Connectionist Speech Recognition of Broadcast News

A.J. Robinson(3), G.D. Cook(4), D.P.W. Ellis(5), E. Fosler-Lussier(6),
S.J. Renals(7) and D.A.G. Williams(8)
(0) Engineering Department, University of Cambridge
(1) International Computer Science Institute
(2) Department of Computer Science, University of Sheffield
(3) Now with SoftSound, formerly (0)
(4) Now with Dragon Systems U.K. R.&D. Ltd, formerly (0)
(5) Now with Columbia University, formerly (1)
(6) Now with Bell Labs, Lucent Technologies, formerly (1)
(7) Still with (2)
(8) Now with Dragon Systems U.K. R.&D. Ltd, formerly (2)

November 22, 2000

Corresponding author: Tony Robinson (ajr@softsound.com)
SoftSound Ltd, St John's Innovation Centre, Cowley Road, Cambridge, CB4 0WS, United Kingdom.

Abstract

This paper describes connectionist techniques for recognition of Broadcast News. The fundamental difference between connectionist systems and more conventional mixture-of-Gaussian systems is that connectionist models directly estimate posterior probabilities as opposed to likelihoods. Access to posterior probabilities has enabled us to develop a number of novel approaches to confidence estimation, pronunciation modelling and search. In addition we have investigated a new feature extraction technique based on the modulation-filtered spectrogram, and methods for combining multiple information sources. We have incorporated all of these techniques into a system for the transcription of Broadcast News, and we present results on the 1998 DARPA Hub-4E Broadcast News evaluation data.

1 Introduction

This paper describes recent advances in a set of interrelated techniques collectively referred to as "Connectionist Speech Recognition". Specifically, it describes those advances made by the SPRACH consortium[1] that we have incorporated into our large vocabulary recognition system. Our system uses connectionist, or neural, networks for acoustic modelling, and this single change from the conventional architecture has led us to investigate new models and algorithms for other components in the speech recognition process. It is the aim of this paper to describe the differences between our system and those of others, both to explain the advantages and disadvantages of the connectionist approach and also to facilitate the transfer of the advances we have made under the connectionist framework to other speech recognition methodologies.

We are concerned with the transcription of North American Broadcast News. This domain, and the associated evaluation programme, is described by Pallett (2000). The final systems in the paper are trained with the full 200 hour Broadcast News acoustic training set released by NIST (containing 142 hours of speech), although the development systems reported through the paper have been trained on approximately half of this data (except where noted). For development we have used the 1997 evaluation data released by NIST (termed Hub-4E-97); for rapid turnaround of experiments we also defined a 173 utterance subset with a duration of approximately 30 minutes, termed Hub-4E-97-subset. This subset results in a slightly higher word error rate (WER) than the full Hub-4E-97 data set. Language models for all experiments were constructed from the transcripts of the acoustic training data (about 1 million words), 150 million words of broadcast news transcripts covering 1992–96, and a variety of newspaper and newswire sources. Unless otherwise stated our systems operated with a vocabulary of 65,532 words and a trigram language model.

The paper is organised as follows: section 2 briefly outlines the connectionist approach to speech recognition, showing how we estimate a stream of posterior probabilities of phone occurrence. This approach is particularly amenable to the incorporation of novel features and we describe the modulation-filtered spectrogram (MSG) features and their benefits in section 3. Another pleasing feature of the connectionist stream-of-probabilities approach has been a computationally efficient means of computing confidence levels that is discussed in section 4. These confidences are applied to the task of pronunciation modelling in section 5. Over the past five years we have developed two new search algorithms based on the stack decoder and these are described in section 6. Finally, all these elements are combined into a complete system that is applied to the Broadcast News domain in section 7.

2 Connectionist speech recognition

We use the phrase "connectionist speech recognition" to refer to the use of connectionist models, or artificial neural networks, as acoustic modelling in a speech recognition system (Bourlard and Morgan, 1994). The acoustic model in a "traditional", generative hidden Markov model (HMM) system is typically composed of a mixture of Gaussian densities and trained to estimate the probability that a given state, q_k, emitted the observation, $\mathbf{x}^t$, *i.e.*, $P(\mathbf{x}^t|q_k)$. However, an alternative approach is to model $P(q_k|\mathbf{x}^t)$, relating the two as:

$$P(q_k|\mathbf{x}^t) \quad = \quad \frac{P(\mathbf{x}^t|q_k)P(q_k)}{P(\mathbf{x}^t)} \quad . \tag{1}$$

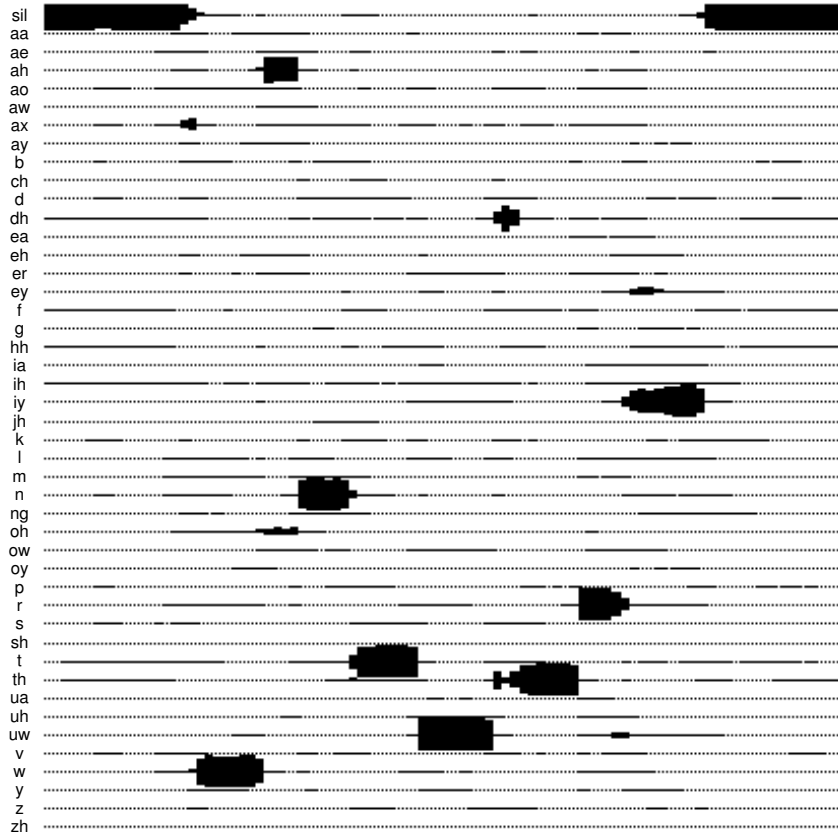[1] http://tcts.fpms.ac.be/sprach/sprach.html

Figure 1: Phone class posterior probabilities for an instance of the word sequence *one two three*. Time follows the horizontal axis and the duration of the utterance is 1.5 seconds.

Since $P(\mathbf{x}^t)$ is independent of any given state sequence, it acts only as a scaling factor and $P(q_k|\mathbf{x}^t)$ may be divided by $P(q_k)$ to obtain what we refer to as a scaled likelihood, $P(\mathbf{x}^t|q_k)/P(\mathbf{x}^t)$. Typically, a phone class is allocated a single state q_k and so the acoustic model is required to estimate relatively few probabilities at each time frame. These posterior probabilities may be used to form a meaningful display of the speech process. Such a display is shown in figure 1 for the word sequence *one two three* with the corresponding phone sequence /w ah n t uw th r iy/. Each horizontal line in the figure corresponds to a phone, q_k, and its width is proportional to the posterior probability of the class. Phones are arranged in alphabetic order with the exception of silence (sil), which appears at the top. Time runs horizontally and the relative positions of each of the eight phones can be easily located in this example. (The most significant confusion occurs for /dh/ and /th/ at the start of *three*.) The diagram highlights the factorization of the speech recognition process into the task of estimating phone class posterior probabilities followed by that of finding the most probable word sequence. Furthermore, the existence of this intermediate stream of probabilities has proven to be useful for the combination of acoustic features, the estimation of confidence measures, the generation of new pronunciations and as a component of fast search algorithms.

2.1 Multi-layer perceptrons and recurrent neural networks

We have made use of two connectionist architectures, a standard feedforward multi-layer perceptron (MLP) (Bourlard and Morgan, 1994) and a recurrent neural network (RNN) (Robinson, 1994). Both take acoustic features at the input, and output phone class posterior probability estimates. Traditional HMMs typically make use of a vector of acoustic features and their first and second temporal derivatives. In comparison, the connectionist acoustic models make use of more acoustic information. In the case of the MLP we commonly use a nine frame window centred on the frame of interest, *i.e.,* the sequence $\mathbf{X}_{t-4}^{t+4} =$
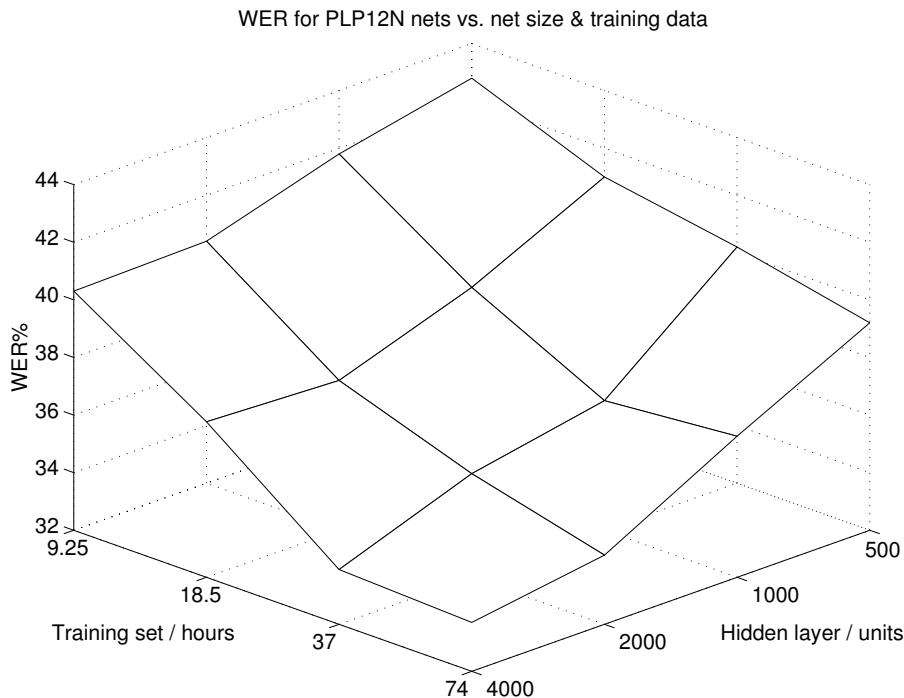
Figure 2: Surface plot of system word error rate as a function of the amount of training data and the hidden layer size.

$\{\mathbf{x}^{t-4}, \ldots, \mathbf{x}^t, \ldots, \mathbf{x}^{t+4}\}$. The recurrent connections of an RNN incorporate information from the start of the sequence to four frames into the future, estimating $P(q_k|\mathbf{X}_1^{t+4})$.[2] We have used three-layer MLPs with a hidden layer of up to 8,000 units (or about 2.5 million parameters) and RNNs, which use feedback to accomplish the hidden layer processing, with a feedback vector of 256 elements (approximately 80,000 parameters). In both cases we use Viterbi training, with the network parameters estimated using stochastic gradient descent. Further details of the MLPs and RNNs that we have used, including size, format and training issues can be found in Bourlard and Morgan (1996) and Robinson et al. (1996).

In any HMM system there is a tradeoff between the number of distinct states and the amount of acoustic training data available to model each state. We have experimented with this balance and found that the connectionist models work best with fewer states (and hence more acoustic information per state) than traditional HMMs. This gives rise to a simpler probability stream to pass to the decoding stage—for most of our work we use monophone models (*i.e.,* no phonetic context), as in figure 1.

2.2 Training set vs number of parameters

Concern has previously been expressed, by ourselves and others, that our context-independent (CI) models could not easily exploit large amounts of training data for tasks such as Broadcast News. To test this, we used Perceptual Linear Prediction (PLP) (Hermansky, 1990) features to train MLPs for every combination of 4 network sizes (hidden layers of 500, 1,000, 2,000 and 4,000 units) and 4 training-set sizes (corresponding to 1/8, 1/4, 1/2 and all of the 74 hours of 1997 Broadcast News training data, `bntrain97`). The performance of these nets, measured in percentage word error rate (WER) on the Hub-4E-97-subset, is plotted as a surface in figure 2. The results show that significant gains are to be had from increasing model size in step with training data over the entire range tested. This strongly suggested that using the full 142 hours of 1997 and 1998 training data and doubling the network size once again to 8,000 hidden units should be a worthwhile effort. Using special-purpose hardware (Wawrzynek et al., 1996), this training run took

---

[2] The use of more than one frame of context violates the usual HMM assumption that the current observation is conditionally independent of past (and future) observations given the HMM state.

4

21 days, totaling some 10^{15} parameter updates. Again the larger networks gave lower error rates than the smaller ones: while our best 4,000 hidden unit net achieved an error rate of 27.2% on Hub-4E-97-subset when combined with the RNN acoustic model, the error rate for the final 8,000 hidden unit net in the same conditions was 25.4%, reflecting both the increased model complexity and the improved training targets made available through the iterative relabelling of the training data with each successive acoustic model.

As training recurrent neural networks is also computationally expensive, the RNNs are also trained on dedicated hardware—in this case an array of 28 TMS320C30's (33 MFLOP each) (Morgan et al., 1992). Memory limitations with this hardware have restricted the RNN size to 256 feedback units.

2.3 Connectionist probability combination

In addition to using two different network architectures, in the course of this work we have trained models using different features, on different amounts of data and representing different balances of acoustic conditions. As the RNN output is (asymmetrically) dependent on previous inputs, we often train pairs of RNNs, with one running forwards in time and the other backwards (having a dependency on all frames to the *end* of input). With all of these probability estimates, a method for combination of probability streams is necessary; the most straightforward combination is framewise. A simple linear combination is an obvious choice, but we have found that a log domain combination consistently outperforms linear ones (Hochberg et al., 1995). For a set of estimators $P_n(q_k|\mathbf{x}^t)$,

$$\log P(q_k|\mathbf{x}^t) = \frac{1}{N} \sum_{n=1}^{N} \log P_n(q_k|\mathbf{x}^t) - B, \tag{2}$$

where B is a normalizing constant so that $\sum_k P(q_k|\mathbf{x}^t) = 1$. It is our experience that each extra model combined in this way reduces the number of errors made by about 8% for up to four models.

3 The Modulation-Filtered Spectrogram

Two different acoustic feature sets were used in this work: the normalized cepstra derived from the modified Perceptual Linear Prediction (PLP) (Hermansky, 1990) as used in the 1997 ABBOT evaluation system (Cook and Robinson, 1998), and a novel representation called the modulation-filtered spectrogram (MSG) (Kingsbury, 1998; Kingsbury et al., 1998).

The modulation-filtered spectrogram was developed as a representation for speech recognition robust to adverse acoustic conditions such as reverberation. It is inspired by several observations concerning the human auditory system, namely:

- the use of peripheral frequency channels whose bandwidth and spacing increase with frequency;

- within each channel, preferential sensitivity to particular modulation frequencies (e.g. below 16 Hz);

- adaptation to the signal characteristics in each channel over a short timescales on the order of hundreds of milliseconds;

- the use of multiple alternative representations for the same basic signal.

The development of MSG features originated in experience with RASTA processing (Hermansky and Morgan, 1994), which also used an auditory-like spectral analysis, and achieved robustness though band-pass filtering of subband energy envelopes. In devising MSG, Kingsbury took the same basic concept of filtering in the modulation domain but experimented with a much larger range of filters, looking specifically for systems that performed well with reverberant speech and with mismatched train/test conditions. Automatic gain control was included to further normalize the characteristics within each channel, mirroring adaptation in humans.

Multiple parallel representations were initially introduced by separating the real and imaginary (quadrature) components of the complex modulation-domain band-pass filter used in early prototypes. The resulting filters approximated a low-pass filter and its differential, and the considerable gains seen in using the

5

separated components mirrors the benefits found when using the deltas of conventional features. As with all the parameters, later experiments systematically varied the filter characteristics of the two parallel banks to optimize performance. Full details of the development of each stage are presented in (Kingsbury, 1998); the final form of the processing is as follows:

1. The speech signal is segmented into 32 ms frames with a 16 ms frame step, each frame is multiplied by a Hamming window, and the power spectrum is computed with a fast Fourier transform (FFT).

2. The power spectrum is accumulated into critical-band-like frequency channels via convolution with a bank of overlapping, triangular filters that have bandwidths and spacings of 1.0 Bark and cover the 100–4,000 Hz range. The critical-band power spectrum is converted into an amplitude spectrum by taking the square root of the filterbank output.

3. The critical-band amplitude trajectories are filtered through two different infinite impulse response (IIR) filters in parallel: a lowpass filter with a 0–16 Hz passband and a bandpass filter with a 2–16 Hz passband. The IIR filters are designed to have at most ± 1 sample of group delay ripple in their passbands.

4. The lowpass and bandpass streams are each processed through two feedback AGC units (Kohlrausch et al., 1992) in series. For the lowpass stream the first AGC has a time constant of 160 ms and the second has a time constant of 320 ms, while for the bandpass stream the first AGC has a time constant of 160 ms and the second has a time constant of 640 ms.

5. Finally, the features from both the lowpass and bandpass streams are normalized to have zero mean and unit variance on an utterance-by-utterance basis.

For comparison, figure 3 shows spectrograms of the same brief speech segment processed by the PLP front-end (shown before cepstral transformation) alongside the corresponding first bank of MSG features (with a modulation passband of 0 to 16 Hz). The automatic gain control and modulation filtering can be seen to greatly emphasize onsets while suppressing sustained energy. It is clear that the two representations are very different.

Experiments to compare the performance of the different front ends were performed using an MLP classifier. Our goal in developing the MLP-based acoustic model was to provide a useful complement to the existing RNN classifier based on the PLP features. Previous experience with combining multiple representations has shown that an additional input is most useful when it is most different, and that combining two systems can show dramatic improvements relative to either system alone—even when one system's isolated performance is significantly worse, as long as it is also significantly different (Wu, 1998).

Further, the MLP system was based on acoustic data bandlimited to 4 kHz, thereby reducing the acoustic distinction between telephone-quality speech and the rest of the data. Although throwing away the upper half of the spectrum slightly reduces the overall system performance, it serves the goal of increasing the differences from the original RNN system. Later experiments with MSG features based on full-band acoustics showed no significant improvement over the bandlimited variants when combined with PLP-based models.

We conducted initial investigations with several different feature sets (PLP, RASTA-filtered PLP (Hermansky and Morgan, 1994) and MSG) to confirm that the MSG features were indeed the best choice for combination. These experiments are summarized in table 1. In each case we trained a 2,000 hidden-unit MLP classifier on the features calculated for one-half of the bandlimited `bntrain97` data (*i.e.,* 37 hours), using target labels from the 1997 ABBOT system. Decodings were performed with relatively aggressive pruning and the test set was the Hub-4E-97-subset described above. For each feature set the WER was computed using both the MLP classifier alone and combined with an RNN baseline system (as in section 2.3). The RNN baseline system consisted of a set of four CI recurrent networks, trained on the `bntrain97` data using PLP cepstral coefficients.

These results confirmed our expectations that MSG features based on 4 kHz bandlimited data offered a significant benefit as part of a combination with the RNN baseline, and were therefore used as the basis of our subsequent work.

As table 2 shows, MSG features (as used with an MLP and based on reduced bandwidth data) were not as good as PLP features from the full bandwidth data (as used with the RNN). However, even in
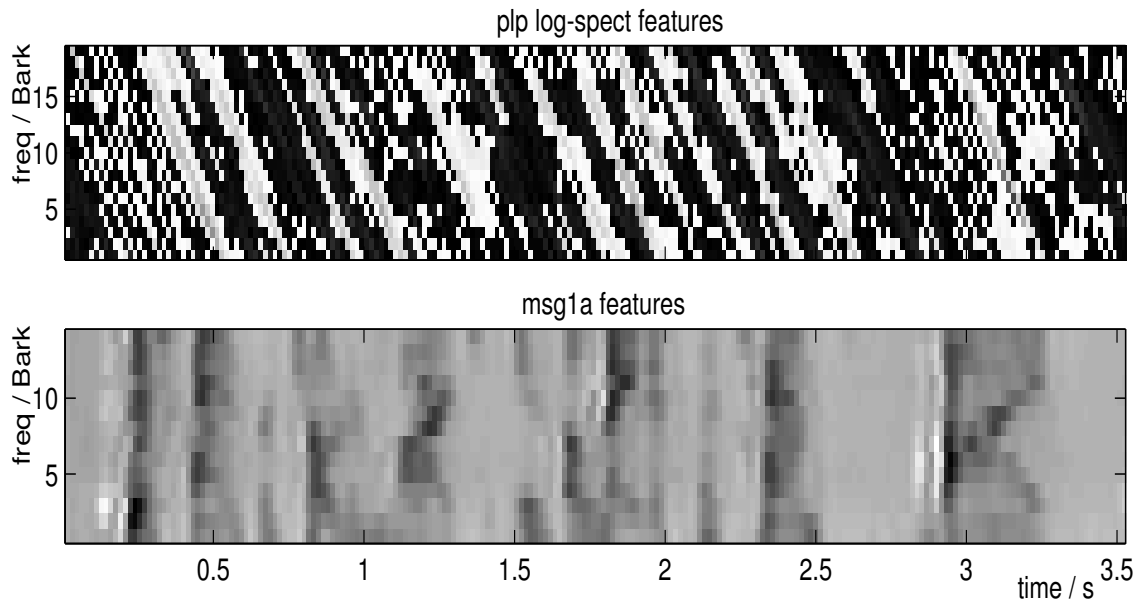
6

Figure 3: Comparison of spectrograms based on PLP and MSG processing, illustrating the effect of modulation-domain filtering. The vertical frequency axis is in Bark units; because the MSG features are based on signals bandlimited to 4 kHz, they extend only to 14 Bark, whereas the PLP features continue to 19 Bark.

Feature	Elements	WER% (alone)	WER% (RNN combination)
RNN baseline			33.2
plp12N	13	36.7	31.1
ras12+dN	26	44.4	32.5
msg1N	28	39.4	29.9

Table 1: Comparison of different acoustic features alone (MLP classifier) and in combination with the RNN baseline system on Hub-4E-subset. plp12N refers to 12th order cepstral coefficients based on PLP-smoothed spectra; ras12+dN refers to 12th order cepstra of RASTA-filtered PLP spectra plus deltas (note that the RASTA filter was not adjusted for the 16 ms frame step); msg1N refers to the MSG features. These error rates are elevated due to relatively aggressive pruning used in the decoding.

this case, the word error rate was significantly improved by combining these two subsystems. The first column of the table shows this improvement. However, the second column shows that almost none of this improvement was obtained on the prepared, studio component of the shows (labelled F0 according to the Hub-4 conventions); rather, as shown in the third column, essentially all of the improvement was achieved for signals that were degraded in some manner.

4 Confidence Measures

A confidence measure on the output of a recognizer may be defined as a function that quantifies how well the recognition model matches some speech data. Further, a confidence measure must be comparable across utterances (*i.e.,* there must be no dependence on $P(\mathbf{X})$). Although confidence measures are most obviously useful for utterance verification (Cox and Rose, 1996; Williams and Renals, 1999), they have been shown to be useful throughout the recognition process. 'Unrecognizable' portions (such as music or unclear acoustics) may be filtered from an unconstrained stream of audio input to a recognizer, reducing both the WER and the computational expense of decoding (Barker et al., 1998). The search space may

system	ALL	F0	non-F0
RNN using PLP	27.2%	14.4%	35.2%
MLP using MSG	29.7%	17.7%	37.2%
Frame level combination	25.4%	14.3%	32.3%

Table 2: Word error rates for RNN subsystem using PLP features, MLP subsystem using MSG features, and combined system. F0 is the studio quality prepared speech condition. Scores are for the Hub-4E-97-subset.

be structured using confidence measures, for example by pruning low confidence phones (Renals, 1996) or by weighting the language model probability (Neti et al., 1997). It is also possible to design model combination weighting functions based on confidence measures (section 7). Finally, confidence measures may be used for diagnostic purposes, by evaluating the quality of model match for different components of a recognition system (Eide et al., 1995; Williams and Renals, 1998).

The posterior probability of a hypothesized word (or phone) given the acoustics is a suitable basis for a confidence measure. While it is not straightforward to estimate this from a large generative HMM system, the connectionist acoustic model performs local posterior probability estimation, without requiring explicit normalization. Such computationally inexpensive estimates may be used to provide purely acoustic confidence measures directly linked to the acoustic model. Their relationship to the acoustic model enables them to be applied at the frame-, phone-, word- and utterance-levels, providing a basis for distinguishing between different sources of recognition error. We have investigated a variety of such confidence measures (Williams and Renals, 1999), including two that were applied in the SPRACH system: duration normalized posterior probability and per-frame entropy. We have compared these to a confidence measure that also incorporates language model information—lattice density.

Duration Normalized Posterior Probability   The connectionist acoustic model[3] makes a direct estimate of the posterior probability of a phone model given the acoustics, $P(q_k^t|\mathbf{X}_1^{t+4})$. This estimate may be regarded as a frame-level confidence measure. By making an independence assumption we can extend this frame-level estimate to the phone- and word-levels. A consequence of such an assumption is an underestimate of the posterior probability for a sequence of frames. To counteract this underestimate, we define a duration normalized posterior probability measure, $\text{nPP}(q_k)$, for a phone hypothesis q_k with start and end frames t_s and t_e using the Viterbi state sequence:

$$
\begin{aligned}
\text{nPP}(q_k) &= \frac{1}{D}\sum_{t=t_s}^{t_e} \log\left(P(q_k^t|\mathbf{X}_1^{t+4})\right) \\
D &= t_e - t_s + 1 \, .
\end{aligned}
\tag{3}
$$

This measure may be extended to a hypothesized word w_j, by averaging the phone level confidence measures for each of the L phones constituent to w_j (Bernardis and Bourlard, 1998):

$$
\text{nPP}(w_j) = \frac{1}{L}\sum_{l=1}^{L} \text{nPP}(q_l) \, .
\tag{4}
$$

Per-Frame Entropy   A more general measure of acoustic model match is the per-frame entropy of the estimated local posterior distribution. This measure, which is directly available from the network output stream, does not require the most probable state sequence, and so may be used for filtering (Barker et al., 1998; Williams and Ellis, 1999) and frame-level model combination prior to decoding. The connectionist acoustic model is based on a discriminative phone classifier; the per-frame entropy confidence measure is based on the hypothesis that a sharply discriminant posterior distribution indicates that the classifier is well matched to the data. We define the per-frame entropy $S(t_s, t_e)$ of the K local phone posterior probabilities,

---

[3]In this section we consider the RNN acoustic model. However, posterior probabilities conditioned on X_{t-c}^{t+c} may also be used.

averaged over the interval t_s to t_e as:

$$S(t_s, t_e) = -\frac{1}{D} \sum_{t=t_s}^{t_e} \sum_{k=1}^{K} P(q_k^t | \mathbf{X}_1^{t+4}) \log \left\{ P(q_k^t | \mathbf{X}_1^{t+4}) \right\} \ . \tag{5}$$

Lattice Density   The above purely acoustic confidence measures may compared with a confidence measure derived using both the acoustic and language models such as the *lattice density* $LD(t_s, t_e)$—the density of competitors in an *n*-best lattice of decoding hypotheses (Hetherington, 1995). This is computed by counting the number of unique competing hypotheses (NCH) that pass through a frame and averaging the counts over the interval t_s to t_e:

$$LD(t_s, t_e) = \frac{1}{D} \sum_{t=t_s}^{t_e} NCH(t) \ . \tag{6}$$

Recognition errors may be caused by unclear acoustics (high levels of background noise) or mismatched pronunciation models. A low value of the duration normalized posterior probability at the state-level, coupled with a high per-frame entropy implies a poor acoustic model match, whereas a high value of the duration normalized posterior probability at the state-level but a low value at the word-level, and a low per-frame entropy, implies a mismatched pronunciation model.

The performance of the three measures for the task of recognizer output verification, as measured by the *unconditional error rate* (UER) of the associated hypothesis test, is illustrated in figure 4. (The UER counts any valid words in rejected utterances as deletion errors, and thus indicates the overall system performance possible with the confidence scheme.) These experiments indicate that the acoustic measures (a) perform better at the phone-level than the word-level; (b) perform better on noisy (FX) rather than planned studio (F0) speech; and (c) perform at least as well as the lattice density confidence measure for the broadcast news data. These purely acoustic confidence measures have been used in the SPRACH system for pronunciation modelling (section 5), search (section 6) and model combination (section 7).

5   Pronunciation modelling

Our two goals when developing pronunciation models were (a) to improve the pronunciation models for common words; and (b) to find baseforms for novel words. We were interested in applying lessons learned from automatic pronunciation modelling work in the Switchboard domain (*e.g.,* Riley et al. (1998)), while not relying on the presence of hand-transcribed phonetic data. Thus, we were seeking a method of introducing new pronunciation variants of existing words suggested by the acoustic models of the recognizer, while trying to limit the number of spurious pronunciations. Modelling patterns of pronunciation variation at the phone level, rather than the word level, allowed us to generalize to infrequently-occurring words.

We were also faced with the task of generating baseforms for words such as *Lewinsky* that did not occur in the ABBOT 96 dictionary—a vocabulary that spans current affairs is crucial for recognizing Broadcast News.

Both of these tasks required considerable new machinery. To reduce effort when implementing disparate functions, we reorganized our pronunciation software around a finite state grammar (FSG) decoder (Figure 5). The modularization of Viterbi alignments into an FSG compilation stage and a decoding stage allowed for novel compilation techniques without having to completely rewrite the decoder. Thus, as long as the output of the procedure was a valid finite state grammar, we could easily implement new pronunciation models (such as a letter-to-phone model for novel words).

5.1   Learning new pronunciations from data

In these experiments, we wanted to expand the set of pronunciations in the baseline lexicon from the ABBOT Broadcast News transcription system (Cook et al., 1997), which contained an average of 1.10 pronunciations for each member of the 65K word vocabulary. Candidate alternative pronunciations were generated by using the neural net acoustic models of the recognizer, in a six step process:

1. Generate canonical phonetic transcriptions with the baseline dictionary.
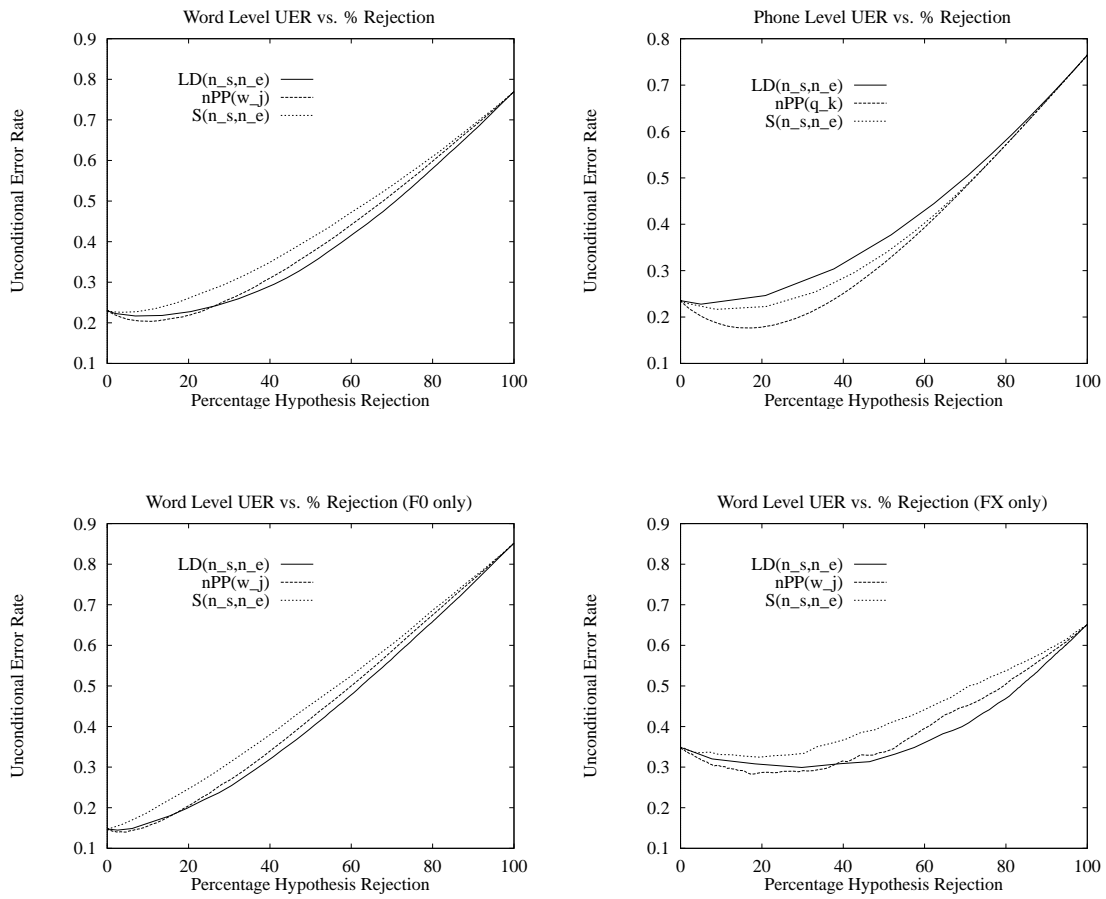
Figure 4: A comparison of confidence measure UERs for word-level hypotheses (*Upper Left*) and hypotheses obtained using phone-level decoding constraints (*Upper Right*) and for word-level hypotheses in the F0 condition (*Lower Left*) and the FX condition (*Lower Right*), on Hub-4E-97.
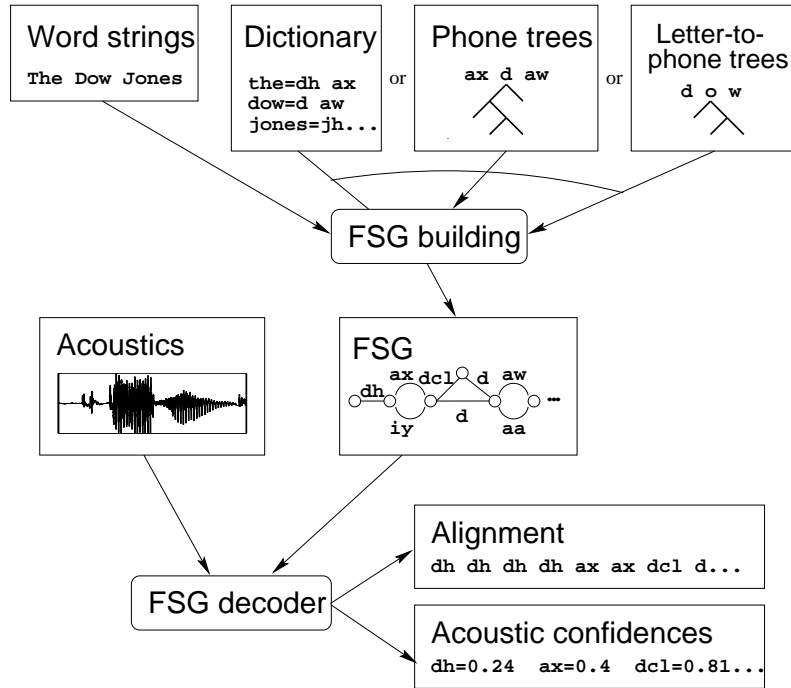
Figure 5: Decoding for different pronunciation models using finite state grammars.

Dictionary	WER (%)	Decode time
Baseline (ABBOT 96)	27.5	21 x RT
Smoothed Trees: no pruning	27.1	72 x RT
log count (SPRACH 98)	26.9	33 x RT
confidence log count	26.6	30 x RT

Table 3: Word error rate on Hub-4E-97-subset for various pruning methods. Decoding time is given in multiples of real-time on a 300 MHz Sun Ultra-30 using the start-synchronous decoder described in section 6.

2. Generate an alternative acoustic-model based transcription via phone recognition.

3. Align the canonical and alternative transcriptions so that each canonical phone is either deleted or matches one or more alternative transcriptions.

4. Train a decision tree to predict P(alternative|canonical,context). Smooth out noise in the distribution by pruning low probability phone pronunciations in the leaves of the d-tree (typically p<0.1).

5. Retranscribe the corpus: for each utterance in the training set, create a network of pronunciation variations using the d-trees, and align with the acoustic models to get a smoothed phone transcription.

6. Collect smoothed pronunciation variants into a new dictionary.

Since the FSG decoder produced both word and phone alignments, the new smoothed phone transcription was easily converted into a new lexicon by collecting all of the pronunciation examples for each word and determining probabilities by counting the frequency of each pronunciation. New pronunciations were then merged with the ABBOT 96 dictionary using simple interpolation.[4]

---

[4]Recognition experiments using deleted interpolation for dictionary merging, where the interpolation coefficients are dependent on the count of the word in the corpus, showed no improvement over simple interpolation.

The interpolated dictionary, with 1.67 pronunciations per word, required an almost four-fold increase in decoding time over the ABBOT 96 dictionary when evaluated on Hub-4E-97-subset (table 3). Therefore, we devoted efforts to selectively pruning pronunciations from the dictionary to increase decoding speed. For the SPRACH system, setting a maximum number of baseforms per word based on the log count of training set instances gave useful reductions in decoding time without sacrificing recognition accuracy.

After the 1998 evaluation, we continued to improve the dictionary. For each baseform in the unpruned lexicon, we computed, over the 1997 BN training set, the average acoustic confidence scores based on the local posterior probabilities from the acoustic model (Section 4). Dictionary baseforms were reselected by the log-count pruning scheme according to their confidence-based rankings rather than by Viterbi frequency counts; this provided a small boost to performance both in terms of decoding time and recognizer accuracy. This is an interesting result: we suspect that choosing pronunciations based on acoustic confidence may lead to less confusability in the lexicon.

5.2 Pronunciation generation for novel words

To develop pronunciations for novel words, we employed a technique similar that of Lucassen and Mercer (1984). In essence, this model is almost identical to the dictionary construction algorithm discussed above: we replaced the mapping between canonical and alternative pronunciations in steps 1-3 with an alignment between the letters of words and the canonical pronunciations in the dictionary.

For all words in the canonical dictionary, we generated an alignment between the letters of the word and its corresponding phones using a Hidden Markov Model: starting from an arbitrary initial correspondence, a mapping between letters and phones was successively re-estimated using the Baum-Welch algorithm. The next step involved constructing d-trees based on this aligned data to estimate the probability distribution over phones given a central letter and the context of three letters to the left and right.

For each novel word that we desired in the vocabulary, we constructed a (bushy) pronunciation graph for that word given its spelling (cf. step 5 above). One could just find the best n paths in the pronunciation graph *a priori* to determine alternative pronunciations. However, we had volunteers speak each word in isolation (by reading word lists), and then used the recorded speech data to find the best path *with respect to the acoustic models*. The matching of this graph to the acoustic models via alignment is the critical gain of this technique; using an off-the-shelf text-to-speech system would likely produce pronunciations with properties different from our baseline dictionary.

The Viterbi alignment of the graph provides the best pronunciation for each word, as well as the acoustic confidence score (equation 3). Using this procedure, we generated 7,000 new pronunciations for novel words in the 1998 SPRACH system. The acoustic confidence measures turned out to be critical: spot checks of the novel word pronunciations with high confidence scores revealed them more linguistically reasonable than the low-confidence pronunciations; thus, we were able to focus hand-correction efforts on the lower-confidence words.

We also employed a number of other pronunciation modelling techniques, which we only summarize here. Multi-word models developed in the vein of Finke and Waibel (1997) did not show any improvement in our final system. Extending the phone d-tree models to allow contextual variation based on the word context, syllable context, speaking rate, language model probability, and other factors did decrease the WER, although insignificantly (0.3% absolute). For further details on these developments, the reader is invited to consult Fosler-Lussier and Williams (1999) and Fosler-Lussier (1999).

6 Large vocabulary search

The search problem in a large vocabulary speech recognition system involves finding the most probable sequence of words given a sequence of acoustic observations, the acoustic model and the language model. An efficient algorithm for carrying out such a search optimally and exhaustively is the Viterbi algorithm (forward dynamic programming), which can find the most probable path through a time/state lattice. However, in the case of LVCSR an exhaustive search requires a very large amount of computation, due to the vocabulary size (typically 65K words). Indeed, once higher than first-order Markov language models are used, the search space expands massively and exhaustive search is impractical. Modifications have been

made to the Viterbi algorithm to combat this including beam search (Lowerre and Reddy, 1980), dynamic memory allocation (Odell, 1995) and multi-pass approaches, in which earlier passes using simpler acoustic and language models restrict the available search space (Schwartz et al., 1996).

These Viterbi-based searches are termed *time-synchronous*: the probabilities of all active states at time t are computed before the probabilities at time $t+1$. A different class of search algorithms is based on stack decoding (Jelinek, 1969), a heuristic search approach related to A* search (Nilsson, 1971). Stack decoders are based on a best-first approach in which partial hypotheses are extended in order of their probability. We have developed two novel search algorithms based on stack decoding, both of which have proven to be well-suited to large vocabulary speech recognition. The NOWAY algorithm (Renals and Hochberg, 1995, 1999) uses a start-synchronous search organisation; the CHRONOS algorithm (Robinson and Christie, 1998) uses a *time first* organisation. Although the details of the two approaches are different they share the advantages of stack decoding (also common to other approaches (Paul, 1992; Gopalakrishnan et al., 1995)):

- Decoupling of the language model from the acoustic model; the language model is not used to generate new hypotheses;

- Easy incorporation of longer term language models (*e.g.,* trigrams, four-grams) and other knowledge sources;

- Search organisation supports more sophisticated acoustic models (*e.g.,* segmental models);

- Relatively straightforward implementation.

Most of the development experiments reported in this paper were carried out using the NOWAY algorithm; the more recent CHRONOS algorithm was used for the final results reported in section 7.

6.1 Search Organisation

The two search algorithms have several common features including a tree-structured lexicon, stacks implemented using a priority queue data structure, a search organization enabling shared computation and the use of beam search pruning.

The start-synchronous decoder processes partial hypotheses in increasing order of reference time. This is implemented as a sequence of stacks, one for each reference time. The stacks are processed in time order, earliest first. All hypotheses with the same reference time are processed in parallel, allowing different language model contexts to share the same acoustic model computations. Hypotheses are extended by all possible (unpruned) one word extensions and the resultant extended hypotheses are pushed onto the stack with the relevant reference time, after a check for "language model equivalent" hypotheses on that stack (figure 6). Hypothesis extensions through the tree are carried out in a breadth-first (time-synchronous) manner.

The time-first decoder employs a single stack for all hypotheses ordered first by reference time, then by log probability. Computation is shared by each stack item representing a language model state (rather than an individual path history) and containing a range of reference times. Figure 7 illustrates this process, all three hypotheses that have the language model history "<START> THE" are pulled from the stack and are propagated through the HMM, resulting in three new hypotheses with language model history "THE SAT" that are inserted onto the stack. The "time-first" designation comes from the order of the HMM dynamic programming computations (numbered *1...12* in italic); the ordering computes the first HMM state for all possible times (subject to pruning) before proceeding with the next state. In conventional time-synchronous all states for time 6 would be processed before time 7 leading to the order $1, 2, 5, 3, ...9, 12$.

6.2 Pruning

Both search algorithms use beam search approaches, augmented by confidence measures and adaptive beam widths. A record of the maximum path probability to each frame is kept and the search is pruned if the current hypothesis extension is less than a fixed fraction of the recorded maximum. To avoid explicit lookahead, an online garbage model (Bourlard et al., 1994) based on the average local posterior probability
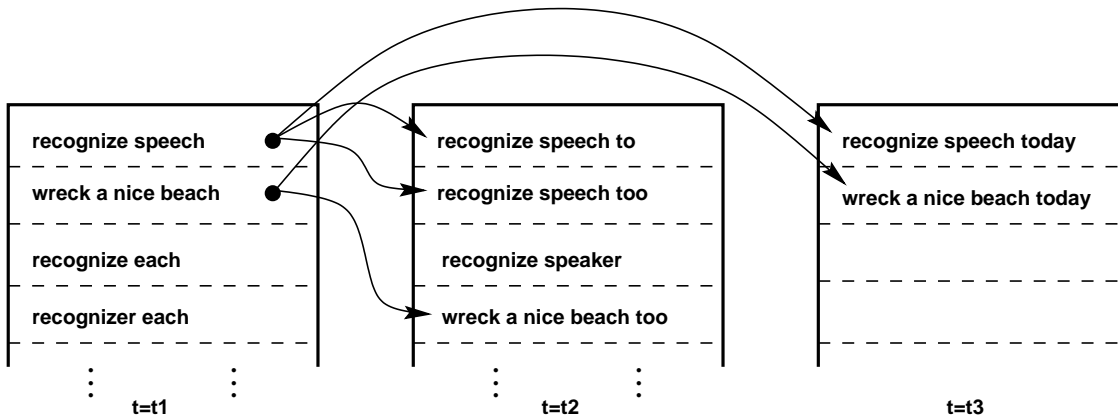
recognize speech

wreck a nice beach

recognize each

recognizer each

t=t1

recognize speech to

recognize speech too

recognize speaker

wreck a nice beach too

t=t2

recognize speech today

wreck a nice beach today

t=t3

Figure 6: Start synchronous search organization.

Stack

Time−first hypothesis extension

Language model state (2 words)	Time range	Path probabilities and history pointers
<START> THE	5−8	Pr(5...8), Hist(5...8)
<START> CAT	6−9	Pr(6...9), Hist(6...9)
<START> SAT	6−10	Pr(6...10), Hist(6...10)
THE CAT	8−11	Pr(8...11), Hist(8...11)
THE SAT	10−12	Pr(10...12), Hist(10...12)
THE MAT	11−14	Pr(11...14), Hist(11...14)

POP

PUSH

O(6) s 1

s 2   ae 5

O(8) s 3   ae 6

O(9) s 4   ae 7   t 10

ae 8   t 11
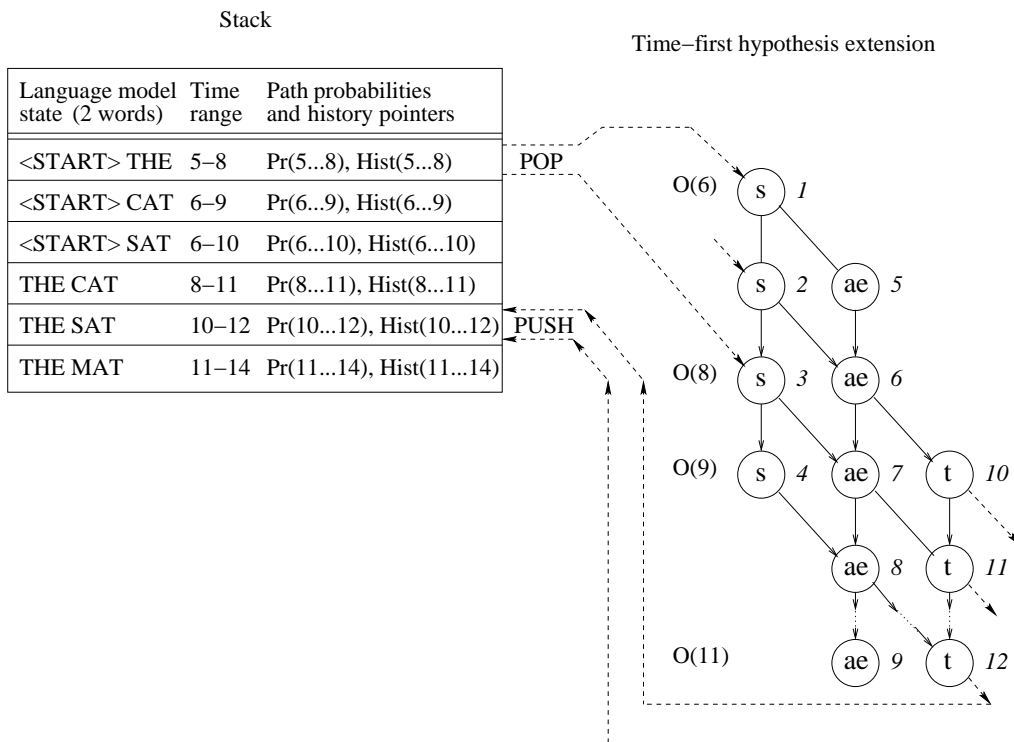
O(11) ae 9   t 12

Figure 7: Time first search organisation

14

estimates of the top k phones is used to control the beam. Additionally an adaptive beamwidth has been employed in the time-first search to prevent a large expansion of the stack size in unfavourable acoustic conditions. This may be as simple as a fixed maximum stack size, but we have found the following scaling scheme to be successful:

$$B(t) = B^* + \gamma \log \left(\frac{N}{n(t) + 0.5} - 1 \right), \tag{7}$$

where $B(t)$ is the log beam width used at time t, B^* is the target beam, $n(t)$ is the number of stack items at time t, N is the maximum stack size (typically 8191) and γ is the weight of the adaptive element in the beam.

Confidence measures have also been used to directly prune the search space. Phone deactivation pruning (Renals, 1996) sets a threshold on the local posterior probability estimates: phones with posterior probabilities below this threshold are pruned from the search. This is an efficient way to reduce the search space by a large amount—typically by over 60% with little or no additional search error.

7   Current system description and results

This section describes the system we used in the 1998 Hub-4E evaluations and our recent improvements to it. The approach taken is to produce multiple acoustic probability streams, and to combine these at both the probability-level and the hypothesis-level. In this manner we are able to take advantage of the properties of different acoustic features and models.

The task for the 1998 Hub-4E evaluation was the transcription of segments from a number of broadcast news shows. This is an unpartitioned task, i.e. no speaker change boundaries or change of acoustic conditions are provided. We used both the segmentation strategy freely available from CMU (Siegler et al., 1997) and that developed by Hain et al. (1998), termed the HTK segmentation method in these experiments. We have developed a *hub* system that is not constrained in any way, and two *spoke* systems: A *10x* system that is constrained to run 10x slower than real-time on a 450MHz Pentium-II PC and a *real-time* system on the same platform.

7.1   Acoustic Models

Both RNN and MLP models were used 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 used PLP acoustic features. The outputs of the forward and backward models were merged in the log domain to form the final CI RNN probability estimates. The MLP had 8,000 hidden units and was trained on all 200 hours of broadcast news training data downsampled to 4 kHz bandwidth. MSG features, as described in section 3, were used in this case.

Context-dependent (CD) RNN acoustic models were trained by factorisation of conditional context-class probabilities (Kershaw, 1996). The joint *a posteriori* probability of context class j and phone class k is given by $P(q_{CD_j}, q_{CI_k} | \mathbf{X}_1^{t+4}) = P(q_{CI_k} | \mathbf{X}_1^{t+4}) P(q_{CD_j} | q_{CI_k}, \mathbf{X}_1^{t+4})$. The CI RNN was used to estimate the CI phone posteriors, $P(q_{CI_k} | \mathbf{X}_1^{t+4})$, and single-layer perceptrons were used to estimate the conditional context-class posteriors, $P(q_{CD_j} | q_{CI_k}, \mathbf{X}_1^{t+4})$.

The input to each of the context-class perceptrons was the internal state of the CI RNN, since it was assumed that the state vector contained all the relevant contextual information necessary to discriminate between different context classes of the same monophone. Phonetic decision trees were used to choose the CD phone classes, and 676 word-internal CD phones were used in the system.

7.2   Language Models and Lexicon

Around 450 million words of text data were 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 the LDC's 1998 release of North American News text data.

The models were trained using version 2 of the CMU-Cambridge Statistical Language Model Toolkit (Clarkson and Rosenfeld, 1997). We built both trigram and four-gram language models for use in the evaluation system. Both these models employed Witten-Bell discounting.

The recognition lexicon contains 65,432 words, including every word that appears in the broadcast news training data. The dictionary was constructed using the phone decision tree smoothed acoustic alignments of section 5.

7.3 The Hub and 10x System

The hub and 10x systems use the same models and algorithms, and differ only in the levels of pruning applied during the search. Five component connectionist acoustic models were used: CI RNN (forward and backward), CI MLP and CD RNN (forward and backward). These were combined at the frame level by averaging the probability estimates in the log domain as described in Section 2.3. This results in three probability streams as illustrated in figure 8. Combining estimates from models with different feature representations takes advantage of the different information available to the different models. In the same manner it is possible to combine information from multiple decoder streams. Evidence suggests that using different acoustic and/or language models results in systems with different patterns of errors, and that the differences can used used to correct some of these errors (Fiscus, 1997). To this end the streams were individually decoded, and the resultant hypotheses were merged using the ROVER (recognizer output voting error reduction) system (Fiscus, 1997) developed at NIST. The component hypotheses were weighted using the word-level nPP confidence measure (equation 4).

The overall recognition process is shown in figure 8 and 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.

5. Decode with a four-gram language model using the CHRONOS stack decoder generating confidence measures

6. Combine hypotheses (using ROVER with confidence-weighting) to produce the final system hypothesis.

The results of this system are shown in table 4. This table suggests that the differences in segmentation strategy have a greater effect than the tighter pruning for the 10x system. However, both changes are very small and both together amount to only a 5–15% relative increase in the number of errors.

7.4 A Real-Time System

We have developed a real-time system by using only the context-independent RNNs in the Hub-4E system. This system has the following features:
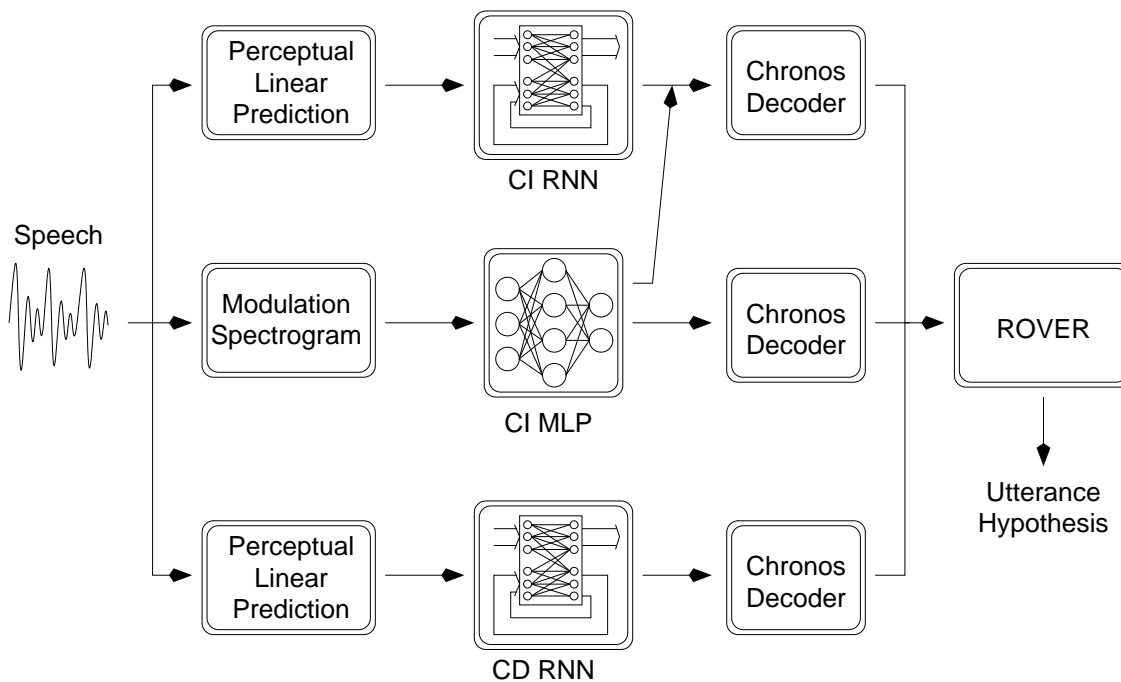
- CMU segmentation.

Figure 8: Schematic of the SPRACH Hub-4E Transcription System.

Segmentation	Test Set	Hub System	10x System
HTK	h4e_98_1	21.2%	
HTK	h4e_98_2	19.6%	
CMU	h4e_98_1	21.9%	22.0%
CMU	h4e_98_2	20.3%	22.9%

Table 4: Results for the Hub and 10x systems on the 1998 Broadcast News evaluation data.

- PLP acoustic features.

- Forward and backward context-independent RNN acoustic models. No MLP or context-dependent models are used.

- CHRONOS decoder.

- Trigram or four-gram language models.

The word error rate and timings for each stage of the real-time system are shown in Table 5.

Operation	h4e_98_1 x real-time	h4e_98_2 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 (four-gram)	0.72	0.66
Total (trigram)	0.97	0.92
Total (four-gram)	1.07	1.01
Word Error Rate	*h4e_98_1*	*h4e_98_2*
Trigram system	27.2	25.9
four-gram system	26.8	25.2

Table 5: Results for the real-time system.

It is encouraging that the single pass decode with a four-gram language model takes only about 10% longer than the trigram language model.

8 Conclusions

This paper has presented several developments to a connectionist speech recognition system. The most salient aspect of the system is the connectionist acoustic model that provides a stream of local phone class posterior probability estimates. This acoustic model is well suited to the incorporation of large amounts of temporal context as well as experimental features with unknown statistical behavior (such as those described in section 3), and we have found that our system performs well using a simple stream of monophone probabilities. The availability of this local posterior probability stream has allowed us to investigate several new approaches to the components of an LVCSR system.

Firstly, the posterior probability stream has proved amenable to the combination of a complementary set of acoustic features and phone classifier architectures. A considerable benefit was obtained through merging the outputs of component connectionist networks via log-domain averaging at the frame level. The most accurate performance was achieved using RNNs trained forward and backward in time on PLP features, and an MLP using MSG features; the MSG stream provided particular gains for acoustics outside of the studio-quality F0 condition.

A computationally efficient search algorithm is especially important for the Broadcast News domain as the recognizer is not only expected to operate under more challenging acoustic conditions, but may also be applied to very long, unsegmented recordings. We have described two novel search algorithms based upon a stack decoder. In addition to the advantages inherent to stack decoding, these algorithms have been able to exploit the stream of local posterior probabilities for pruning the search. A consequence of these decoder advances is that our system is able to operate in real time with reasonable memory requirements—a feature we have found useful as we move to decoding large quantities of audio data for information retrieval and extraction tasks. At the addition of modest computational expense, we have found small performance gains through the use of four-gram language models.

Local phone class posterior probability estimates have proven to be a good basis for measures of confidence. The availability of these measures at the frame-, phone-, word- and utterance-levels coupled with

their explicit link to the recognition models may well prove useful for distinguishing between decoding errors due to unclear acoustics and those due to mismatched pronunciation models. We have presented results that indicate that for the task of utterance verification on Broadcast News data these measures perform at least as well as the more conventional lattice density measure that is derived from both the acoustic and language models. These confidence measures were shown to be useful for the tasks of filtering 'unrecognisable' acoustics from the input to a recognizer and for evaluating alternative baseforms in an automatic pronunciation learning process. Maintaining an up to date vocabulary is clearly beneficial in the Broadcast News domain, and we have presented a set of techniques for both improving the pronunciation models of words existing in the vocabulary and for learning the pronunciations of novel words. Given appropriate acoustic data, these techniques should allow the construction of an automatically-updated pronunciation lexicon that can track a changing vocabulary without manual intervention.

A system based on these techniques was included in the 1998 Hub-4E evaluations. Its word error rate was relatively some 30% higher than the best-performing systems, much of which may be attributed to the absence of any local adaptation to speaker characteristics and acoustic conditions in the test or training data. Partially, this is because we chose to focus on other areas, but it should be noted that while the generative HMM approach estimates a joint distribution over all variables, the connectionist approach does not estimate the distribution of the acoustic data. A consequence of this is that Gaussian mixture-based HMM systems are very amenable to local adaptation through schemes such as maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995), which cannot be applied to connectionist systems. Current approaches to connectionist speaker adaptation are based on gradient-descent (Neto et al., 1995), although Fritsch et al. (1998) have recently presented a promising method using a somewhat different connectionist acoustic model.

The complete recognizer was easily adapted to run within 10x slower than real-time, and a fully real-time system was possible with only about a 25% relative increase in word errors.

In summary we have outlined a simple, efficient speech recognition system based on a connectionist acoustic model. This single departure from traditional HMM techniques has prompted us to explore several new approaches to the various components of an LVCSR system. Our results suggest that these methodologies are well suited to the challenges presented by the Broadcast News domain.

Acknowledgments

References

Barker, J., G. Williams, and S. Renals (1998). Acoustic confidence measures for segmenting broadcast news. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 2719–2722. 4, 3

Bernardis, G. and H. Bourlard (1998). Improving posterior confidence measures in hybrid HMM/ANN speech recognition systems. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 775–778. 3

Bourlard, H., B. D'hoore, and J.-M. Boite (1994). Optimizing recognition and rejection performance in wordspotting systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 373–376. 6.2

Bourlard, H. and N. Morgan (1994). *Continuous Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers. 2, 2.1

Bourlard, H. and N. Morgan (1996). Hybrid connectionist models for continuous speech recognition. In C.-H. Lee, F. K. Soong, and K. K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition: Advanced Topics*, Chapter 11, pp. 259–283. Kluwer Academic Publishers. 2

Clarkson, P. and R. Rosenfeld (1997). Statistical language modeling using the CMU-Cambridge toolkit. In *Eurospeech-97*, pp. 2707–2710. 7.2

Cook, G., D. Kershaw, J. Christie, and A. Robinson (1997, February). Transcription of broadcast television and radio news: The 1996 Abbot system. In *Proceedings of DARPA Speech Recognition Workshop*, pp. 79–84. Morgan Kaufmann. 5.1

Cook, G. and T. Robinson (1998). Transcribing broadcast news witht the 1997 ABBOT system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 917–920. 3

Cox, S. and R. Rose (1996). Confidence measures for the SwitchBoard database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 511–514. 4

Eide, E., H. Gish, P. Jeanrenaud, and A. Mielke (1995). Understanding and improving speech recognition performance through the use of diagnostic tools. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 221–224. 4

Finke, M. and A. Waibel (1997). Speaking mode dependent pronunciation modeling in large vocabu lary conversational speech recognition. In *Eurospeech-97*, pp. 2379–2382. 5.2

Fiscus, J. (1997). 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*. 7.3

Fosler-Lussier, E. and G. Williams (1999). Not just what, but also when: Guided automatic pronunciation modeling for broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*. 5.2

Fosler-Lussier, J. E. (1999). *Dynamic Pronunciation Models for Automatic Speech Recognition*. Ph. D. thesis, University of California, Berkeley. 5.2

Fritsch, J., M. Finke, and A. Waibel (1998). Effective structural adaptation of LVCSR systems to unseen domains using hierarchical connectionist acoustic models. In *Proceedings of the International Conference on Spoken Language Processing*. 8

Gopalakrishnan, P. S., L. R. Bahl, and R. L. Mercer (1995). A tree-search strategy for large vocabulary continuous speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 572–575. 6

Hain, T., S. Johnson, A. Tuerk, P. Woodland, and S. Young (1998). Segment generation and clustering in the HTK broadcast news transcription system. In *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pp. 133–137. 7

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America 87*, 1738–1752. 2.2, 3

Hermansky, H. and N. Morgan (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing 2*(4), 578–589. 3, 3

Hetherington, L. (1995). New words: Effect on recognition performance and incorporation issues. In *Eurospeech-95*, pp. 1645–1648. 3

Hochberg, M. M., S. J. Renals, A. J. Robinson, and G. D. Cook (1995). Recent improvements to the AB-BOT large vocabulary CSR system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 401–404. 2.3

Jelinek, F. (1969). Fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development 13*, 675–685. 6

Kershaw, D. (1996). *Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System*. Ph. D. thesis, Cambridge University Engineering Department. 7.1

Kingsbury, B. E. D. (1998). *Perceptually-inspired Signal Processing Strategies for Robust Speech Recognition in Reverberant Environments*. Ph. D. thesis, University of California, Berkeley, CA. 3

Kingsbury, B. E. D., N. Morgan, and S. Greenberg (1998). Robust speech recognition using the modulation spectrogram. *Speech Communication 25*, 117–132. 3

Kohlrausch, A., D. Püschel, and H. Alphei (1992). Temporal resolution and modulation analysis in models of the auditory system. In M. E. H. Schouten (Ed.), *The Auditory Processing of Speech: From Sounds to Words*, pp. 85–98. Berlin, Germany: Walter de Gruyter and Co. 4

Leggetter, C. and P. Woodland (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language 9*, 171–186. 8

Lowerre, B. and R. Reddy (1980). The Harpy speech understanding system. In W. A. Lea (Ed.), *Trends in Speech Recognition*. Prentice Hall. 6

Lucassen, J. and R. Mercer (1984). An information theoretic approach to the automatic determination of phonemic baseforms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 42.5.1–4. 5.2

Morgan, N., J. Beck, P. Kohn, J. Bilmes, E. Allman, and J. Beer (1992). The Ring Array Processor (RAP): A multiprocessing peripheral for connectionist applications. *Journal of Parallel and Distributed Computing 14*, 248–259. 2.2

Neti, C., S. Roukos, and E. Eide (1997). Word-based confidence measures as a guide for stack search in speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 883–886. 4

Neto, J., L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and A. Robinson (1995). Speaker adaptation for hybrid HMM-ANN continuous speech recognition system. In *Eurospeech-95*, pp. 2171–2174. 8

Nilsson, N. J. (1971). *Problem Solving Methods of Artificial Intelligence*. New York: McGraw-Hill. 6

Odell, J. J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*. Ph. D. thesis, University of Cambridge. 6

Pallett, D. (2000). The broadcast news task (fix). *Speech Communication*. Submitted to this issue. 1

Paul, D. B. (1992). An efficient A* stack decoder algorithm for continuous speech recognition with a stochastic language model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 25–28. 6

Renals, S. (1996). Phone deactivation pruning in large vocabulary continuous speech recognition. *IEEE Signal Processing Letters 3*, 4–6. 4, 6.2

Renals, S. and M. Hochberg (1995). Efficient search using posterior phone probability estimates. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 596–599. 6

Renals, S. and M. Hochberg (1999). Start-synchronous search for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing 7*, 542–553. 6

Riley, M., W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos (1998). Stochastic pronunciation modelling from hand-labelled phonetic corpora. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, Netherlands, pp. 109–116. 5

Robinson, T. (1994). The application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks 5*, 298–305. 2.1

Robinson, T. and J. Christie (1998). Time-first search for large vocabulary speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 829–832. 6

Robinson, T., M. Hochberg, and S. Renals (1996). The use of recurrent networks in continuous speech recognition. In C.-H. Lee, F. K. Soong, and K. K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition: Advanced Topics*, Chapter 10, pp. 233–258. Kluwer Academic Publishers. 2

Schwartz, R., L. Nguyen, and J. Makhoul (1996). Multiple-pass search strategies. In C.-H. Lee, F. K. Soong, and K. K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition – Advanced Topics*, Chapter 18, pp. 429–456. Kluwer Academic. 6

Siegler, M. A., U. Jain, B. Raj, and R. M. Stern (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *Proceedings of the Speech Recognition Workshop*, pp. 97–99. Morgan Kaufmann. 7

Wawrzynek, J., K. Asanovic, B. Kingsbury, J. Beck, D. Johnson, and N. Morgan (1996). SPERT-II: A vector microprocessor system. *IEEE Computer 29*(3), 79–86. 2.2

Williams, G. and D. Ellis (1999). Speech/music discrimination based on posterior probability features. In *Eurospeech-99*, Volume II, pp. 687–690. 3

Williams, G. and S. Renals (1998). Confidence measures for evaluating pronunciation models. In *ESCA Workshop on Modeling pronunciation variation for automatic speech recognition*, Kerkrade, Netherlands, pp. 151–155. 4

Williams, G. and S. Renals (1999). Confidence measures from local posterior probability estimates. *Computer Speech and Language 13*, 395–411. 4

Wu, S.-L. (1998). *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*. Ph. D. thesis, University of California, Berkeley. 3