Proceedings of the Institute of Acoustics

SOME EXPERIMENTS IN PHONEME-BASED CONTINUOUS SPEECH RECOGNITION USING THE MESSIAH DATABASE

A P Breen[1], *S Kapadia*[1], *S J Whittaker*[1], *S J Young*[2]

[1]British Telecommunications Research Laboratories, Martlesham Heath, Ipswich, Suffolk IP5 7RE

[2]Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ

ABSTRACT

This paper describes a set of speaker dependent recognition experiments undertaken at British Telecom Research laboratories. The aim was to evaluate the usefulness of phoneme-based recognition for continuous speech understanding and to investigate the factors which influence performance. The methods used were vocabulary independent and the phoneme models were constructed from 3-state left-right Hidden Markov Models. Training was performed using hand labelled phrases from the 'Messiah' database of phonetically rich sentences. Performance was evaluated by measuring both unconstrained phoneme recognition accuracy on a separate segment of the phonetically rich sentences and also by measuring word and semantic accuracy on a speech understanding task. Results will be presented for one, two and three modes per mixture continuous density models, context dependent broad class models and clustered context dependent models.

1 INTRODUCTION

An important factor in the development of advanced interactive speech systems is the integration of high quality continuous speech recognition into Spoken Language Systems (SLS). The techniques described in this paper have been developed to combine pattern matching and parsing techniques in the interpretation of the semantically complex responses which can be expected from users of such systems. In any complex service provided to naive users, it is likely that a proportion of responses will be 'poorly formed' even if expressing valid semantics, either in terms of deviation from predicted syntax or via the insertion of non-vocabulary items and non-speech sounds. The recognition and linguistic processing elements of an SLS should therefore cope as a matter of course with such deviations by providing confidence ranked alternative semantic explanations for responses, perhaps using information about dialogue context to limit the scope of recognition to predicted responses.

An off-line speech understanding testbed is described in which a token-passing formulation [8] is used to integrate frame synchronous DTW and HMM pattern matching techniques with context-free grammars for *a priori* match constraint and *a posteriori* chart parsing. The Messiah phonetically rich database is described and phoneme based recognition performance discussed. Comparison is made between the performance of a phoneme based speech understanding system and a wholeword DTW version.

2 SYSTEM CONFIGURATION

2.1 TOKEN PASSING

Young, Russell and Thornton [8] discuss Token Passing as a method for the implementation of recognition systems in which high level processing and the imposition of match constraints are independent of the pattern matching scheme used. Each speech unit is implemented as a *'word model'* (potentially a sub-word) which handles all intra-model score propagation. In a connected word system *word instances* are produced as instantiations of word models at the correct points in a constraint grammar. Inter-word propagation is controlled by the passing of tokens and associated cumulative scores into word models and the flow out from them. *A priori* match constraints are straightforwardly implemented as the assimilation of scores from all word models emitting valid tokens at each node in the

grammar at each time frame and the propagation of the best score (most probable token) into valid successors to that node. *Word link records* store the traceback and score information required for later processing.

The similarities between connected recognition algorithms, for example one-pass and level building have been thoroughly explored [2] and it can be shown that the use of a token passing formalism allows them to be viewed simply as topological variants. More sophisticated context-free constraints may be straightforwardly implemented.

The token passing methodology allows linguistic processing to be independent of the choice of speech unit. Evaluation of the testbed was undertaken using wholeword DTW templates and phoneme-based HMMs

2.2 TESTBED ARCHITECTURE

Within the ALVEY VODIS project [1,7] a speech understanding architecture was developed (see figure 1). At each state of the dialogue a prediction of user responses is passed to the *Context-driven Connected-speech Recognition* (C^2R) subsystem by the *Dialogue Controller*. A match constraint grammar is generated from the rulebase and frame-synchronous matching proceeds. The lattice generated is then transformed to form a chart which is parsed using a second set of rules to provide an explanation of the semantics of the phrase. The results of the parse are expressed in terms of a *frame* whose hierarchy and terminal *slots* are stripped to leave only information bearing elements. This frame is used by the dialogue controller to determine the next stage of the interaction. Previous work using whole-word data [8] investigated the performance of the speech understanding system on syntactically aberrant sentences. Improvements in performance over the single strong grammar case was observed where a weak match and strong parse grammar were specified.

The C^2R subsystem used for the experiments described provides enhancements over that used within the VODIS project. A portable software based system has been developed which is not dependent on any particular pattern matching hardware. Provision is made for the inclusion of wildcards implicitly within word models or explicitly via placement in weak grammars. Wildcarding is a simple fixed penalty scheme which allows thresholding over regions of poor match quality given certain durational constraints. Wildcards may be used to allow propagation through a poorly matching word model to encourage continuity through a connected algorithm. Alternatively, they may explain non-vocabulary items given a weak grammar, allowing the identification of segments of syntactically valid speech.

During parsing, partial spans of the chart are penalised by the application of a parser threshold penalty to unexplained segments. This allows the parsing of incomplete edges. The balance of wildcard and parser penalties has a strong influence on the composition of the chart and its explanation. The selection of wildcard values is closely related to the distribution of score densities for vocabulary items, being a tradeoff between a tendency of the wildcard to misfire and the misrecognition of incorrect vocabulary items. Within an HMM scheme, *sink models* [6] offer the possibility of a more formally based approach.

Sink models are models which are trained on a wide variety of speech and non-speech data rather than on data corresponding to a single lexical item. They have been used for matching isolated words in extraneous speech, keyword spotting and the location of non-vocabulary words [5,6]. In these experiments, sink models are used for the explanation of all non-vocabulary items within a connected word recognition scheme.

3 TRAINING AND TEST DATASETS

3.1 TEST PROBLEM

The test set comprises 40 sentences obtained from transcripts of interactions with British Rail enquiry services. The problem is divided into two subsets, a syntactic set (SYN) which fulfils a broad parse grammar and a non-syntactic set (NON_SYN) within which 40% of vocabulary items are unknown. For example,

CONTINUOUS SPEECH RECOGNITION USING THE MESSIAH DATABASE

Syntactic:

Yes, from Dundee to Edinburgh this afternoon about 3 o'clock

Non-syntactic:

Er, like the times of trains from London to York from about five o'clock onwards if that's possible

3.2 WHOLEWORD TELEPHONY DATABASE

The telephony dataset was obtained over a local dialup line using a BT Speech card system for prompt control and digitisation [3]. The training set consisted of a single set of isolated utterances from a single speaker, manually endpointed. 72 word and phrase templates were created covering the 112 vocabulary words represented in the test grammars. The test set comprised the 40 test sentences collected over the same line during the same session.

3.3 MESSIAH SUBWORD DATABASE

This database was constructed at BTRL to aid research in subword recognition. It contains ten speakers (five males and five females) and was recorded digitally at a sampling rate of 20kHz in a silence cabinet using a high quality head set microphone. The database contains a number of different types of speech data, the largest component being a set of 239 phonetically rich phrases.

Phrases from one male speaker constituted the training and test sets used in the speaker dependent recognition experiments to be described in the next section.

In designing the phrase set, the following criterion was applied; *at least one instance of most the legally occurring diphones in the RP dialect of British English should be contained in as few phrases as possible.* In addition the following *caveats* were borne in mind.

a) The phrases should be of moderate length and not overly complex.

b) The phrases should be semantically plausible and not contain words with ambiguous or unfamiliar pronunciations.

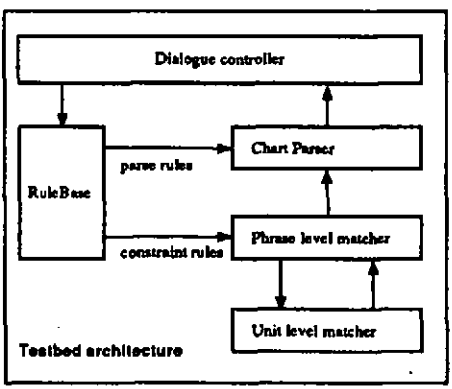

Figure 1: Testbed Architecture

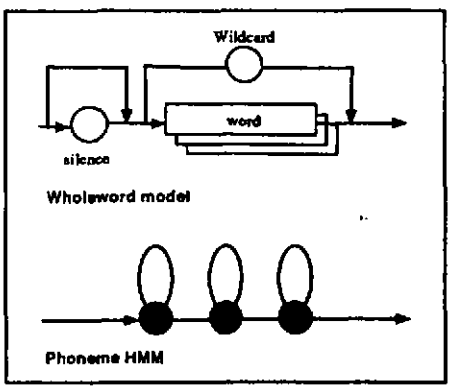

Figure 2: "Word Model" Structures

The diphone was chosen as the most practically useful basic unit of analysis, in the sense that a complete set of diphones is practically realisable and guarantees phonetic richness. The diphone represents a compromise between incorporating context and maintaining a manageable number of units.

Once the phrases were constructed the following labelling strategy was adopted. An assumed broad phonetic transcription was constructed for each phrase in the set using SAM-PA machine readable symbols. Each phrase was analysed in terms of its diphone content and the syllabic and intersyllabic phoneme clusters present. Global statistics for phoneme and diphone occurrence were obtained.

The speech data was annotated such that, for each phrase for every speaker, there existed an orthographic label and a non-time aligned transcription. The non-time aligned transcriptions contained extra symbols in addition to the SAM-PA symbols. These extra symbols were used to label non-speech sounds such as lip smacks, coughs, extraneous noise and silence. In addition to the non-time aligned transcriptions, 25% of the data for two speakers (one male and one female) was hand annotated down to the phonemic level, again using SAM-PA symbols.

4 VOCABULARY INDEPENDENT TRAINING

This section briefly describes a series of speaker dependent recognition experiments to examine the performance of different subword units on the vocabulary independent Messiah database. Results are presented for three different types of subword unit on data downsampled to 8 KHz:

a) Context independent (phoneme) units (CI). Results are presented for models with 1,2 and 3 modes per mixtures for each state.

b) Context dependent 'broad class' units (CBC).

c) Acoustically clustered context dependent units.

The same basic HM model topology was used in all the experiments. The model topology was as follows: a five state (two non-emitting entry and exit states), left - right, no-skip, continuous density HM model. All the experiments were conducted using the same set of features, in the form of a 17 element feature vector.

4.1 TRAINING

The training procedure consisted of two stages. The first stage had two phases. The first phase was as follows:

a) Uniformly segment each hand labelled phoneme from the database for seeding individual model states.

b) Cluster data if more than one mode per mixture is specified.

c) Re-estimate means and variances for each state and each model.

d) Viterbi align to new models, re-segment data.

Steps c) and d) was repeated until convergence or until a fixed maximum number of iterations had been reached. In the second phase, these models were further optimised using Baum-Welch (BW) re-estimation on the hand labelled seed data. The re-estimation was halted once the log probability failed to increase above some small factor ε.

The second stage of training used embedded re-estimation to still further improve the HM models:

a) Use the BW iteration in embedded re-estimation on all training data using the non-time aligned transcriptions.

b) Repeat the process for N iterations (typically N was either three and four).

4.2 RECOGNITION

Recognition was performed as follows: A test phrase was compared against a syntax network of subword models. The decoded sequence of sub-word models expressed by the syntax network was chosen to be the state sequence

which exhibited the highest likelihood (Viterbi recognition). The results of the Viterbi decoding were tested by DP aligning the recognised symbols with non-time aligned transcriptions of the test phrases. In the following results of recognition experiments are given as

%correct = Correctly aligned phonemes/Total number of phonemes

%accuracy = (Correctly aligned phonemes - Insertions)/Total number of phonemes

Other information such as hits (correctly aligned phonemes), deletions, substitutions and insertions are also given. The experiments were conducted with an equal experimentally derived inter-model probability.

4.3 EXPERIMENTAL RESULTS FOR CONTEXT INDEPENDENT UNITS

Table 4.1 given below tabulated the results of three different recognition experiments conducted with context independent units. The experiments differed solely in the number Gaussian PDFs used. The results demonstrate that recognition performance increases with the number of Gaussian modes per mixture in each state.

No. Modes	%Correct	%Accuracy.	Hits	Deletion	Substitution	Insertion
1	55.4	44.4	496	39	360	70
2	59.8	49.7	535	51	309	90
3	61.6	50.2	551	49	295	102

Table 4.1 Context Independent Phoneme Recognition

4.4 EXPERIMENTAL RESULTS FOR CONTEXT DEPENDENT BROAD CLASS (CBC) UNITS

This experiment attempted to introduce a restricted amount of context dependency into the sub-word models. The restrictions were imposed to reduce the number of required units to manageable levels. The broad class context dependent units were designed to accommodate the left and right contexts of each phoneme, where the contextual information was defined as the phonological 'manner' of articulation. The five broad classes were *Plosive, Fricative, Nasal, Glide and Vowel*, resulting in a total of 1125 CBC models. Two methods of training were adopted:

a) Using the iterative training method described in section 4.1 to produce CBC models with a single Gaussian PDF per state.

b) Models were trained as for a) and then hand interpolated with context independent units.

The procedure for b) was as follows. Each M-Gaussian CBC model was combined with its corresponding single Gaussian CI model, such that the resulting model was a CBC with M+1 Gaussian modes per mixture. The modal weights were calculated from the respective occurrences of the contexts. Table 4.2 tabulates the results of recognition experiments conducted in this way.

No. Modes	%Correct	%Accuracy	Hits	Deletion	Substitution	Insertion
case (a)	59.9	46.8	536	42	317	157
2	60.3	47.6	520	31	311	110
3	61.1	47.5	527	35	300	116
4	62.1	49.5	536	36	290	109

Table 4.2 Context Dependent Phoneme Recognition

4.5 ACOUSTICAL CLUSTERING EXPERIMENTS

In this experiment context dependent triphone models [4] were acoustically clustered. Contexts not explicitly present in the training data were predicted by clustering similar contexts. The clustering strategy adopted a tree struc-

ture which ensured that at any node in the tree above the leaf nodes maximally descriminable clusters were produced. Table 4.3 shows the results of experiments conducted with increasing numbers of clustered models.

No. Models.	%Correct	%Accuracy	Hits.	Deletion	Substitution	Insertion
73	60.4	43.9	530	34	314	145
123	62.9	46.0	552	32	294	148
148	63.1	45.9	554	28	296	151
198	64.4	47.3	565	24	289	150

Table 4.3 Phoneme Recognition with Acoustical Clustering

4.6 CONCLUSIONS

Table 4.4 summarises the results given above, showing the top three best results from the above experiments. The results show that the best performance on the test data was produced using 3 modes per mixture context independent units. It is likely that the context dependent models performed less well due to under-training, although this effect is decreased when acoustical clustering is employed, at the price of an slightly increased insertion rate.

Type of unit	No. of units	%Correct	%Accuracy.
CI (3 Mix.)	48	61.6	50.2
CI+CBC	1125	62.1	49.5
Acoustical Clusters	198	64.4	47.3

Table 4.4 Summary of the best three scores (ranked in order of highest accuracy).

5 SPEECH UNDERSTANDING

Testbed performance was evaluated using the Context-Independent (2 modes per mixture) HMM's described above and compared with benchmark wholeword DTW performance. The two cases are not directly comparable as wholeword performance was investigated for network speech, as opposed to the clean Messiah data of the subword case. Two basic context-free grammars were used: a strong grammar for both matching and parsing, and a weak phrase level grammar imposing *function_word-content_word* ordering but no high level structure. The particular detail of the weak match grammars, used solely for matching, varied slightly from case to case with a wildcard or sink model being placed optionally between phrases.

Performance figures quoted are frame accuracies. This is obtained by dynamic programming of the resultant parse frame and a parse obtained from reference orthography. Both frames are automatically stripped by the parser prior to parsing such that the remaining slots correspond roughly to items of information.

5.1 WHOLEWORD MATCHING

Results are shown for two grammar combinations:

a) Full : Strong match, strong parse

b) Weak: Weak match, strong parse

The weak match grammar includes explicit phrase level wildcards and use of implicit silence modelling.

	Full	Weak
SYN	95.8%±2.5	92.6%±2.7
NON-SYN	50.9%±4.7	68.9%±4.5

Table 5.1 Wholeword Testbed Performance

The results demonstrate the ability of the system to use weak *a priori* match constraints to raise performance on non-syntactic data at the cost of a minor degradation in performance in the syntactic case. This is despite the poor channel quality and the simplicity of the training scheme used. Performance was observed to be strongly dependent on the selection of wildcard penalties and parser thresholds (peak performance is given).

5.2 SUBWORD MATCHING

Experiments were undertaken using explicit placement of wildcards, and using the following sink models, all of which were 8-state with 1- and 2-state skips allowed.

- SPC8 trained on speech delimited by silence
- WRD trained on words from orthography
- NSP trained on silence/non speech

Results are shown for weak grammar combinations.

a) Weak_wild : explicit placement of phrase level wildcards and looped silence/breath noise models.

b) Weak_SPC8 : explicit placement of SPC8 model and NSP.

c) Weak_WRD : explicit placement of WRD model and NSP. _a and _b options denote topological variants.

	Full	Weak_wild	weak_SPC8	weak_WRD_a	weak_WRD_b
SYN	97.8±1.5	97.8±1.5	88.2±3.9	82.8±3.9	90.2±3.1
NON-SYN	*	60.2±4.8	66.3±4.8	67.7±4.7	63.0±4.8

*denotes parser unable to complete

Table 5.2 Subword Testbed Performance

The results show that a high level of performance is possible on the speech understanding task using subword modelling. Weakened *a priori* match constraints and the use of wildcards are able to produce good non-syntactic performance without degrading syntactic behaviour. The use of sink model allows further improvement on the non-syntactic data with some degradation in syntactic performance. The figures shown demonstrate that a tradeoff exists between syntactic and non-syntactic performance amongst sink model variants and that this is strongly influenced by the topology of the weak match grammar. Use of sink models and good alignment of subword units led to performance which was only weakly dependent on the parser penalty. Less tuning is therefore required than in the wholeword DTW case.

6 CONCLUSIONS

A speech understanding system based on traditional pattern matching techniques has been demonstrated. Syntax control and parsing techniques have been used which are independent of the size of recognition unit and the form of frame-synchronous pattern matching utilised.

CONTINUOUS SPEECH RECOGNITION USING THE MESSIAH DATABASE

Vocabulary independent training has been used to allow rapid implementation of task specific grammars. Techniques have been demonstrated which allow the system to successfully interpret the syntactically and semantically aberrant data which could reasonably be expected from naive users of complex systems. It should therefore be possible to obtain high performance within context driven speech understanding systems where language models are determined from dialogue simulations and user studies.

ACKNOWLEDGEMENT

Acknowledgement is made to the Director of Research and Technology, British Telecom for permission to publish this paper.

HMM training was undertaken using the HTK suite of software written by Dr S J Young. Special thanks are due to Neil Russell for the initial testbed implementation and to Sheila Barbone, Mary Lumkin, Jill Jacobs, Geoff Walker and Jo Salter for their assistance with the database definition, collection and preparation activities.

BIBLIOGRAPHY

[1] Bruce I P.
Engineering an Intelligent Voice Dialogue Controller.
Computer-aided Engineering Journal Feb 1987

[2] Godin C, Lockwood P.
DTW schemes for Continuous speech recognition
Computer Speech and Language, Vol 3, No2 pp169-198, 1989

[3] Hunter P J, Watts M O
A Speech Card for Provision of Interactive Speech Systems
Digital Signal Processing: components and applications seminar, ERA 88-0386, Nov 1988

[4] Lee K F
Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System
PHD Thesis. CMU-CS-88-148. Carnegie Mellon University 1988

[5] Rose R C, Paul D B.
A Hidden Markov Model Based Keyword Recognition System
vol 1, pp129-132, ICAASP 1990

[6] Wilpon J G, Rabiner L R, Lee C H, Goldman E
Automatic Recognition of Vocabulary Word Sets in Unconstrained Speech Using Hidden Markov Models
(preprint of paper) AT&T Bell Laboratories 1989

[7] Young S J, Proctor C E
The Design and Implementation of Dialogue Control in Voice Operated Database Inquiry Systems
Computer Speech and Language Vol 3, No 4 pp329-358, 1990

[8] Young S J, Russell N H, Thornton J H S
Token Passing : a simple conceptual model for connected speech recognition systems.
Technical Report CUED/F/INFENG/TR38, Cambridge University

Proceedings of the Institute of Acoustics

AN OPTIMAL SUB WORD UNIT HMM NETWORK FOR OPEN VOCABULARY CHINESE SPEECH RECOGNITION

D.N.L.Howell

Howell Associates,
Anelgate House, Jubilee Drive, Malvern, Worcs. WR13 6DQ

1 INTRODUCTION

Sub word unit based, large vocabulary speech recognition systems like Armada[1] and Sphinx [2] are based on hidden Markov models (HMM) of various sub word units (SWU) of speech. These are simply compiled into sequences of models that then represent words. These word models are then used in making decisions about what was said. Coarticulations between the SWUs are represented by replicating SWU models for each of the many different types of context. A statistical smoothing algorithm such as deleted interpolation is used to solve problems of under training inherent in using large numbers of model replications.

The algorithm described in this paper generates a network of Chinese sub syllable models that are linked together as part of the HMM training procedure. The SWU model parameters are an integral part of the network as a whole and coarticulations between SWUs are modelled as a natural consequence of reestimating the network.

2 INITIAL SUB WORD UNIT HMMs

Where the boundaries of a sub syllable unit lie may be difficult or impossible to assess. An easier question to answer is whether a particular SWU exists in an interval of speech and to give its rough position within that interval. In order to obtain initial HMM parameters to bootstrap the reestimation process the following procedure was performed:

For each SWU;

Syllables I_i^c of speech containing the SWU S_i of interest in different contexts(c) were time aligned using an asymetric dynamic time warping alignment to the first interval. This produced a minimum distance, optimal alignment path between the first and subsequent intervals.

Data frames from examples of the same SWU class were well aligned whereas other speech frames within the syllable were spuriously aligned.

The aligned frames were then averaged, segmented and the mean and variance for each segment calculated. Thus, N data clusters were obtained for each of the N states with means μ and diagonal covariance matrix Σ.

$$\mu(s_j) = [\frac{1}{N}\sum_{t=n}^{m}(C_t)]_j \qquad (1)$$

$$\Sigma(s_j) = [\frac{1}{N}\sum_{t=n}^{m}(C_t)^2]_j \qquad (2)$$

where:
s_j = jth state of the HMM.
C_t = t th frame of averaged template.
m = start frame of segment.
n = end frame of segment.

All the allowed transition probability parameters $a_{i,j}$ were set to 1.0.

The same syllables, with the forward-backward algorithm were used to obtain maximum likelihood reestimates of the HMM parameter set for the SWU. A prototype HMM after reestimation contained a good model of the SWU of interest embedded in a generalised model of its context. Bakis [3] style HMMs were used as the inclusion of skip transitions allowed the network to organise itself for alternative prononciations in different contexts. This was found to be particularly important at SWU boundaries.

3 REESTIMATING THE DISCRIMINATIVE NETWORK

We now have a set of SWU models each embedded in a general model of context. Much of this context model will be irrelavent to the reestimated network. What is now required is a maximum likelihood solution to the SWU connection problem that will prune away irrelavent states and transitions, leaving a network of states modelling the units themselves together with transitions and coarticulations between units.

Sub-Syllable units were fully connected according to a transcription of the training data. i.e. if SWU B followed SWU A then all the states in the initial SWU model A with embedded context were connected to all the states in SWU B with embedded context. The parameter expectation values of the composite network were then accumulated using current parameter values and forward (α) and backward (β) accumulated probability passes . As one might expect, parameters between the areas of the SWU models representing context that were not pertinent to the SWU sequence produced low expectations values during reestimation whereas parameters in embedded areas that were pertinent produced higher expectations.

Expectation values for all legal SWU sequences were produced in this manner and then used to the calculate reestimated parameter values for the network . The end result was a network of labelled hidden Markov states, connected via transition probabilities. Irrelavent states and transitions were pruned away by setting a low threshold on all transition probabilities. If reestimated transition probabilities fell below this threshold then they were deleted. If no transitions then led to a state then it too was deleted.

4 THE SPEECH DATA

All sets of syllable data contained 25 examples of each syllable class for recognition and a completely separate set of 25 for training. Recognition was performed on the recognition set for all the results quoted.

- set 1 = shuai(tone 4) shuang(tone 3) huang(tone 2) huai(tone 4) shua(tone 3) hua(tone 2).
- set 2 = gang gao gong gui guo geng kang kao kong kui kuo dang dao dong dui duo deng (tone 1).
- set 3 = gang gao gong gui guo geng kang kao kong kui kuo dang dao dong dui duo deng (tone 1,4).
- four tones = geng zhun dui jiong jun zhuo gui zhou gong zun xiu (tone 1,2,3,4).

5 EXPERIMENTS

The data listed in section 5 was used to build HMMs of whole syllables and demisyllable based networks based on the techniques described in sections 3 and 4. These were then compared and assessed.

6 MAIN RESULTS

6.1 Algorithm Evaluation

The following table shows the main results of the evaluation of algorithms discribed in this paper. The figures quoted are percentage error rates.

				pair	trio	set1	set2
Isolated syllable HMMs[1]				16	38	10	18
Isolated syllable SHMM[2]				1	13	4.8	-
Demi-syllable based network:							
ff[3]	an[4]	seg[5]	df[6]				
n	n	u	n	-	-	14	21
n	y	u	n	-	-	9	12
n	y	u	y	-	-	8	-
n	y	a	n	-	-	4	-
y	y	a	n	-	-	3	-

[2]8 state Bakis model hidden Markov model per class.

[3]8 state semi-hidden Markov model using Poisson duration statistics.

[4]ff: During initial 'startup model' building, the first frame of training data was always assigned its own state.

[5]an: The mean amplitude of the data frames was added as an extra channel and subtracted from the other channel values.

[6]seg: During initial 'startup model' building, initial state output means and variances were taken from: u= uniform segmentation of time aligned training data. a= asymetric segmentation of time aligned training data.

[7]df: Amplitude normalised data frames included the mean amplitude of the previous frame.

7 DISCUSSION AND CONCLUSIONS

The novel algorithms and software developed during this project represent a major advance in Chinese speech recognition. The accuracy of the recognition is in excess of that acheived during other published studies. It is difficult to compare the performance with that of similar systems developed for other languages but the 3% error rate of the best version of our system stands up well when superficialy compared with others. For example without the help of a grammar the best HMM SPHINX phone-based system has a word accuracy of 50-60%. The performance figure for our system was obtained on the most difficult subset of the Mandarin syllabary but used isolated citation form syllables rather than fluently spoken but more discriminable English speech. A more exacting test will be on a fluently spoken vocabulary containing the complete syllabary.

The use of the onset and rhyme as the two phonological entities for recognition have given us the advantage of robust units that are reasonably stable over contextural variations. Variations that do occur are modelled well using the HMM network generated using the techniques for building and connecting the SWU HMMs. Phone based recognisers developed elsewhere must replicate models to accomodate different contexts, whereas our chosen demi-syllable models account for this variation within their structure.

8 ACKNOWLEDGEMENTS

This work was assisted by the Alvey Directorate as project MMI054. The research was done while the author was working as principal researcher for Sindex Speech Technology Ltd. The work was conducted at the Speech Research Unit at RSRE Malvern.

The author would like to thank the members of the SRU and in particular Martin Russell and Roger Moore for their help and assistance during this project. I would also like to thank Dr Paul Thompson at the School of Oriental and African Studies for his invaluable insight into the phonology of spoken Mandarin.

9 REFERENCES

1. M J Russell et al,'The ARM Continuous Speech Recognition System' ICASSP90 Conf Proc May 1990.
2. K-F Lee, Large Vocabulary Speaker Independant Continuous Speech Recognition: the SPINX System. PhD Thesis, Carnegie Mellon University,1988.
3. R. Bakis, Spoken Word Spotting Via Centisecond States, IBM Technical Disclosure Bulletin Vol 18 No10 March 1976 pp. 3479-3481.