Recent Results from the *ARM* Continuous Speech Recognition Project

Martin Russell and Keith Ponting

Speech Research Unit
RSRE, Malvern, Worcs WR14 3PS, UK

Introduction

This paper describes some of the most recent work on continuous speech recognition using phoneme-level hidden Markov models (HMMs) which has been conducted at the UK Speech Research Unit as part of the ARM (Airborne Reconnaissance Mission) project [11]. The goal of the project is automatic recognition of spoken airborne reconnaissance reports. The project draws on many years of research undertaken in the UK by the Joint Speech Research Unit and the current RSRE Speech Research Unit, and also on the work on continuous speech recognition using sub-word HMMs which has been conducted under the current DARPA programme, particularly at MIT Lincoln Laboratory [5] and Carnegie Mellon University [3], and by the speech groups at IBM and BBN.

The project began with the definition, implementation and evaluation of a simple speaker-dependent "baseline" system. This was then systematically improved and assessed in order to measure the performance gain resulting from each enhancement. The most recent version of the speaker-dependent ARM system scores an average 86.8% word accuracy with no syntax on the 497 word vocabulary ARM task. An overview of the development of the speaker- dependent ARM system is presented in [11] and more detailed information about particular stages in the evolution of the system can be found in a set of separate reports [7, 6, 10, 8, 9]. For completeness, the ARM task and the most recent version of the speaker- dependent ARM system are both described in the present paper. The paper goes on to report work in progress in two areas: initial work towards the development of a speaker-independent version of the ARM system, and a study of the performance of versions of the speaker-dependent ARM system from the viewpoint of the number of system parameters.

The development of a speaker-independent version of the ARM system is a current goal of the project. This has necessitated the collection of a new 340 speaker speech corpus which includes recordings of ARM reports for each subject. Orthographic annotation of this corpus is proceeding in parallel with the development of a baseline speaker-independent version of the recognition system through a process of "annotation by forced recognition". At the time of writing over 2000 sentences from 120 speakers have been labelled in this way and systems trained on 10, 20, 30, 40, 50 and 60 male speakers have been evaluated. This work is described in more detail below.

For a fixed size of training set, the number of system parameters is clearly an important consideration in the design of any statistically based speech recognition system. The final section presents a study of some of the systems which have been evaluated as part of the ARM project from the perspective of number of parameters. The results show a range of sizes of parameter set which are large enough to fully exploit the training data in terms of accurate modelling of speech patterns and at the same time small enough to be supported by the training set. Experiments using clustered triphones with state-specific covariance matrices, which were stimulated by these results, are also reported.

The Airborne Reconnaissance Mission Task

Texts of simulated airborne reconnaissance reports were created using an automatic sentence generator based on a finite state syntax (perplexity 6) and 497 word vocabulary, defined by the Royal Aerospace Establishment (RAE), Farnborough UK. A typical ARM report is as follows:

"Inflight report 1-alpha/268. Target map ref foxtrot kilo 9012, correction 2435. Sighting at zero one oh eight zulu. New target defended strip. Less than 13 helicopters, type possibly hip. Runways heading northwest wholly damaged, SAM defences to west intact. TARWI 7/8ths at 2000, end of message"

The start (first four sentences) and end (final sentence) of the report specify a mission reference number, target location, time of sighting, target category and weather conditions respectively and are tightly structured. The remaining central part of the report, which describes what can be seen from the aircraft, is relatively free format.

The Speaker-Dependent ARM System

The development of the speaker-dependent ARM system is described in detail in [11]. This section is concerned with a description of its most recent version (ARM version 7 of [11]).

The acoustic front-end for the ARM system is based on a conventional filterbank analyser with 27 critical band spaced filters covering frequencies up to 10kHz and producing 100 frames per second. The mean channel amplitude of each filterbank frame is subtracted from all components of that frame, and a cosine transform is then applied. The 17 dimensional representation consisting of cosine coefficients 1 to 16 and mean filterbank channel amplitude forms the acoustic front-end parameterisation for the ARM system (see [8]). The frame-rate is reduced by approximately 50% using the variable frame rate technique described in [7, 6].

Acoustic-phonetic processing in the current speaker-dependent version of the system uses a set of approximately 1500 HMMs (the precise number depends on the speaker) consisting of:

- Four single state "non-speech" HMMs to model non-speech sounds in regions of the test data between spoken sentences.
- Six word-level HMMs for the commonly occuring short words "air", "at", "in", "of", "oh" and "or". The number of states in these word-level HMMs is equal to three times the number of phonemes in the baseform transcription of the corresponding word.
- Approximately 1490 three-state HMMs, one for each word-internal triphone [12] which occurs in the ARM vocabulary. Since the baseform pronunciations of ARM vocabulary words vary between speakers in the speaker dependent system, the precise number of triphone HMMs is different for each speaker.

All HMM states are identified with single multivariate Gaussian state output probability density functions with diagonal (co)variance matrices. A single "grand" covariance matrix is shared by all states [4, 9].

Words in the ARM vocabulary are related to phonemes through a dictionary of "baseform" phonemic transcriptions (one transcription per word). In the current, speaker-dependent, system this dictionary is modified for each speaker. The modifications are concerned with broad differences, for example between "northern British English" and "southern British English", rather than with fine details of the speakers pronunciation.

Parameter estimation is based on standard subword HMM training procedures in which sentence level HMMs are constructed from phoneme-level HMMs (using the dictionary of baseform pronounciations). These are then mapped onto the sentence level acoustic data using the forward- backward algorithm to obtain contributions to the new model parameter estimates. Training is done in 3 stages: estimation of the parameters of context-insensitive monophone-HMMs, estimation of the parameters of context-sensitive triphone-HMMs (using the monophone HMM parameters as initial statistics), and estimation of the grand (co)variance matrix.

Performance of the speaker-dependent ARM system

The system was trained and evaluated separately on three speakers. For each speaker, 37 spoken ARM reports (224 sentences, approximately 15 minutes of speech), labelled orthographically at the sentence level, were used to estimate the parameters of the phoneme-level HMMs, and 10 reports (540 words) were used as a test set. Recognition is performed using a one-pass dynamic programming algorithm with beam search and partial-traceback [1]. In experiments conducted in autumn 1989 the system scored an average 86.8% word-accuracy without syntax (93.8% words correct) [11].

The "baseline" speaker independent ARM system

A current goal of the ARM project is to develop a speaker-independent version of the system. This will involve two stages: the creation of a set of speaker-independent triphone HMMs, and the development of adaptation techniques, such as those described in [2], which will enable the parameters of these models to be adapted for new speakers. This section reports on the first of these two stages. As in the speaker-dependent work, this phase of the project has started with the implementation of a simple "baseline" speaker-independent system. This is obtained by training the system described in the previous section using a corpus of ARM reports spoken by a number of speakers. This has necessitated the recording of a new speech corpus which includes recordings of ARM reports spoken by a large number of speakers.

The "Speaker Independent" SIA Speech Corpus

The SIA corpus contains recordings of 340 speakers, each speaking the following material:

- 3 ARM reports
- 6 extracts from ARM reports. These extracts consist of the centre sections of the reports which describe what the "observer" can see from the aircraft. These sections of the reports are less constrained than the initial and final parts, and consequently contain a richer variety of phonemic contexts.
- 10 sentences generated from an air-traffic control application
- 10 "TIMIT like" English sentences.

Only the first two sets of recordings are used in the current phase of the project, the remainder are intended for future work. As with the earlier speaker-dependent database, all recordings were made digitally on video cassette (44kHz sample rate) in a sound proof room using a Shure SM10 head-mounted microphone.

Annotation of the SIA corpus

Although it is possible to estimate triphone HMM parameters using speech labelled at the report level, in practice it desirable that the training material should be labelled at a finer level. In the present experiments annotation is nominally at the sentence level, however segments of speech which are separated by long portions of non-speech are labelled as distinct items. Thus if a subject speaks reports as a sequence of fluent sentences, the data will be labelled at the sentence level, but if a long pause occurs in the middle of a sentence, that sentence will be labelled as two separate segments. Labelling of the speech corpus is proceeding in parallel with the development of the baseline speaker- independent ARM system through a process of "forced recognition". New reports are labelled by the ARM recognition software using the current best speaker-independent models in conjunction with a report-specific syntax which allows non-speech models to occur between words but ensures recognition of the correct word sequence. The results of this automatic labelling process are checked manually and corrected if necessary. Thus, reports spoken by the first ten training speakers were labelled using speaker-dependent triphone HMMs, and reports spoken by subsequent groups of training speakers were labelled using triphone HMMs trained on all previous speakers.

At the time of writing 360 reports from 120 speakers have been labelled in this way and recognition systems trained on 10, 20, 30, 40, 50 and 60 male speakers have been evaluated. For each speaker in the training set, all three ARM reports were used as training material.

Performance of the baseline speaker independent ARM system

Figure 1 shows percentage word accuracy with no syntax as a function of number of training speakers for a set of ten test subjects, none of whom were in the training set. The training and test speakers are all male. It is clear from the figure that there are two modes of performance.

For the eight best speakers, recognition accuracy increases with number of training speakers for training sets with up to 40 speakers, after which it is approximately constant. The average word accuracy for these 8 subjects with models trained on 60 speakers is 59.2%, with individual scores ranging from 38.5% to 76.5%.

For the remaining 2 speakers the performance of the system is badly degraded, with an average word accuracy of -38.6%. No obvious reason for this poor performance is apparent from listening to the recordings, for example the speaking styles of

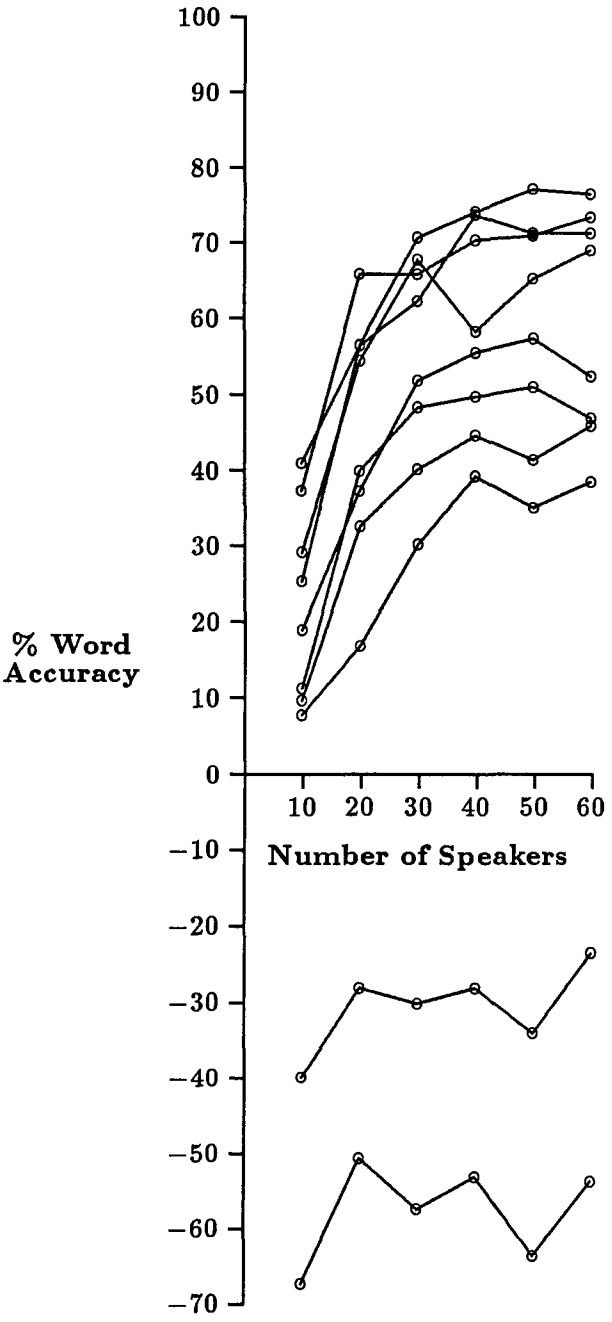

Figure 1: *Speaker-independent word accuracy without syntax as a function of number of training speakers for 10 male test speakers using HMMs trained on male speakers.*

these two speakers are, subjectively, no more atypical than those of the other 8 speakers. Current investigations are concentrating on the possibility that some components of the ARM system are more sensitive to speaker diffences than was anticipated.

Performance as a Function of Size of Parameter Set

During the development of the speaker dependent ARM system several factors were varied which change the number of system parameters. These include the acoustic front-end parameterisation, the number of HMMs and their topologies, and the use of shared or state-specific covariance matrices. Although the effect on performance which results from a particular change is normally attributed to its appropriateness in terms of speech pattern modelling, there will also be effects due to the ability of the training set to support different numbers of parameters. For example, early results showed that for monophone HMMs there was sufficient training material to support state-specific covariance matrices and that the introduction of a shared "grand" covariance matrix resulted in poorer performance [8]. By contrast the introduction of triphone HMMs with state-specific covariance matrices resulted in either a small increase or a significant decrease in performance, because of the large number of system parameters, and large improvements in recognition accuracy were not observed until a shared covariance matrix was used [9]. In terms of number of parameters these two cases represent extremes in the development of the system, but the results suggest that it would be fruitful to look at the performance of a range of versions of the system from the perspective of number of parameters.

Figure 2 shows %word accuracy with no syntax as a function of number of system parameters for 22 speaker- dependent systems which were evaluated as part of the ARM project. To a first approximation small, medium and large numbers of parameters correspond to monophone HMM systems with alternative acoustic front-end parameterisations [8], clustered triphone systems [10], and triphone systems with state-specific covariance matrices [9] respectively. No distinction has been made between means, variances and transition probabilities in the calculation of parameter set size. The figure clearly suggests an underlying effect of parameter set size, with poor performance resulting both from small numbers of parameters, which do not permit sufficiently accurate modelling of the speech patterns, and large numbers of parameters which cannot be supported by the training set. The figure indicates that an acceptable balance between detailed modelling and trainability is achieved with sets of between 20,000 and 100,000 parameters.

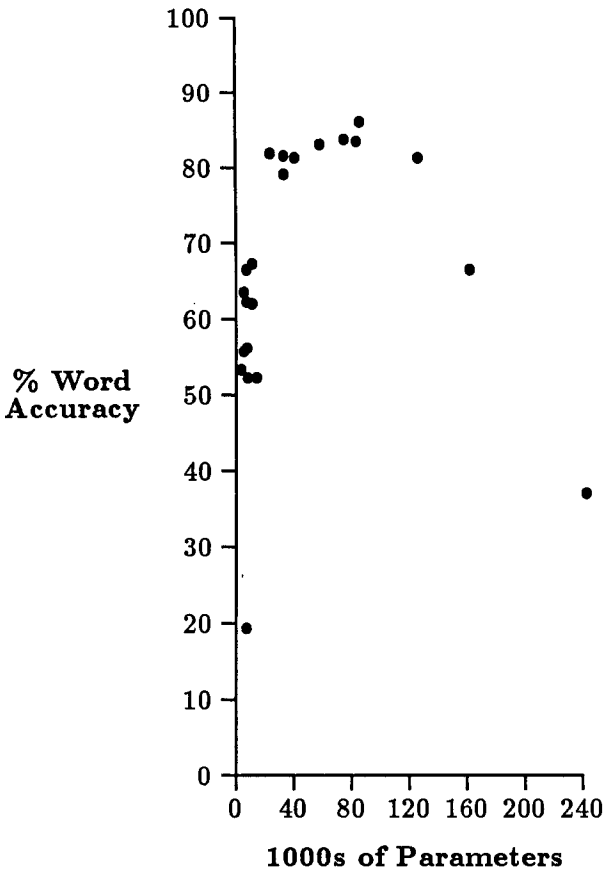

Figure 2: *Performance of various versions of the ARM system for speaker SJ as a function of size of parameter set.*

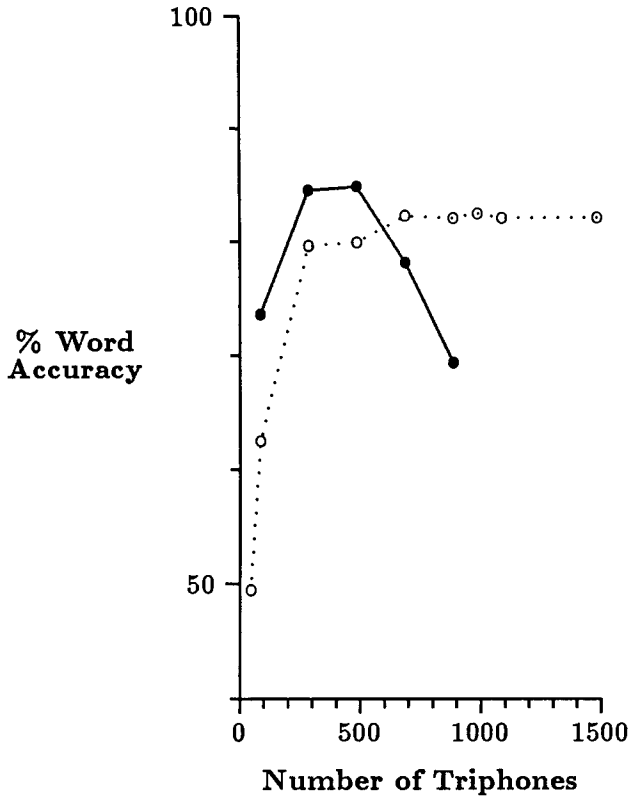

Figure 3: *Word accuracy with no syntax as a function of number of triphones averaged over 3 speakers. Shared (dotted line) and state-specific (solid line) covariance matrices.*

Further Experiments with Clustered Triphones

The results of the previous section suggest a modification to the triphone clustering experiments described in [10], which demonstrate that the number of triphone HMMs in the ARM system can be reduced from 1500 to 300 by clustering with no significant drop in performance. The HMM sets in these experiments have a single shared covariance matrix. If state-specific covariance matrices had been used, the number of parameters for sets of 280 and 480 triphones would have been 47,317 and 79,917 respectively. According to the previous section, these sizes of parameter set can be supported by the training data. Hence one would predict that improved performance would result from the use of state-specific covariance matrices for sets of 280 and 480 triphone HMMs.

The dotted line in figure 3 is taken from [10] and shows %word accuracy as a function of number of triphones for sets of triphones with shared covariance matrices. The solid line shows new results and is the corresponding graph for sets of triphones with state-specific covariance matrices. As predicted by figure 2, the overall best performance is obtained from sets of 280 and 480 triphones with state-specific covariance matrices. The poorer performance obtained with sets of 80, 680 and 880 triphones with state specific covariance matrices is also consistent with the results shown in figure 2.

The fact that the superior performance of the sets of 280 and 480 triphone HMMs with state specific covariance matrices was predicted from figure 2 confirms that an understanding of the size of parameter set which can be supported by a given training set is important in the design of this type of system.

Acknowledgement

The author wishes to acknowledge the contributions to the ARM project which have been made by all members of staff at the Speech Research Unit.

References

[1] J S Bridle, M D Brown and R M Chamberlain, "A one-pass algorithm for connected word

recognition", IEEE-ICASSP, 899-902, 1982.

[2] S J Cox and J S Bridle, "Unsupervised speaker adaptation by probabilistic spectrum fitting",ICASSP 89, Glasgow, Scotland, 1989.

[3] K-F Lee, "Large vocabulary speaker-independent continuous speech recognition: the SPHINX system", PhD Thesis, Carnegie Mellon University, 1988.

[4] D B Paul, "A speaker-stress resistant isolated word recognizer", ICASSP'87, Dallas, TX, 1987.

[5] D B Paul, "The Lincoln robust continuous speech recognizer", ICASSP 89, Glasgow, Scotland, 1989.

[6] S M Peeling and K M Ponting, "Further experiments in variable frame rate analysis for speech recognition", RSRE memorandum 4336, 1989.

[7] K M Ponting and S M Peeling, "Experiments in variable frame rate analysis for speech recognition", RSRE memorandum 4330, 1989.

[8] M J Russell, D Lowe, M D Bedworth and K M Ponting, "Improved Front-End Analysis in the ARM System: Linear Transformations of SRUbank", RSRE memorandum 4358, February 1990.

[9] M J Russell and K M Ponting, "Experiments with Grand Variance in the ARM Continuous Speech Recognition System", RSRE memorandum 4359, February 1990.

[10] M J Russell, K M Ponting, S R Browning, S Downey and P Howell, "Triphone Clustering in the ARM System", RSRE memorandum 4357, February 1990.

[11] M J Russell, K M Ponting, S M Peeling, S R Browning, J S Bridle, R K Moore, I Galiano and P Howell, "The ARM Continuous Speech Recognition System", ICASSP'90, Albuquerque, New Mexico, April 1990.

[12] R M Schwartz, Y L Chow, O A Kimball, S Roucos, M Krasner and J Makhoul, "Context-Dependent Modelling for acoustic-phonetic recognition of continuous speech", ICASSP 85, Tampa, April 1985.