Proceedings of the Institute of Acoustics

THE DEVELOPMENT OF THE SPEAKER INDEPENDENT ARM SPEECH RECOGNITION SYSTEM

Martin J Russell

Speech Research Unit, DRA Malvern, St. Andrews, Great Malvern, England

1. INTRODUCTION

This paper describes the development of a system for automatic recognition of spoken Airborne Reconnaissance Mission (ARM) reports using phoneme-level hidden Markov models (HMMs). The system is speaker-independent, with no explicit speaker enrolment. The results extend previous work on speaker-dependent recognition of spoken ARM reports [5]. The work is presented in detail in [1]

It has been demonstrated in [12] and elsewhere that good speaker-independent performance can be achieved for tasks similar to ARM using phoneme-level HMMs trained on task-specific speech from a large population of speakers. The first stage in the development of the speaker-independent ARM system was to create a multi-speaker corpus of spoken ARM reports [4], and to use this to train the best speaker-dependent ARM system in order to obtain a "baseline" speaker-independent system. This baseline system is described in section 3. The system was developed using speech from male speakers only, therefore in the context of this paper, "speaker-independent" should be interpreted as "male-speaker-independent". The performance of the baseline system (section 4) was not entirely as predicted. For two of the ten speakers in the evaluation set performance was extremely poor and apparently independent of number of training speakers. An investigation of the behaviour of the system for these two speakers led to modifications to the front-end parameterisation. Also, in order to facilitate changes to the front-end parameterisation an altenative variable frame rate (VFR) analysis scheme was adopted. This is described in section 5.

The next two sections report the results of routine enhancements which were made to the baseline system. In section 6 it is shown that the use of a delta-cepstrum front-end representation results in improved speaker-independent performance. Section 7 reports the results of using a word transition penalty. This was prompted by the observation that the errors made by the speaker-independent system were unduly biased towards word insertions. This version of the system (SI-ARM version 5), with VFR analysis applied directly to the SRUbank representation, a delta-cepstrum-based front-end representation and an appropriately chosen word insertion penalty, scores an average word accuracy of 72.5% on the 10 speaker evaluation set.

Section 8 presents a reassessment of the two competing VFR analysis schemes in the context of SI-ARM version 5. There is no significant difference between the performances of the two competing schemes.

At this point it was decided to evaluate the system on a larger, unseen test set consisting of spoken ARM reports from 80 male subjects. The results of this evaluation are presented in section 9. The system scores an average word accuracy of 74.1% with no explicit syntactic constraints.

Some conclusions which have been drawn from this work are presented in section 10.

2. THE "SI89" 321 SPEAKER CORPUS

The "SI89" corpus consists of recordings of speech from 321 subjects (230 male and 91 female). All of the subjects were DRA Malvern staff who responded to a site notice requesting volunteers to participate in the production of the corpus. The recordings were made digitally on video cassette (44.1kHz sample rate) in a sound proof room using a Shure SM10A head-mounted microphone. Details of the recording procedure

and equipment used have been presented elsewhere [7]. Each of the subjects recorded a number of different types of material, including 3 ARM reports. The latter were used in the current experiments. The complete corpus is described in [4].

2.1 The Airborne Reconnaissance Mission Reports

The form of the ARM reports has been described elsewhere [1, 5] but is repeated here for completeness. The reports were created using an automatic sentence generator based on a finite state syntax and 497 word vocabulary. A typical ARM report is as follows:

"Inflight report one dash alpha slash two six eight. Target map ref foxtrot kilo niner zero one two, correction two four three fife. Sighted at zero one oh eight zulu. New target defended strip. Less than thirteen helicopters, type possibly hip. Runways heading northwest wholly damaged, SAM defences to west intact. TARWI seven eighths at two thousand, end of message"

2.2 The Speaker-Independent ARM Pronunciation Dictionary

A "speaker independent" pronunciation dictionary expresses each word in the 500 word ARM vocabulary as a sequence of phoneme-level symbols. For the majority of the words the dictionary contains single baseform phonemic transcriptions. The main exceptions to this rule are the six short words "air", "at", "in", "of", "oh" and "or" which are allocated their own unique word-level symbols. The dictionary also includes two "compound" words: "a few" and "a number" (the words "a", "few" and "number" only occur in these contexts in the ARM application).

3. THE BASELINE SPEAKER-INDEPENDENT ARM SYSTEM

The baseline speaker-independent system was obtained by training version 7 ([5]) of the speaker-dependent system on spoken ARM reports from the "SI89" corpus. The baseline system is described below for completeness.

3.1 Front-end acoustic analysis

Front-end acoustic analysis is derived from the SRUbank filterbank analyser in its default configuration (27 filters spanning the range 0 to 10kHz, 100 frames per second). The mean channel amplitude $m(\vec{v}_t)$ of each frame $\vec{v}_t$ is subtracted from each component of $\vec{v}_t$ and the resulting vector is rotated using a discrete cosine transform to obtain a new feature vector $\vec{w}_t$. A 17 dimensional vector $\vec{x}_t$ is obtained from the first 16 coefficients of $\vec{w}_t$ (excluding coefficient 0) plus $m(\vec{v}_t)$. This is the *CC16* parameterisation from [15]. The sequence $(\vec{x}_t)$ is then compressed using the VFR analysis algorithm described in [11] with threshold 350. This gives a new sequence $(\vec{o}_t)$. For each (VFR) time t, the 18^{th} component $o_t{}^{18}$ of $\vec{o}_t$ is set equal to D_t, the number of SRUbank feature vectors which were replaced by $\vec{o}_t$ during VFR analysis.

3.2 Acoustic-Phonetic Processing

Acoustic-phonetic processing uses a set of 1495 HMMs. The model set consists of (i) 4 single state "non-speech" HMMs to cope with non-speech sounds in regions of the test data between spoken sentences, (ii) 6 word-level HMMs for the commonly occuring short words "air", "at", "in", "of", "oh" and "or" (the number of states in each of these word-level HMMs is equal to three times the number of phonemes in the baseform transcription of the corresponding word), and (iii) a set of 1485 3 state triphone HMMs, one for each word-internal triphone in the ARM vocabulary. All HMM states are identified with single multivariate Gaussian state output probability density functions sharing the same "grand" diagonal (co)variance matrix.

3.3 HMM Training and Recognition

3.3.1 Training and Test Data. The training, evaluation and test sets are disjoint sets comprising 3 ARM reports each from 61, 10 and 80 male subjects respectively from the "SI89" corpus. These sets are specified fully in [1].

3.3.2 HMM Training. Monophone HMMs were obtained using training material labelled orthographically at the sentence level only. Standard sub-word HMM training procedures were used in which sentence level

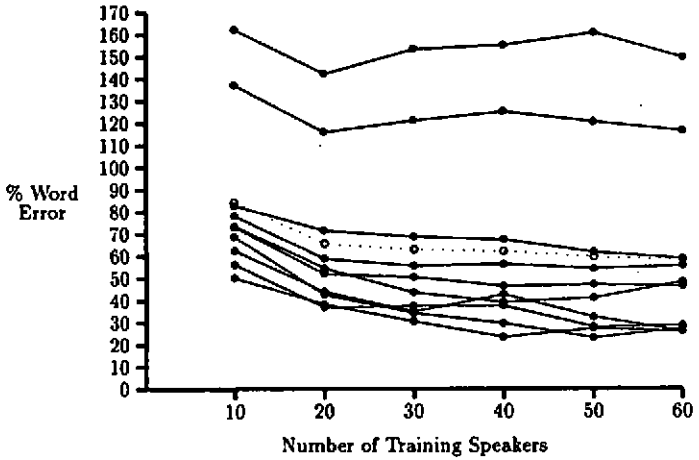

Figure 1: *Performance of the "baseline" speaker independent ARM system as a function of number of training speakers (% word errors without explicit syntax). The solid and dotted lines show the scores for individual speakers and averaged over all speakers respectively.*

HMMs were constructed from phoneme-level HMMs using the dictionary of baseform transcriptions. These models were then mapped onto the sentence level acoustic data and contributions to the model parameter estimates computed. For the initial iteration this mapping was linear, but for subsequent iterations the standard "forward-backward" algorithm was used.

The monophone HMMs provided initial estimates for the parameters of the triphone HMMs. These were then optimised with respect to the training set using standard sub-word HMM training procedures. Three further iterations of the training algorithm were used to estimate a "grand" covariance matrix [8].

3.3.3 Recognition. This comprised a one-pass dynamic programming algorithm with beam search and partial traceback [3]. Results are presented in terms of % words wrong and % word errors, computed as follows, using dynamic programming to align the true transcription of the test data with the output of the recogniser:

$$\text{words wrong} = \frac{S+D}{N} \times 100, \text{ word errors} = \frac{S+D+I}{N} \times 100 \quad (1)$$

where N is the number of words in the test set, and S, D and I are the number of words substituted, deleted and inserted respectively.

4. PERFORMANCE OF THE "BASELINE" SYSTEM

Figure 1 shows % word error with no explicit syntax as a function of number of training subjects for each of the 10 speakers in the evaluation set. It is clear from the figure that there are two modes of performance. For the 8 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 error for these 8 subjects with models trained on 61 speakers is 39.5%, with individual scores ranging from 58.7% to 25.8%. For the remaining two speakers the performance of the system is badly degraded, with an average word error of 132.8% . Furthermore, for these speakers there is no clear correlation between number of training speakers and performance. Following standard terminology, these two speakers will be referred to as "goats"

This system, trained on 61 male speakers, will be refered to as *SI-ARM* version 1 and scores an average of 58.1% word error on the evaluation set with no explicit syntactic constraints.

5. "SHEEPING THE GOATS" - IMPROVING PERFORMANCE FOR SPECIFIC SUBJECTS

The investigation into the poor performance of the system for two of the subjects focussed on two components of the system: the VFR analysis procedure and the cosine transform. The former was motivated by the fact that the parameters of the VFR were chosen as a result of speaker-dependent experiments [11], and the latter by the concern that higher cosine coefficients might correspond to speaker-specific properties of the speech signal.

5.1 Effect of Variable Frame-Rate Analysis

An experiment was conducted to re-assess the effect of variable frame rate analysis in the context of the speaker-independent ARM system. The experiment showed that the optimal values of the VFR threshold are similar to those for the speaker-dependent system [5]. The best performance, 57.5% word errors, is obtained with a threshold of 450, but this is not significantly better than the figure of 58.1% word errors obtained with the original VFR threshold of 350, and the performance is worse with the lower threshold of 250, for which fewer acoustic vectors are discarded during the VFR process (see [1]). Hence the poor performance of the baseline speaker independent system is not due to VFR analysis.

5.2 Modifications to the Variable Frame-Rate Analysis Procedure

An important difference between the front-end processing scheme described above and that used in the most recent version of the speaker-dependent ARM system is that in the latter system VFR analysis is applied immediately after the filterbank analysis and before the cosine transform [11]. This improves recognition accuracy in the speaker dependent system [11] and it allows one to fix the metric and threshold in the VFR analysis algorithm for the SRUbank parametrisation and to ignore possible interactions between subsequent transformations and VFR analysis. An experiment was therefore conducted to investigate the effect of applying VFR analysis immediatly after SRUbank analysis. The experiment uses the same training and evaluation sets as in the previous section. The new scheme, applied with an optimal threshold of 1100, results in an average word error of 61.4% and a reduction in the number of frames to 35.6% of the original. However, the best performance obtainable with the original scheme is 57.5% average word errors with 37.4% of the original data (with a VFR threshold of 450). The results are presented in full in [1]. Although the original scheme performs best, it was decided to adopt the new VFR scheme during the development of the system because of its convenience, and to repeat the comparison of the two VFR schemes for the final version of the system. The new VFR scheme, with a threshold of 1100, was used in all subsequent experiments.

This version of the system, with VFR analysis applied after filterbank analysis, is SI-ARM version 2.

5.3 Effect of the Cosine Transform

Inspection of the average values of cosine coefficients 1 to 16 after VFR analysis reveals that there is some separation between the values for the two "goats" and those for the remaining speakers for some of the higher cosine coefficients [1]. Although these differences are small, they may still lead to relatively large differences in probability because the grand variances for high cosine coefficients will also be small [8].

Therefore the contributions to the observation probabilities due to individual cosine coefficients were measured. Acoustic patterns corresponding to spoken ARM reports from each of the test speakers were aligned with the correct sequence of HMMs using the Viterbi algorithm. If $\vec{o} = \vec{o_1}, ..., \vec{o_T}$ denotes the sequence of feature vectors corresponding to a particular utterance, and $\sigma = \sigma_1, ..., \sigma_T$ is the corresponding optimal state sequence, then the contribution $log(P_t(\vec{o}, \sigma))$ to the joint log probability of $\vec{o}$ and σ for (VFR) time t is given by:

$$log(P_t(\vec{o}, \sigma)) = -\sum_{d=1}^{16} \frac{(m_i^d - o_t^d)^2}{(v_i^d)^2} + constant \qquad (2)$$

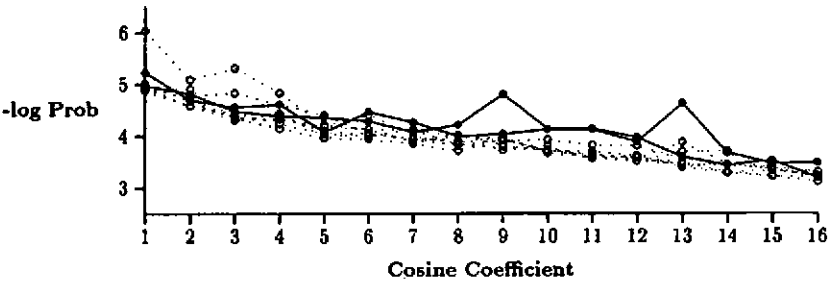

Figure 2: *Average cosine coefficient channel -log probabilities over a single ARM report for the two "goats" (solid lines) and the remaining speakers in the evaluation set (dotted lines)*

$$= \sum_{d=1}^{18} log(P_t^d(\vec{o}, \sigma)) + constant \quad (3)$$

where $\vec{m_t}$ and $\vec{v_t}$ are the mean and variance vectors associated with state σ_t and $P_t^d(\vec{o}, \sigma)$ is the joint probability of the d^{th} cosine coefficient in $\vec{o_t}$ and state σ_t.

The contributions $P_t^d(\vec{o}, \sigma)$ due to the individual cosine coefficients are independent because of the assumption that the covariance matrix is diagonal. Figure 2 shows average values of $-log(P_t^d(\vec{o}, \sigma))$ for $d = 1, ..., 16$ computed over a single *ARM* report for each of the ten evaluation speakers. The figure shows large differences between the graphs for the two "goats" and the graphs for the other speakers for cosine coefficients 9 and 13 and also some of the other higher cosine coefficients. This suggests that these coefficients are particularly sensitive to speaker-dependent factors which distinguish the two "goats" from the other evaluation speakers. For example, the peak in the average value of $-log(P_t^9(\vec{o}, \sigma))$ for one of the "goats" suggests that there is periodic structure, with period 6 channels, in the filterbank-analyser output frames for that speaker, and that this structure is characteristic of the speaker. Since the number of channels in the SRUbank representation is 27, this means that one would expect to see four equally spaced peaks in the spectrum. Observation shows that this type of structure does indeed occur in the SRUbank data for this speaker, in particular in regions of the data which correspond to the "shwa" vowel. We believe that this factor, plus the fact that this speaker displays a tendency to centralise vowels, acounts for the poor performance.

5.4 Effect of Reducing the Number of Cosine Coefficients

As a consequence of this work, the number of cosine coefficients in the acoustic front-end parameterisation was reduced from 16 to 8. Hence, since the mean channel amplitude and variable frame-rate analysis frame-count parameters were retained, the dimensionality of the new front-end is 10. The resulting system is SI-ARM version 3, and scores an average word error of 51.0% (25.1% words wrong) on the evaluation set (no syntax). The use of this lower-dimensional representation leads to a substantial improvement in the performance for the two "goats", as predicted. However, the average word error for the remaining 8 speakers increases from 40.8% with the 16 cosine coefficient front-end to 44.3%. Although the performance for the "goats" is improved, it is still worse than that for any of the other speakers [1].

6. DELTA CEPSTRUM

Previous experiments [15] showed that with a front-end parameterisation based on 8 cosine coefficients, performance was improved by including time-difference, or "delta cepstrum" information. This is the *CC8δ* front-end from [15]. There is also other evidence that the use of the delta cepstrum offers advantages in speaker-independent recognition [12]. Hence a delta-cepstrum was added to the front-end described in the

previous section. This gives a 20 dimensional parameterisation where the feature vector o_t at VFR time t is defined by: $o_t^d = w_t^d$, $d = 1,...,8$, $o_t^9 = m(\bar{v}_t)$, $o_t^{10} = D_t$, and $o_t^d = o_{t+\delta}^{d-10} - o_{t-\delta}^{d-10}$, $d = 11,...20$.

The average % word error on the evaluation set falls from 51.0% without delta cepstrum to 36.1% with delta cepstrum, confirming the value of the delta cepstrum in a speaker-independent system. The results are shown in more detail in [1]. This version of the system is refered to as *SI-ARM* version 4.

7. WORD TRANSITION PENALTIES

The patterns of errors in the systems described above are biased towards word insertions. For example, the average word error of 36.1% for SI-ARM version 4 can be broken down into substitution, deletion and insertion rates of 7.2%, 10.2% and 18.8% respectively. The standard solution to this problem is to use a "word transition penalty" [14, 10]. This is normally a fixed, system-wide, "word transition probability" by which state sequence probabilities are multiplied whenever a transition into a new word occurs. It is usual to refer to a word transition penalty because log arithmetic is used in the recognition algorithm. The penalty is then the negative logarithm of the word transition probability.

The use of a word transition penalty leads to a substantial improvement in recognition accuracy. For example, with a word transition penalty of 30, the average % word error and % words wrong is 27.5% and 18.2% respectively, compared with 36.1% and 17.4% with no word transition penalty. It was found that the precise value of the word transition penalty is not critical [1]. Based on this result, a word transition penalty of 30 is used in SI-ARM version 5. The improvement in performance resulting from the use of a word transition penalty in the current speaker-independent experiments is much greater than that observed in the speaker-dependent experiments reported in [10].

8. FINAL COMPARISON OF ALTERNATIVE VFR SCHEMES

The altenative VFR scheme, which was dropped in section 5.2, was re-evaluated for SI-ARM version 5. In this scheme, VFR analysis is applied after the *CCδ* transform rather than before. The best word error rate obtained with this VFR scheme is 27.3%, which is not significantly different from the score (27.5%) for VFR analysis applied directly to the SRUbank representation [1]. Hence it was decided to retain the latter scheme. It was noted that the performance of SI-ARM vesion 5 is much less sensitive to VFR threshold than early versions [1].

9. FINAL EVALUATION OF SI-ARM

The final version of the speaker-independent system (SI-ARM version 5) has the following characteristics: Initial front-end analysis uses the SRUbank filterbank analyser in its default configuration (27 critical-band spaced filters spanning frequencies up to 10 kHz, 100 frames per second). Each SRUbank vector is amplitude normalised, and the mean channel amplitude is stored as an additional 28^{th} channel. VFR analysis is applied directly to the SRUbank output with a VFR threshold of 1100. Secondary front-end analysis uses a cosine transform to rotate the SRUbank data after VFR analysis. The final front-end acoustic vector at time t is a 20 dimensional delta-cepstrum comprising: cosine coefficients 1 to 8 at time t, the mean SRUbank channel amplitude at time t, the VFR count at time t, and the differences between the previous 10 parameters at times $t+1$ and $t-1$. Acoustic-phonetic modelling is based on a set of 1495 HMMs comprising 4 single state "non-speech" models, 6 "word-level" models of short common words and 1485 triphone models. Acoustic-phonetic decoding uses the "one-pass" dynamic programming based decoding algorithm with a word insertion penalty of 30.

The final evaluation of this system uses the unseen test set of recordings from 80 male speakers. The system scores 25.9% word errors (15.9% words wrong), corresponding to substitution, deletion and insertion rates

of 4.7%, 11.2% and 10.0% respectively (see table 1).

	Percentage score	Number of words
Words correct	84.1%	10,903
Word accuracy	74.1%	
Words wrong	15.9%	2,062
Word errors	25.9%	3,355
Mismatch	4.7%	613
Deleted	11.2%	1,449
Inserted	10.0%	1,293

Table 1: *Performance of the final version the speaker-independent ARM system on the 80 male speaker test set (12,965 words).*

10. CONCLUSIONS

A number of interesting conclusions can be drawn from these experiments.

The level of performance reported in this paper has been achieved with a system which is basically very simple. In particular the state output pdfs associated with the HMM states are single multivariate Gaussian pdfs with diagonal covariance matrices. Results from other laboratories suggest that this result could be improved by replacing these simple pdfs with multiple component Gaussian mixture densities.

Comparison of the final versions of the speaker-dependent and -independent ARM systems shows that many of the empirically derived parameters are similar in both systems. A significant exception to this rule is that the front-end parameterisation based on the first 16 cosine coefficients, which is used successfully in the speaker-dependent system, includes coefficients which are sensitive to speaker specific factors and hence leads to poor results for particular speakers in the speaker independent system. A further difference is that the use of the delta-cepstrum, which did not result in significant improvements in recognition accuracy in the speaker-dependent system, does so in the speaker-independent system.

Two alternative VFR schemes have been considered. However, in terms of word accuracy, both the benefits of VFR and any significant differences between the two schemes diminish as the basic performance of the system increases. The results suggest that in more sophisticated systems the main benefit of VFR analysis is likely to be reduced computation.

Finally, the average word error in SI-ARM version 5 (27.5%) is approximately 50% of the word error achieved by the baseline system (58.1%). However, the main contribution to this improvement is a large reduction in word error for just two of the speakers in the evaluation set. For these two speakers the average word error falls from 132.8% (version 1) to 32.2% (version 5), a reduction to less that 25% of the original word error rate. Thus the improvement in performance is not uniform over all speakers in the evaluation set, but is concentrated on a relatively small subset.

References

[1] M J RUSSELL, "The development of the speaker independent ARM continuous speech recognition system", RSRE memorandum 4473, January 1992.

[2] "SCRIBE - Spoken Corpus Recordings In British English : Text of Speech Material" SCRIBE Document SCRIBE-23, *Available from the Speech Research Unit, RSRE, Malvern.*

[3] J S BRIDLE, M D BROWN & R M CHAMBERLAIN, "A one-pass algorithm for connected word recognition", IEEE-ICASSP, 899-902, 1982.

[4] S R BROWNING, J MCQUILLAN, M J RUSSELL & M J TOMLINSON, "Texts of material recorded in the SI89 speech corpus", SP4 Research Note number 142, RSRE, February 1991.

[5] M J RUSSELL, K M PONTING, S M PEELING, S R BROWNING, J S BRIDLE & R K MOORE, "The *ARM* Continuous Speech Recognition System", Proc. ICASSP'90, Albuquerque, New Mexico, April 1990.

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

[7] M J RUSSELL, R K MOORE, M J TOMLINSON & J C A DEACON, "RSRE Speech Database Recordings 1983 : Part II Recordings made for Automatic Speech Recognition Assessment and Research", RSRE Report No. 84008, May 1984.

[8] M J RUSSELL & K M PONTING, "Experiments with Grand Variance in the *ARM* Continuous Speech Recognition System", RSRE Memorandum Number 4359, 1990.

[9] M J RUSSELL, K M PONTING, S R BROWNING, S DOWNEY & P HOWELL, "Triphone Clustering in the *ARM* System", RSRE memorandum 4357, February 1990.

[10] K M PONTING & S M PEELING, "Word transition penalties in the *ARM* continuous speech recognition system", RSRE memorandum 4362, 1990.

[11] S M PEELING & K M PONTING, "Variable frame-rate analysis in the ARM continuous speech recognition system", Speech Communication 10, pp 155-162, 1991.

[12] K-F LEE, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX System", PhD Thesis, Carnegie Mellon University, 1988.

[13] D B PAUL, "Speaker-stress resistant continuous speech recognition", Proc ICASSP'88, New York, 1988.

[14] D B PAUL, "The Lincoln Robust Continuous Speech Recogniser", Proc ICASSP'89, Glasgow, Scotland 1989.

[15] M J RUSSELL, D LOWE, M D BEDWORTH & K M PONTING, "Improved Front-End Analysis in the *ARM* System: Linear Transformations of SRUbank", RSRE Memorandum Number 4358, 1990.

[16] J WELLS et al., "Specification of SAM Phonetic Alphabet (SAMPA)", included in: P WINSKI, W J BARRY & A FOURCIN (Eds), "Support Available from SAM Project for other ESPRIT Speech and Language Work", The SAM Project, Department of Phonetics, University College London.