An Overview of the SPRACH System for the Transcription of Broadcast News

Gary Cook[(1)], *James Christie*[(1)], *Dan Ellis*[(2)], *Eric Fosler-Lussier*[(2)], *Yoshi Gotoh*[(3)], *Brian Kingsbury*[(2)], *Nelson Morgan*[(2)], *Steve Renals*[(3)], *Tony Robinson*[(1)] *and Gethin Williams*[(3)]

[(1)] Cambridge University Engineering Department
[(2)] International Computer Science Institute
[(3)] Sheffield University, Dept. of Computer Science

ABSTRACT

This paper describes the SPRACH system developed for the 1998 Hub-4E broadcast news evaluation. The system is based on the connectionist-HMM framework and uses both recurrent neural network and multi-layer perceptron acoustic models. We describe both a system designed for the primary transcription hub, and a system for the less-than 10 times real-time spoke. We then describe recent developments to CHRONOS, a time-first stack decoder. We show how these developments have simplified the evaluation system, and led to significant reductions in the error rate of the 10x real-time system. We also present a system designed to operate in real-time with negligible search error.

1. INTRODUCTION

This paper describes a broadcast news transcription system developed in collaboration by several research groups as part of a European Union sponsored project known as SPRACH (for SPeech Recognition Algorithms for Connectionist Hybrids). This paper presents an overview of the complete system, further details of MLP acoustic modelling and automatic pronunciation modelling are described in the companion papers [1, 2].

The layout of the rest of this paper is as follows: first we describe the Hub-4E evaluation system. This includes a description of the acoustic features, acoustic and language models, and the recognition procedure. Next we describe the modifications to this system necessary to ensure that it runs in less-than 10 times real-time. Section 3 introduces the CHRONOS decoder, outlining the search strategy employed. We then describe some recent features added to CHRONOS to allow the SPRACH system to use a single pass recognition procedure. Finally we describe a system designed to operate in real-time.

2. EVALUATION SYSTEM

This section describes the SPRACH broadcast news evaluation system. This uses multiple acoustic models to produce three acoustic probability streams. Search is performed on each of these streams to provide three hypotheses, and these are merged to form the final system output. The following sections describe the system in detail.

2.1. Acoustic Segmentation & Features

Acoustic segmentation for the hub system is performed using the method developed by Cambridge University HTK group [3]. The less-than 10 times real-time spoke system used the CMU segmentation tools [4] to perform acoustic segmentation. Two sets of acoustic features are used: PLP, 12^{th} order cepstral coefficients derived using perceptual linear prediction and log energy, and MSG, modulation-filtered spectrogram features [5] derived from data that is first downsampled at 8kHz. The modulation-filtered spectrogram is a robust speech representation for automatic speech recognition. The robustness of the representation is based on two signal-processing strategies modelled after human speech perception. The first strategy is the emphasis of changes in the spectral structure of the speech signal (measured with critical-band-like resolution) occurring at rates of 16Hz or less. The second is adaptation to slowly-varying components of the speech signal that functions as a form of automatic gain control (AGC). 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

The SPRACH system uses both recurrent neural network (RNN) and multi-layer perceptron (MLP) models to estimate *a posteriori* context-independent (CI) phone class probabilities. Forward-in-time and backward-in-time RNN models were trained using the 104 hours of broadcast news training data released in 1997. These models use PLP acoustic features. The outputs of the forward and backward models are merged in the log domain to form the final CI RNN probability estimates. The MLP has 8000 hidden units and was trained on all 200 hours of broadcast news training data and uses MSG features [1].

Context-dependent (CD) RNN acoustic models were trained by factorisation of conditional context-class probabilities [6]. The joint *a posteriori* probability of context class j and phone class i is given by $y_{ij}(t) = y_i(t)y_{j|i}(t)$. The CI RNN estimates $y_i(t)$, and single-layer perceptrons are used to estimate the conditional context-class posterior, $y_{j|i}(t)$. The input to each module is the internal state of the CI RNN, since it is assumed that the state vector contains all the relevant contextual information necessary to discriminate between different context classes of the same monophone. Phonetic decision trees were used to chose the CD phone classes, and the SPRACH system uses 676 word-internal CD phones.

2.3. Language Models and Lexicon

Around 450 million words of text data was used to generate back-off n-gram language models. Specifically these models were estimated from:

- broadcast news acoustic training transcripts (1.6M),
- 1996 broadcast news language model text data (150M),
- LA Times/Washington Post texts (12M), Associated Press World Service texts (100M), NY Times texts (190M) — all from 1998 release of North American News text data.

The models were trained using version 2 of the CMU-Cambridge Statistical Language Model Toolkit [5]. We built both trigram and 4-gram language models for use in the evaluation system. Both these models employed Witten-Bell discounting.

Table 1 shows the language models sizes (in terms of number of n-grams) and perplexity on the 1997 evaluation data. For use in the 10x real-time spoke we pruned the trigram model using the Seymore-Rosenfeld method [7]. Despite the large increase in perplexity of the pruned model only a very small word error rate increase (0.1%) was observed.

Model	*N-grams*	*Perplexity*
Trigram	bigrams: 7.7 million trigrams: 25.6 million	174.3
4-gram	bigrams: 7.7 million trigrams: 25.6 million 4-grams: 34.4 million	164.3
Pruned Trigram	bigrams: 5.8 million trigrams: 13.2 million	190.2

Table 1: Language models sizes and perplexity on the 1997 Hub-4E evaluation test set.

The recognition lexicon contains 65,432 words, including every word that appears in the broadcast news training data. The dictionary was constructed using phone decision tree smoothed acoustic alignments. Full details of the automatic pronunciation modelling used are give in [2].

2.4. Hypothesis Combination

As described in Section 2.2 the SPRACH system uses a set of three different acoustic models. In order to use each of these models a method for combining their estimates is necessary. Frame level acoustic combination is effective for a set of estimators with the same output classes. However, it is more problematic to combine hypotheses with different output classes, such as context-independent and context-dependent acoustic models.

One method for combining models with different output classes is to combine at the hypothesis level as opposed to the acoustic probability level. We have employed the NIST recogniser output voting error reduction (ROVER) system for hypothesis combination [8]. ROVER may be operated either as a purely voting system, or in a mode in which confidence scores are taken into account. We have used the local phone posterior probability-based confidence measure as the confidence score for ROVER. The confidence measure is based purely on the connectionist acoustic model, the duration normalised log phone posterior probability [9]:

$$nPP(q_k) = \frac{1}{D} \sum_{n=n_s}^{n_e} \log \left(p(q_k|\vec{x}^n) \right) . \tag{1}$$

This is the log domain average of the posterior probability estimates, over the duration D of the phone q_k. A word-level correlate then be constructed from the phone-level confidence estimates:

$$nPP(w_j) = \frac{1}{L} \sum_{l=1}^{L} nPP(q_l) \tag{2}$$

where L is the number of phones in word w_j.

2.5. Hub System

A schematic of the system can be seen in Figure 1. Since the SPRACH system employs multiple acoustic models, a number of recognition passes are required. The recognition process can be summarised as follows:

1. Automatic data segmentation using the HTK method.
2. PLP and MSG feature extraction.
3. Generate acoustic probabilities:
   (a) Forward and backward CI RNN probabilities;
   (b) Forward and Backward CD RNN probabilities;
   (c) MLP probabilities.
4. Merge acoustic probabilities to produce three final acoustic models:
   (a) Merged forward and backward CI RNN and MLP probabilities;
   (b) Merged forward and backward CD RNN probabilities;
   (c) MLP probabilities.

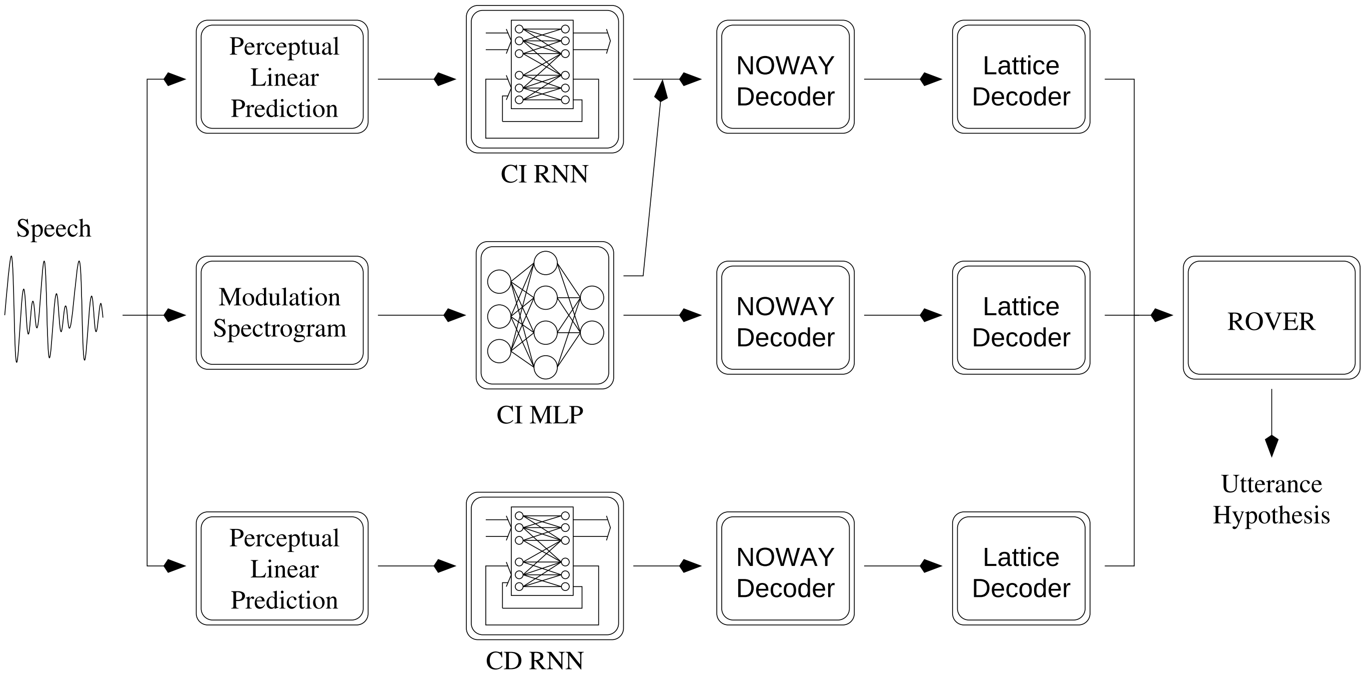

Figure 1: Schematic of the SPRACH Hub-4E Transcription System.

5. Decode using the NOWAY stack decoder and a trigram language model to produce lattices for each of the three acoustic models.
6. Decode lattices using a 4-gram language model to produce three 1-best hypotheses.
7. Generate confidence scores for each of the hypotheses.
8. Combine hypotheses (using ROVER) to produce the final system hypothesis.

The word error rates obtained by the SPRACH system on each of the 1998 evaluation test sets are shown in Table 2. The perplexity of the 4-gram language model used for lattice decoding is 165 on h4e_98_1, and 176 on h4e_98_2. The out-of-vocabulary rates are 0.47 and 0.50 respectively.

Condition	*h4e_98_1*	*h4e_98_2*
Overall	21.7	20.0
F0	11.6	13.6
F1	24.7	23.8
F2	32.4	28.4
F3	33.8	18.9
F4	15.5	23.0
F5	27.9	15.7
FX	29.6	48.3
Female	22.3	18.7
Male	20.4	20.4

Table 2: Official results for the SPRACH Hub-4E Transcription Hub System.

2.6. Less-than 10 x Real-Time Spoke System

The SPRACH 10x real-time system is very similar to the hub system described above. The differences are:

- The 10x system uses the CMU segmenter available from NIST to perform acoustic segmentation.
- A pruned trigram language model is used. This was done to reduce the amount of memory required for search.
- Lattice decoding with a 4-gram language model was not performed.

In addition much tighter pruning was required for the 10x real-time system. This was necessary because the system runs on a 450 MHz Pentium II machine, and performance is very poor when the process size exceeds the physical memory size.

As can be seen from Table 3 the system runs in considerably less-than 10 times real-time. The word error rates are shown in Table 4.

Operation	*h4e_98_1*	*h4e_98_2*
Acoustic segmentation/features	0.17 xRT	0.17 xRT
Acoustic probability generation	3.00 xRT	3.06 xRT
Search (3 probability streams)	3.12 xRT	2.40 xRT
Confidence scores/ROVER	0.37 xRT	0.32 xRT
Overall	6.70 xRT	5.95 xRT

Table 3: Timings for the SPRACH Hub-4E less-than 10 x Real-Time Spoke System.

3. RECENT DEVELOPMENTS

This section describes recent developments with the SPRACH system, based on the use of the CHRONOS decoder. CHRONOS is a time-first [10] Viterbi stack decoder with a tree based lexicon for large vocabulary CSR. Time-first indicates that the Viterbi updates of HMM state probabilities proceeds as follows:

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

with the constraint that the HMMs are left-to-right. The search traverses the tree-structured pronunciation lexicon in this manner, sharing computations between words with common pronunciation prefixes. This approach also allows large branches of the tree to be pruned out of the search in one decision, and is very memory efficient.

Continuous recognition is achieved by growing a tree of word hypotheses, where each node corresponds to an element on the stack (of which there is only one), which is ordered on time. An adaptive beam-width is used to limit the stack size. Processing involves popping the hypothesis at the top of the stack, extending it by each word in the lexicon, and pushing all resulting hypotheses back onto the stack. The finite-state property of N-gram language models can be exploited by only keeping the most probable hypothesis for each unique language model history. Within the time-first framework it is beneficial to group hypotheses with common histories into a single stack item, thus grouping future extensions into a single computation.

Condition	*h4e_98_1*	*h4e_98_2*
Overall	26.2	23.8
F0	16.0	17.1
F1	27.1	27.5
F2	38.4	33.0
F3	45.7	25.9
F4	18.7	26.4
F5	37.0	22.9
FX	36.3	53.8
Female	26.9	23.2
Male	24.3	23.5

Table 4: Official results for the SPRACH Hub-4E less-than 10 x Real-Time Spoke System.

A record of the best path probability to every frame is maintained and the search is pruned if the current hypothesis is less likely than a fixed fraction of the highest path probability. An online garbage model [11] is used to control the beam and so to limit the growth of novel path extensions [12]. Processing is complete when there are no items remaining on the stack.

We have recently added functionality to CHRONOS to allow its use for broadcast news evaluations. To this end we have implemented support for arbitrary n-gram language models, state-based decoding, and word level confidence score output. With these facilities it is possible to replace the NOWAY and lattice decoder stages of the hub system with a single pass using CHRONOS. In addition, the time-first search strategy results in significant reduction in search times, and so we have been able to produce a system for the less-than 10 times real-time spoke which uses a 4-gram language model and has only a very small search error.

Segments	*Test Set*	*Hub System*	*<10 xRT*
HTK	*h4e_98_1*	21.2% (21.6%)	
	h4e_98_2	19.6% (19.7%)	
CMU	*h4e_98_1*	21.9%	22.0% (26.2%)
	h4e_98_2	20.3%	20.9% (23.8%)

Table 5: Comparison of Hub and Spoke systems with the CHRONOS decoder.

Comparison between the CHRONOS hub and less-than 10 times real-time systems is shown in Table 5 (results with NOWAY are shown in brackets.[1]). From these results it can be seen that almost all the increase in error rate seen in the less-than 10 times real-time system is due to the use of the CMU segmentation tools as opposed to HTK segmentation. Run times for the less-than 10 times real-time system are 8.2 times real-time for h4e_98_1, and 7.8 times real-time for h4e_98_2.

4. A REAL-TIME SYSTEM

We have developed a real-time system based on the acoustic and language models used for the Hub-4E evaluation. The real-time system has the following features:

[1] These are not the same as those in Table 2. There was a minor bug in the original hub system which caused incorrect word durations to be printed. This had a small effect during scoring.

- CMU segmentation.
- PLP acoustic features.
- Forward and backward context-independent RNN acoustic models. No MLP or context-dependent models are used.
- CHRONOS decoder.
- Trigram or 4-gram language models.

The word error rate and timings for each stage of the real-time system are shown in Table 6.

	h4e_98_1	*h4e_98_2*
Operation	*x real-time*	*x real-time*
Segmentation	0.10	0.10
Feature extraction	0.07	0.07
Acoustic probabilities	0.17	0.18
Search (trigram)	0.63	0.59
Search (4-gram)	0.72	0.66
Total (trigram)	0.97	0.92
Total (4-gram)	1.07	1.01
Word Error Rate	*h4e_98_1*	*h4e_98_2*
Trigram system	27.2	25.9
4-gram system	26.8	25.2

Table 6: A Real-Time System using the CHRONOS decoder.

5. CONCLUSIONS

This paper has described the SPRACH broadcast news transcription system, and presented results from the 1998 DARPA Hub-4E evaluation. The system makes use of multiple acoustic probability streams, and we chose these to be as diverse as possible by using both PLP and MSG features. The use of diverse feature representations was found to reduce error rates for non-studio speech. We have used computatiomly efficient confidence scores based on the *a posteriori* phone class probabilities produced by connectionist acoustic models. These confidence scores were used in conjunction with ROVER for hypothesis combination.

In addition to describing the evaluation system we have outlined the time-first search procedures employed by the CHRONOS decoder. We have shown that using this search method it is possible to run a system in less-than 10 times real-time with negligible increase in error rate. By reducing the number of acoustic models we have also shown that it is possible to run a system in real-time.

References

1. Nelson Morgan, Dan Ellis, Eric Fosler, Adam Janin, and Brian Kingsbury. Reducing errors by increasing the error rate: MLP Acoustic Modeling for Broadcast News Transcription. To appear in DARPA Broadcast News Workshop, 1999.
2. Eric Fosler-Lussier and Gethin Williams. Not just what, but also when: Guided automatic pronunciation modeling for Broadcast News. To appear in DARPA Broadcast News Workshop, 1999.
3. T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, and S.J. Young. Segment generation and clustering in the HTK broadcast news transcription system. In *Proceedings of the Broadcast News Transcription and Understanding Workshop*, 1998.
4. Siegler Matthew A, Uday Jain, Bhiksha Raj, and Ricahrd M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. of the Speech Recognition Workshop*, pages 97–99. Morgan Kaufmann, February 1997.
5. Brian E. D. Kingsbury, Nelson Morgan, and Steven Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1–3):117–132, August 1998.
6. D.J. Kershaw. *Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System*. PhD thesis, Cambridge University EngineeringDepartment, 1996.
7. K. Seymore and R. Rosenfeld. Scalable Backoff Language Models. In *Proceedings of the International Conference on Spoken Language Processing*, 1996.
8. J. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.
9. Gethin Williams and Steve Renals. Confidence measures derived from an acceptor HMM. In *Proceedings of the International Conference on Spoken Language Processing*, November 1998.
10. Tony Robinson and James Christie. Time-first Search for Large Vocabulary Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 829–832, 1998.
11. H. Bourlard, B. D'hoore, and J.-M. Boite. Optimising Recognition and Rejection Performance in Wordspotting Systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 373–376, 1994.
12. S. Renals and M. Hochberg. Start-synchronous search for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 7, 1999. To appear (preprint available at ftp://ftp.dcs.shef.ac.uk/share/spandh/pubs/renals/sap99-preprint.ps.gz).