably in recent years, many of the systems operate in 130-300 times real-time [1]. Many applications of automatic transcription of broadcast news, eg. closed-caption subtitles for television broadcasts, require real-time operation. This paper describes a connectionist-HMM system for broadcast news transcription, and the modifications to this system necessary for real-time operation. We show that real-time operation is possible with a relative increase in word error rate of about 12%.

1. INTRODUCTION

The automatic transcription of broadcast news is a difficult problem because of the wide variety of acoustic conditions commonly found in such data. A typical show may include studio quality speech, speech corrupted with background music or background speakers, and speech over telephone channels. In addition, speaking styles range from planned speech from native speakers, to spontaneous speech from non-native speakers. Each of these factors mean that the error rate of state-of-the-art systems is around 23%. In order to achieve this level of performance many systems use multiple recognition passes, increasing the complexity of the acoustic and/or language models with each subsequent pass. This strategy can require considerable computer resources and means that many systems operate in 130-300 times real-time.

Applications of automatic transcription of broadcast news such as closed caption or teletext subtitles require real-time operation. Other applications such as audio indexing of archived material often require the transcription of huge amounts of data. Although for such applications the transcription can be performed off-line, the large amount of data necessitates near real-time operation. This paper describes the ABBOT system used for the 1997 DARPA Hub-4E Broadcast News Evaluation, and the modifications to the system required to achieve real-time operation. We show the effect of each of these modifications on word error rate, and show that real-time operation of the ABBOT system is possible with a small (around 12%) relative increase in error rate.

The layout of this paper is as follows. We first describe the ABBOT broadcast news evaluation system. This includes a description of the training data and the methods used for both acoustic and language modelling. Next we describe the modifications required for the system to operate in real-time. This includes the use of the `chronos` decoder [2] which employs a time-first search strategy, and we outline the search procedure. We then present results examining the effect of the modifications on error rate. Finally we present results for a complete system which operates in real-time.

2. THE ABBOT SYSTEM

This section briefly describes the ABBOT system used for the 1997 DARPA Hub-4E Broadcast News Evaluation. A more complete description of the system can be found in [3].

2.1. Acoustic Feature Representation

Two sets of acoustic features are used by the ABBOT evaluation system: MEL+, a 20 channel mel-scaled filter bank with energy, degree of voicing, and pitch [4], and PLP, 12th order cepstral coefficients derived using perceptual linear prediction and log energy [5]. The MEL+ and PLP features were computed from 32 msec windows of the speech waveform every 16 msec. To increase the robustness of the system to environmental conditions, the statistics of each feature channel were normalised to zero mean with unit variance over each segment.

2.2. Acoustic Models

ABBOT is a connectionist-HMM system which uses recurrent neural networks (RNN) for acoustic modelling [6]. The RNN acoustic models estimate the *a posteriori* probability of each of the phone classes given the acoustic features. These *a posteriori* probabilities are converted to scaled likelihoods and used as the observation probabilities in an HMM framework. The standard ABBOT system uses context-independent monophone classes, but the evaluation system uses word-internal context-dependent triphones. A hierarchical technique is used to estimate the context-dependent phone probabilities [7]. Single layer perceptron (SLP) models are used to estimate conditional context-class probabilities. These are then multiplied with the context-independent probabilities to give context-dependent phone probabilities. The context-dependent phone classes are chosen using phonetic decision trees [8]. This technique allows fast context training since it does not require new RNN acoustic models (which are computationally expensive to train).

The acoustic training data available consists of approximately 74 hours of data from both television and radio shows. This has been used to produce four RNN acoustic models sets, each set containing a model trained with the features presented forward in time, and a model trained with the features presented backward in time. This produces different acoustic models due to the fact that the RNN is time-asymmetric. As can be seen from Table 1 different models have been used for wideband and telephone bandwidth data. The probability estimates of the models are combined in the log domain. The combination of multiple acoustic models has been shown to provide significant reductions in error rate [9]. Both the wideband and telephone bandwidth models are augmented with SLP context class models to produce context-dependent phone probabilities. 697 word-internal context-dependent models are used for wideband data, and 604 models for the telephone bandwidth data. The different number of contexts is due to the different data used to build the models and the phonetic decision trees.

Model	Features	Parameters	Training Data
WIDEBAND-1	PLP	174k	All
WIDEBAND-2	PLP	84k	Clean
TELEPHONE-1	MEL+	84k	All (35 hours)
TELEPHONE-2	PLP	84k	All (35 hours)

Table 1: Training data and model size for the ABBOT RNN acoustic models.

The data used to train the telephone bandwidth models consists of both wideband and telephone bandwidth data. This was necessary due to the relatively small amount of telephone bandwidth training data available. To compensate for the use of wide bandwidth data during training the models were adapted to telephone bandwidth by means of a linear input network (LIN). The LIN creates a linear mapping to transform the acoustic features [10]. During recognition these transformed features are used as input to the RNN models.

2.3. Language Model and Lexicon

The ABBOT evaluation system uses both trigram and 4-gram backoff language models. These were trained from the LDC broadcast news training texts, the transcriptions of the broadcast news training data, the 1995 non-financial newswire (H4) texts, the 1995 financial newswire (H3) texts, and the 1995 Marketplace training data transcriptions. The language models were constructed using version 2.03 of the CMU-Cambridge SLM Toolkit [11]. The Witten-Bell discounting method was used for both the 4-gram and trigram models. The language models contained 7.0 million bigrams, 24.1 million trigrams, and 4.7 million 4-grams. The recognition lexicon contains the most common 65,532 words from the broadcast news training texts.

2.4. Recognition Procedure

The evaluation system uses the full set of acoustic models shown in Table 1 to produce context-dependent phone class probabilities. A two pass search strategy is used to find the most probable hypotheses. The first pass uses a trigram language model and is used to produce lattices. The `noway` stack decoder is used for this first pass. A stack based lattice to n-best decoder is then used to produce 1-best hypotheses from the lattices. A 4-gram language model is used for this second pass.

3. REAL-TIME SYSTEM

Although the ABBOT evaluation system runs in far less time than typical Gaussian mixture based systems, a number of changes are required to achieve real-time operation. There are three main differences between the ABBOT system used for the Hub-4E evaluation and the real-time system.

1. The real-time system uses context-independent acoustic models, whereas the evaluation system uses word-internal context-dependent models.
2. The real-time system uses a single pass recognition procedure with a trigram language model. The evaluation system uses a two pass recognition procedure and a 4-gram language model.
3. The real-time system uses the `chronos` decoder which employs a time-first search strategy. The evaluation system use the `noway` start-synchronous stack decoder.

The effect of using context-independent acoustic models and a single pass recognition procedure is shown in Table 2. The test data is the 1997 Hub-4E evaluation test set which consists of 173 minutes of data from various television and radio news programs [12]. Note that results are for manually segmented data — the DARPA evaluation required automatic segmentation.

The use of context-dependent acoustic models results in an 8.8% relative increase in word error rate (which is statistically significant at $p = 0.001$). However, it is not possible to use context-dependent models in a real-time system because the generation of context-dependent acoustic probabilities requires 1.3 x real-time[1].

The use of a 4-gram language model via lattice to n-best decoding results in a small (2.4%) but statistically significant (at $p = 0.05$) reduction in error rate for both the context-dependent and context-independent systems. However, it is not possible to incorporate

[1] All timings are for a 170MHz UltraSparc1 with 448 Mbytes of RAM.

lattice decoding into a real-time system since this requires 1.78 times real-time.

Operation	Time	
CI acoustic probs	0.28 x real-time	
CD acoustic probs	1.3 x real-time	
Operation	**Error Rate**	**Time**
CI noway search	28.4	6.01 x real-time
CD noway search	26.1	5.30 x real-time
CI lattice search	27.7	1.78 x real-time
CD lattice search	25.5	1.78 x real-time
System	**Error Rate**	**Time**
CI trigram (one pass)	28.4	6.29 x real-time
CD 4-gram (two pass)	25.5	8.31 x real-time

Table 2: The effect of context-independent and context-dependent acoustic models on both recognition time and word error rate.

The final two rows of Table 2 compare the error rate and time of the full evaluation system and a system with context-independent acoustic models and a single pass recognition procedure. The times presented are the total times required for all operations. The modifications to the system result in a relative increase in error rate of 11.4% and a decrease in time of 24.3%. The recognition times shown in Table 2 highlight an advantage of connectionist acoustic models. The direct estimation of *a posteriori* probabilities allows simple phone deactivation pruning in which models are pruned if their *a posteriori* probability is below a threshold [13]. This can be applied in addition to standard likelihood-based beam search, and typically reduces search time by a factor of 6. This allows the full evaluation system to operate in 8.31 x real-time, an order of magnitude less than typical Gaussian mixture systems.

3.1. Reducing the Search Time

We have shown the effects of using context-independent acoustic models and a single pass recognition strategy on both error rate and recognition time. As can be seen from Table 2 the single pass context-independent system is still not able to run in real-time. To achieve real-time operation a time-first search strategy is employed via the chronos decoder. This section briefly describes the main differences between the standard ABBOT decoder, noway, and chronos.

Noway is a stack decoder that uses time-synchronous propagation of a tree structured lexicon to extend word hypotheses. There exists one stack per time frame. The hypotheses of the earliest unprocessed stack are extend and pushed onto later stacks. The extension of word hypotheses is independent of the word history, therefore the probabilities may be factored. This allows a single pass of the extension process to extend all the hypotheses ending at a given time [13].

Chronos is also a stack decoder, however, the engine that performs the word extensions is radically different. Instead of performing a time-synchronous search a time-first search is used [2]. This reordering of the search space is most clearly seen in the case of a N state left-to-right HMM with T observations. Time-synchronous search computes the total log-likelihood to state i at time t, $\phi_i(t)$, as:

```
for t = 1 to T
    for j = 1 to N
        $\phi_j(t) = \max_{i \leq j} (\phi_i(t-1) + \log a_{ij}) + \log b_j(O(t))$
```

Whereas the time-first search reorders the loops to give:

```
for j = 1 to N
    for t = 1 to T
        $\phi_j(t) = \max_{i \leq j} (\phi_i(t-1) + \log a_{ij}) + \log b_j(O(t))$
```

The advantage of this approach is apparent when a tree structured lexicon is used since when performing a depth-first tree walk many states are in common with those previously calculated. Thus the word-extension engine can take a word string hypothesis and the range of times at which it can end, and propagate these through the pronunciation tree. Every non-pruned leaf node produces a new word string hypothesis extended by one word and an associated range of time for which it is active. The result is that many paths that have the same word history can be extended at once. However, this comes at the cost of coupling-in the word history and so forsaking noway's ability to extend many independent hypotheses simultaneously.

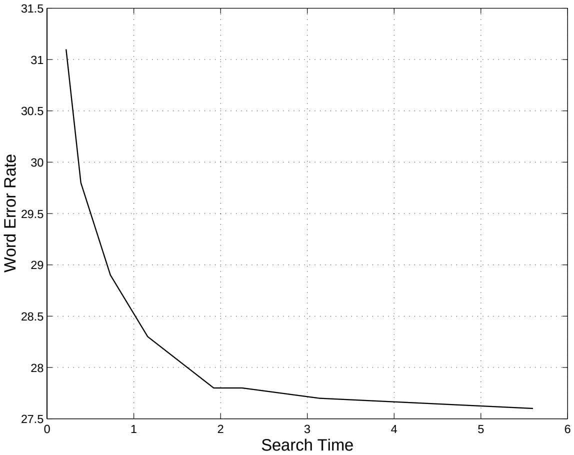

Figure 1: Search time versus word error rate on the 1997 Hub-4E evaluation test data for the chronos decoder.

There are other advantages to the chronos search, the memory required is very small enabling operation on smaller computers or better cache usage on workstations. In addition, the inner loop in chronos only requires access to arrays $\phi_i(t-1)$ and

$\phi_i(t)$ which is fundamentally faster than an inner loop that requires traversing a partially pruned tree.

The effect of using the time-first search strategy can be seen in Figure 1 which plots error rate versus search time for a context-independent system. As can be seen, the `chronos` decoder allows the search time to be reduced from 6 x real-time to real-time with only a small increase in search error (2.5%).

3.2. Real-time System Evaluation

We have compared the performance of three systems. SYSTEM-1 is the full context-dependent evaluation system. SYSTEM-2 uses the modifications to the evaluation system described in Section 3. SYSTEM-3 uses a reduced number of context-independent acoustic models, those designated WIDEBAND-2 and TELEPHONE-2 in Table 1. This reduces the time required to generate acoustic probabilities and allows more time to be spent on search.

System	Error Rate	Time (x real-time)		
		Probs	Search	Total
SYSTEM-1	25.5	1.30	7.08	8.31
SYSTEM-2	28.4	0.28	0.74	1.02
SYSTEM-3	31.6	0.05	0.88	0.93

Table 3: Error rates and time for the evaluation and real-time systems.

The results in Table 3 show that SYSTEM-2 runs in real-time with an 11.3% relative increase in error rate compared to the full evaluation system (SYSTEM-1). It can also be seen that reducing the complexity of the acoustic models has a far more dramatic effect on error rate than increased pruning during search. The reduced acoustic models of SYSTEM-3 result in an increase in error rate of 23.9%.

4. CONCLUSIONS

This paper has described the ABBOT broadcast news evaluation system, and the modifications necessary to achieve real-time operation. We have shown that the full evaluation system is capable of running in less than 10 x real-time. Moreover we have shown that real-time operation is possible with less than 12% relative increase in word error rate. Real-time performance is possible due to both the compact nature of the acoustic models, and the use of a time-first search strategy.

5. ACKNOWLEDGEMENTS

This work was partially funded by ESPRIT project 20077 SPRACH.

6. REFERENCES

1. W. Ligget and W. Fisher. Insights from the Broadcast News Benchmark Tests. *DARPA Broadcast News Transcription and Understanding Workshop*, pages 16–22, February 1998.
2. Tony Robinson and James Christie. Time-First Search for Large Vocabulary Speech Recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.
3. G.D. Cook and A.J. Robinson. The 1997 ABBOT System for the Transcription of Broadcast News. *DARPA Broadcast News Transcription and Understanding Workshop*, pages 49–54, February 1998.
4. A.J. Robinson. Several Improvements to a Recurrent Error Propagation Network Phone Recognition System. Technical Report CUED/F-INFENG/TR.82, Cambridge University Engineering Department, September 1991.
5. H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.
6. A.J. Robinson, M.M. Hochberg, and S.J. Renals. The Use of Recurrent Neural Networks in Continuous Speech Recognition. In C. H. Lee, K. K. Paliwal, and F. K. Soong, editors, *Automatic Speech and Speaker Recognition – Advanced Topics*, chapter 19. Kluwer Academic Publishers, 1995.
7. D.J. Kershaw, M.M. Hochberg, and A.J. Robinson. Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, Cambridge, MA 02142-1399, 1996.
8. D.J. Kershaw. *Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System*. PhD thesis, Cambridge University Engineering Department, September 1996.
9. M.M. Hochberg, G.D. Cook, S.J. Renals, and A.J. Robinson. Connectionist Model Combination for Large Vocabulary Speech Recognition. In *Neural Networks for Signal Processing*, volume IV, pages 269–278, 1994.
10. J. Neto, L. Almeida, M.M. Hochberg, C. Martins, L. Nunes, S.J. Renals, and A.J. Robinson. Speaker Adaptation for Hybrid HMM-ANN Continuous Speech Recognition Systems. In *Eurospeech*, pages 2171–2174, September 1995.
11. P.R Clarkson and R. Rosenfeld. Statistical Language Modelling with the CMU-Cambridge Toolkit. In *EuroSpeech*, 1997.
12. W.M. Fisher, W.S. Liggett, A. Le, J.G. Fiscus, and D.S. Pallett. Data Selection for Broadcast News CSR Evaluations. *DARPA Broadcast News Transcription and Understanding Workshop*, pages 12–15, February 1998.
13. S. Renals and M. Hochberg. Efficient Evaluation of the LVCSR Search Space Using the NOWAY Decoder. *International Conference on Acoustics, Speech, and Signal Processing*, 1:149–152, 1996.