REAL-TIME RECOGNITION OF BROADCAST RADIO SPEECH

G.D.Cook† *J.D. Christie*† *P.R. Clarkson*† *M.M. Hochberg** *B.T. Logan*† *A.J. Robinson*†
C.W. Seymour†

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

*Nuance Communications, 333 Ravenswood Avenue,
Building 110, Menlo Park, CA. USA.

ABSTRACT

This paper presents a real-time speech recognition system used to transcribe broadcast radio speech. The system is based on Abbot, the hybrid connectionist-HMM large vocabulary continuous speech recognition system developed at the Cambridge University Engineering Department [1]. Developments designed to make the system more robust to acoustic variability and to improve performance when decoding spontaneous speech are described. Modifications necessary to increase the speed of the system so that it operates in real-time are also described. Recognition results and latency figures are presented for speech collected from broadcast news segments on BBC Radio 4.

1. INTRODUCTION

To date, most research on very large vocabulary continuous speech recognition has focused on clean, read speech from a single domain such as North American business news. The introduction of the *Switchboard* corpus has encouraged research into recognition of spontaneous speech covering a wide variety of domains, ranging from crime to air pollution [2]. Error rates for this task reflect the difficulty of recognising spontaneous speech, with state-of-the-art systems achieving around 50% word error rates [3].

For speech recognition technology to become widely used, systems must not only be capable of handling speech from a variety of environments (different microphones, noise, etc.) and domains (read business news, spontaneous speech, etc.), but in many real-world situations they are also required to operate in real-time. As a first step in this direction, this paper describes recent developments to Abbot, a hybrid connectionist-HMM large vocabulary continuous speech recognition system [1]. These developments are designed to

- make the system more robust to acoustic conditions such as background noise and microphone mismatch,
- improve performance when decoding spontaneous speech, and
- increase the speed of the system so that it operates in real-time.

To evaluate the performance of the system, we used radio speech recorded from BBC Radio 4.

We present results for read studio speech, for spontaneous studio speech, and for spontaneous telephone speech. We show that the system is capable of operating in real-time. We also investigate the effect of context-dependent acoustic models on both word error rates and decode times for this domain.

2. SYSTEM DESCRIPTION

The system is based on the Abbot large vocabulary continuous speech recognition system developed for the recent ARPA evaluations. For real-time transcription of radio broadcasts, a number of modifications to the original system have been required. The basic components of the real-time system are shown in figure 1 and are briefly described in the following sections. Note that the three basic components may be performed on different processors to bound the recognition time to that of the slowest process; in this case, the decoder.

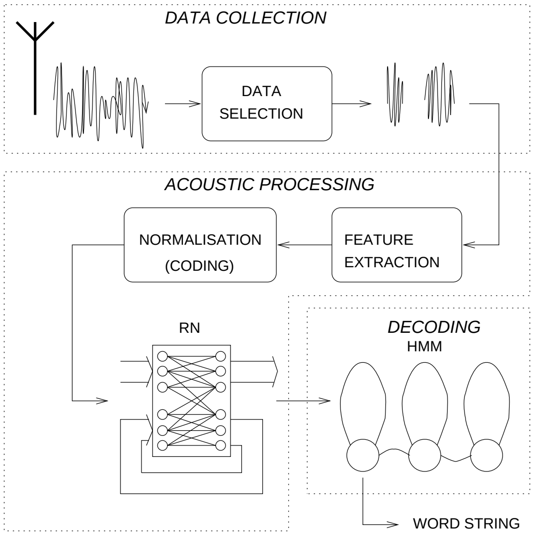

Figure 1. Real-time hybrid connectionist-HMM speech recognition system.

2.1. Data Collection

An Audiolab 8000T tuner with a five-element external aerial is used for FM radio reception. The audio signal is digitised at 16 kHz sample rate and 16 bit resolution by a Silicon Graphics Indigo workstation. For studio-recorded speech, the received signal is reasonably clean: the NIST tool *wavmd* reports typical signal-to-noise ratios of around 45 dB. Any significant mismatch between the acoustic characteristics of the radio speech and training data is likely to be due to differences in microphone response, studio acoustics, and any processing applied to the signal prior to transmission.

Segment boundaries are marked on the incoming speech using an energy measure. The energy in a 64ms frame is compared with the average energy over the last 5 seconds. If the ratio is less than a threshold, the frame is marked as silence. If greater than 0.5 seconds of silence is detected, a segment boundary is marked. This segmentation process is not essential, but is used to reduce the memory requirements of the decoder.

2.2. Acoustic Processing

2.2.1. Feature Extraction

The acoustic waveform is segmented into 32 millisecond frames every 16 milliseconds. The original version of ABBOT used a 20 channel mel-scaled filter bank with voicing features (MEL+). However, experiments with the ARPA 1995 hub 3 adaptation data have shown that Perceptual Linear Predictive (PLP) [4] cepstra coding is more robust to microphone mismatch. The results in table 1 are from 32 utterances spoken by two talkers. The row labelled relative WER increase indicates the relative increase in error rate from the Sennheiser microphone. PLP results in a mean performance gain of 19.1% over MEL+.

microphone	S/N	PLP	MEL+
Sennheiser-HMD410	38	18.5	18.5
Apple Plaintalk	13	56.3	64.6
Microsoft Sound System	14	33.9	46.6
SunMicrophone II	15	40.2	45.6
Audio-Technica AT859QMLx	17	26.8	34.8
Crown PCC-170	15	42.2	51.8
Sony ECM-55B	25	24.2	37.5
mean: (far field)		34.6	42.8
relative WER increase		101%	153%

Table 1. Results on 1995 Hub 3 Adaptation Data

2.2.2. Normalisation

The stream of feature vectors is normalised by converting each input channel into a zero mean, unit variance signal and then byte coding the resulting stream. This achieves data compression, robustness to convolutional noise, and a scaled vector appropriate for processing by the connectionist model.

The normalisation procedure usually demands knowledge of the statistics of all features over an entire utterance. For real-time operation, however, this approach is inadequate since a delay equal to the length of the current utterance is necessarily introduced. The solution has been to employ a simple running average of the past frames.

2.2.3. Acoustic Modelling

The recurrent neural network (RNN) provides a mechanism for modelling the context and the dynamics of the acoustic signal. In the real-time system, the RNN is used to map the sequence of acoustic feature vectors to a local (in time) estimate of the posterior probabilities of the phones given the acoustic data. This acoustic model replaces the standard mixture Gaussian models used in traditional HMMs and has the advantage of achieving good performance using no (or very little) context-dependent modelling.

A Viterbi based procedure is used to train the acoustic model. Each frame of training data is assigned a phone label based on an utterance orthography and the current model. The backpropagation-through-time algorithm is then used to train the recurrent network to map the acoustic input vector sequence to the phone label sequence. The labels are then reassigned and the process iterates [5].

While our standard evaluation system uses a merging of four acoustic models, this is not feasible for the real-time system described here. Hence we use just a single front-end based on PLP features as previously described.

Due to the compact connectionist architecture, generating the frame-by-frame posterior probabilities is achieved in faster than real-time on many standard workstations. A limited context-dependent implementation provides better acoustic models for clean speech resulting in fewer errors and much faster decoding [6]. We evaluated the use of a context-dependent acoustic model on broadcast radio data. This resulted in a 7.8% reduction in word error rate, but the system is no longer able to operate in real-time.

2.3. Language Model

The style of most of the speech recorded from the radio was very different from the style of text found in the corpora which are traditionally used to build language models, which frequently contain newspaper text, often focussing on business news. Such corpora tend to contain American English text, whereas we are aiming to recognise British English speech. A language model trained on American English text would bias the system against recognising common British English words and phrases, and particularly British place names. Furthermore, the style of the language used in broadcast radio speech is very different from that used in newspaper text. In particular, common phrases such as "You're listening to ..." and "This is John Smith reporting from ..." would not be found in the training corpora, and would lead to recognition errors. In addition, hesitations such as "um", "er", etc. and false starts, where a speaker begins to say one word, and then changes their mind are a major source of error.

In order to circumvent this problem, a language model was constructed by combining 100 million words of general text from the 1995 ARPA hub 4 language modelling data, and the British National Corpus [11]. The British National Corpus contains 100 million words of British English, from a wide variety of sources, of which 10 million words are transcribed spontaneous speech. Such text should hopefully match the target domain more closely.

The results of using this language model are compared with those which are obtained using the 1994 ARPA standard 20k trigram language model.

2.4. Decoder

The recognition search procedure was implemented using the NOWAY decoder [7]. This decoder, which uses a start-synchronous stack decoder approach, makes direct use of the posterior probabilities estimated by the recurrent network in phone deactivation pruning, offering a considerable speedup. Recent enhancements to the NOWAY decoder are described briefly below and in more detail in [8].

Since the language model is only applied at word ends during the search, log probability estimates within words are raised relative to word ends. This information can be exploited to achieve a more efficient search by specifying the beamwidth within words to be narrower than at word ends. This modification results in a speedup of a factor of 1.5–2.0, with little or no search error.

The decoder was modified to incorporate new sentences within an utterance. A sentence break was specified to have an acoustic realisation as a pause model with a minimum duration (typically 20 frames).

3. RESULTS

3.1. Test Data

We evaluated the system on broadcast radio speech recorded from BBC Radio 4. The test data was recorded on 22nd November, 1995, and is the programme "World at One". This is a daily news programme covering national and international news and issues. It is comprised of read speech, studio interviews, and interviews conducted over the telephone.

Speech	Duration	percentage
Read studio	9 mins 15 secs	26%
Spontaneous studio	4 mins 57 secs	14%
Spontaneous telephone	21 mins 12 secs	60%

Table 2. Composition of test data

The composition of the programme used is shown in table 2. As can be seen, the majority of the data is spontaneous speech. The spontaneous studio speech is from interviews between the news reader[1] and politicians or reporters. The telephone speech comprises mainly interviews with members of the general public and contains a far greater number of hesitations and false starts than the spontaneous studio speech.

3.2. Acoustic Models

Separate recurrent-neural-network acoustic models were trained for the wide-band studio speech and for telephone speech. The acoustic model used for wide-band speech was trained on the speaker independent training data from the wsjcam0 corpus [9]. This consists of 92 speakers reading business news from the Wall Street Journal. To train the telephone speech acoustic model, the wsjcam0 waveforms were bandpass filtered to simulate a telephone channel. The cutoff frequencies used were chosen to match the bandwidth of British telephone channels. The lower cutoff is 300 Hz and the upper cutoff is 3.4 kHz.

3.3. Context Independent Results

The results when using context-independent acoustic models and the 1994 ARPA standard 20k trigram language model can be seen in table 3. Decoding is performed in real-time for the studio speech on a HP 735/99 workstation. Telephone speech requires a greater decoding effort and takes approximately 1.4 times real-time. Acoustic processing is performed on a separate workstation and is real-time in all cases.

[1]This is an anchorperson in the US.

Speech	Sub.	Del.	Ins.	WER
Read studio	36.5	7.4	5.4	49.0
Spontaneous studio	44.5	14.2	4.2	62.9
Spontaneous telephone	56.5	7.4	9.6	73.4

Table 3. Results using context-independent acoustic models and the ARPA 1994 standard 20k trigram language model

The word error rates are high in all cases. This is due to conversational nature of the speech, microphone/channel mismatch, and the use of an inappropriate language model. The speech to be decoded is primarily spontaneous and covers a wide range of topics, while the acoustic and language models have been trained on read business news from the Wall Street Journal. In addition, the out-of-vocabulary (OOV) rate is very high; 9.0% and 9.1% for the studio and telephone speech, respectively. This is much more significant than typical values for read business news (1.3%–1.6%) [10].

The results in table 4 have been generated using the language model described in section 2.3. This results in a small reduction in word error rate for all of the types of speech. We believe the use of spoken text in the language model results in only a small improvement because of the relatively small amount of text.

Speech	Sub.	Del.	Ins.	WER
Read studio	35.2	6.5	6.3	48.0
Spontaneous studio	41.1	13.5	5.1	59.7
Spontaneous telephone	51.6	10.8	7.2	69.7

Table 4. Results using context-independent acoustic models and a language model generated from the British National Corpus and the ARPA 1995 hub 4 language modelling data

We are currently investigating methods of optimally combining language models generated from different source domain texts. This will enable us to take advantage of the large amount of business news text available, while also reflecting the spontaneous nature of most broadcast news.

3.4. Speed / Error Rate Trade Off

In order to achieve real-time performance it was necessary to increase the decoder pruning. We investigated the effect of this extra pruning on error rate.

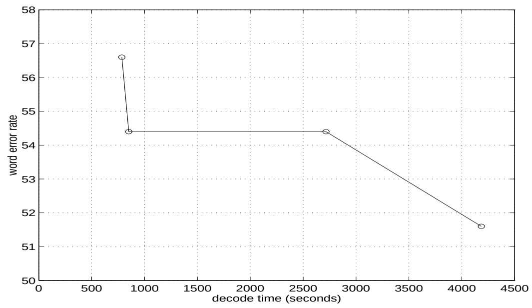

Figure 2. Decode time versus error rate for the context-independent system and the ARPA 1994 standard 20k trigram language model

For real-time performance the error rate is 54.4%. This is reduced to 51.6% when using evaluation pruning levels, however, decode time is increased almost five times.

4. ONLINE RECOGNITION

We established that the system can decode broadcast radio speech in real-time. This introduces a new parameter of interest — the time taken for a word to filter through the system and appear as text. This is a measure of latency and would be of interest in, for example, an interactive system where it is important to have not only un-interrupted transcription, but as brief a lag as possible between the utterance of a word and the emergence of its transcription.

Readings of latency were taken using live news. Every 30 seconds the word being uttered was noted and the time until its transcription (correct or otherwise) appeared was measured. The audio collection, acoustic feature computation, normalisation, and acoustic processing all ran on a Silicon Graphics Indigo, while the decoder ran on an HP 735/99. A plot of the measurements are given in figure 3 and indicate a mean latency of 6 seconds with a standard deviation of 5 seconds.

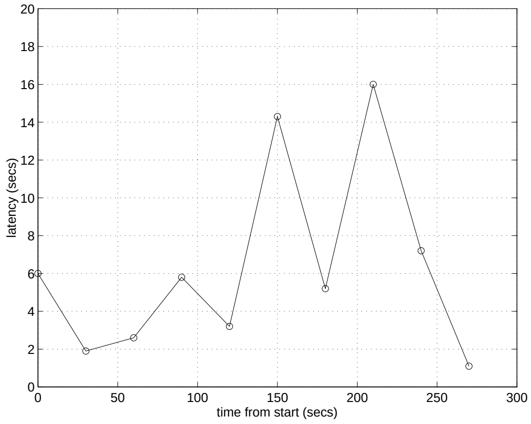

Figure 3. Latency versus time for the context-independent system and the ARPA 1994 standard 20k trigram language model

When the acoustic data is messy (unclear articulation or background noise), the decoder efficiency is reduced because many hypotheses have similar scores. The system is, therefore, prone to backlog. The latency figures show that, although often falling behind (up to 16 seconds at one stage), the system successfully regains ground and, after four and a half minutes of constant decoding, the latency is only 1.1 seconds. This condition of falling behind is a result of the system running very close to the real-time decode threshold. It is worth noting, however, that the latency did not ever fall below one second. This can be considered as the time taken for acoustic data to be processed and mapped to a word string in the case where there is no backlog. The initial delay is caused by the normalisation process. This requires initial prior statistics, and this is achieved by buffering the first five seconds of data.

5. CONCLUSIONS

There is clearly a long way to go before the performance of recognition systems on broadcast speech approaches that obtainable on read business news. Significant improvements in language models are required to more closely match the source and target domains. Because of the greater diversity of broadcast news, larger vocabularies are also required in order to reduce the OOV rates.

Acoustic mismatch is another major source of error when recognising broadcast speech — available training data consists of read business news. How can we adapt acoustic models to varying broadcast conditions and spontaneous speech? These are open issues we are currently investigating.

6. ACKNOWLEDGEMENTS

This work was partially funded by ESPRIT basic research grant 6487, WERNICKE. We would like to acknowledge S.G. Cooper, D.J. Kershaw, S.J. Renals, S.R. Waterhouse, and P.S. Zolfaghari for their contributions to the work in this paper.

REFERENCES

[1] M.M. Hochberg, G.D.Cook, S.J. Renals, A.J. Robinson, and R.S. Schechtman. The 1994 ABBOT Hybrid Connectionist-HMM Large-Vocabulary Recognition System. *Proc. of Spoken Language Systems Technology Worshop, ARPA*, 1995.

[2] J. Godfrey, E. Holliman, and J. McDaniel. SWITCHBOARD: Telephone Speech Corpus for Research Development. *Proc. ICASSP*, pages 517–520, 1992.

[3] P. Jeanrenaud *et al.* Reducing Word Error Rate on Conversational Speech from the Switchboard Corpus. *Proc. ICASSP*, pages 53–56, 1995.

[4] H. Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. *J. Acoust. Soc. Am.*, 87:1738–1752, 1990.

[5] A.J. Robinson. An application of Recurrent Nets to Phone Probability Estimation. *IEEE Transactions on Neural Networks*, 5(2):298 – 305, March 1994.

[6] D.J. Kershaw, M.M. Hochberg, and A.J. Robinson. Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System. Technical Report CUED/F-INFENG/TR217, Cambridge University Engineering Department, July 1995.

[7] S. Renals and M. Hochberg. Efficient Search using Posterior Phone Probability Estimates. *ICASSP*, pages 596–599, 1995.

[8] S. J. Renals and M. M. Hochberg. Efficient Evaluation of the LVCSR Search Space Using the NOWAY Decoder. *Proc ICASSP*, 1996.

[9] J. Fransen, D. Pye, A. J. Robinson, P. C. Woodland, and S. J. Young. WSJCAM0 Corpus and Recording Description. Technical Report CUED/F-INFENG/TR.192, Cambridge University Engineering Department, 1994.

[10] D. Pye, P. C. Woodland, and S. J. Young. Large Vocabulary Multilingual Speech Recognition using HTK. *Proc. EUROSPEECH*, pages 181–184, 1995.

[11] Oxford University Computing Services. *Users Reference Guide for the British National Corpus*, May 1995.