Computer Speech and Language (1997) **11,** 73–89

Multilingual large vocabulary speech recognition: the European SQALE project

S. J. Young,* M. Adda-Dekker,† X. Aubert,‡ C. Dugast,‡ J.-L. Gauvain,† D. J. Kershaw,* L. Lamel,† D. A. Leeuwen§ D. Pye,* A. J. Robinson,* H. J. M. Steeneken,§ P. C. Woodland*

**Cambridge University Engineering Department, Cambridge University, Trumpington Street, Cambridge, CB2 1PZ, U.K., †LIMSI-CNRS, B.P.133, 91403, Orsay, France, ‡Philips GmbH, P.O. Box 1980, D-52021, Aachen, Germany, §TNO Human Factors Research Institute, P.O. Box 33, 3769 ZG Soesterberg, The Netherlands*

Abstract

This paper describes the SQALE project in which the ARPA large vocabulary evaluation paradigm was adapted to meet the needs of European multilingual speech recognition development. It involved establishing a framework for sharing training and test materials, defining common protocols for training and testing systems, developing systems, running an evaluation and analysing the results. The specifically multilingual issues addressed included the impact of the language on corpora and test set design, transcription issues, evaluation metrics, recognition system design, cross-system and cross-language performance, and results analysis. The project started in December 1993 and finished in September 1995. The paper describes the evaluation framework and the results obtained.

The overall conclusions of the project were that the same general approach to recognition system design is applicable to all the languages studied although there were some language specific problems to solve. It was found that the evaluation paradigm used within ARPA could be used within the European context with little difficulty and the consequent sharing amongst the sites of training and test materials and language-specific expertise was highly beneficial.

1. Introduction

Significant advances have been made in recent years in the area of large vocabulary speaker independent continuous speech recognition. For American English, current laboratory systems are capable of transcribing continuous speech from any speaker with average word error rates of between 5% and 10% (Bahl *et al.*, 1995; Dugast *et al.*, 1995; Gauvain, Lamel & Adda-Decker, 1995), and with adaptation, this can be improved

0885–2308/97/010073 + 17 $25.00/0/1a960023

further (Woodland *et al.*, 1995). However, comparable results are generally not available for other languages because, regardless of their native language, many research groups have focussed their system development efforts on American English. One of the main reasons for this has been the existence of the US ARPA CSR programme, and more importantly for *outsiders*, the annual ARPA CSR benchmark tests (Kubula *et al.*, 1994; Pallett *et al.*, 1994, 1995).

Although the original motivation for these evaluations was to compare the performance of the ARPA-sponsored contractors, in practice, they have provided much more than just a simple means of monitoring progress. Each successive evaluation involves testing on a set of previously unknown speakers in a known domain which has typically been read newspaper text. All participants are provided with training data and other infrastructure materials such as standard language models, and the test protocols and results analysis are carefully prescribed. Prior to each evaluation, a working group defines the targets and the basic materials needed to develop and test systems. The post-mortem analysis of the results helps all participants to understand the strengths and weaknesses of their and other systems. The net result is a very strong technology pull.

This paper describes the SQALE (Speech Quality Assessment for Linguistic Engineering) Project. The aim of the project which was sponsored by the European Commission was to adapt the ARPA evaluation paradigm described above to meet the needs of European multilingual speech recognition development. It thus involved establishing a framework for sharing training and test materials, defining common protocols for training and testing systems, developing systems, running an evaluation and analysing the results. To achieve these objectives, there are many issues that must be dealt with when moving to different European languages. These include the impact of the language on corpora and test set design, transcription issues, evaluation metrics, recognition system design and results analysis (Moore, 1988; Steeneken & van Veldon, 1989; Eskenazi, Mariani & Bornerand, 1991). The project started in December 1993 and finished in September 1995. A dry-run evaluation was conducted in February 1995 and the actual evaluation was conducted during April and May 1995.

The SQALE project was coordinated by the Netherlands Human Factors Research Institute (TNO) who were responsible for defining protocols in cooperation with the other partners, supplying test data, monitoring the evaluation and analysing the results. The other partners in the project were Cambridge University Engineering Department (CUED) in England, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) in France and the Man–Machine-Interface group with Philips Research Laboratories (Philips) in Germany. CUED had two systems: CU-CON a connectionist system and CU-HTK a HMM system. Since all partners had previously participated in the ARPA evaluations and had therefore already developed systems for American English, this language was used as a baseline for studying cross-language issues. During the SQALE project, each partner ported their system to one or more of the three European languages tested: British English, French and German. Table I shows the languages that each system was tested on.

The remaining sections of this paper describe the various aspects of the SQALE project in more detail. Section 2 describes the evaluation paradigm and how the various cross-language issues were dealt with in preparing the training materials, designing the tests and analysing the results. Section 3 gives a brief overview of the four recognition systems used and discusses the language specific problems encountered during the cross-

TABLE I. Languages tested by each SQALE system

	CU-CON	CU-HTK	LIMSI	Philips
American	●	●	●	●
British	●	●	●	
French		●	●	●
German		●	●	●

TABLE II. Characteristics of acoustic model training data. All data was recorded at 16 kHz sample rate using 16 bit precision

	American	British	French	German
Corpus	WSJ0	WSJCAM0	BREF-80	PHONDAT
Source	*Wall Street Journal*		*Le Monde*	Phonetic Balanced + Train Queries
No. of speakers	84	92	76	155
No. of sentences	7·2k	8·1k	5·1k	15·2k
Corpus size	14·0 h	13·4 h	9·2 h	13·0 h
No. of spoken words	131k		90k	125k
No. of distinct words	9084		13 850	1725
Av. freq/words	14·4		6·46	72·4
Microphone(s)	Sennheiser HMD414-6		Shure SM10	Various

language development. Section 4 describes the evaluation results and presents cross-system and cross-language comparisons. Finally, Section 5 discusses the main issues arising from the project and Section 6 presents overall conclusions.

2. Evaluation framework

As noted above, a principle aim of the SQALE project was to make both cross-language and cross-system comparisons within a European context. However, time and resource limitations meant that not all the sites could build and test systems for all languages. Hence American English was used to provide a common baseline for comparison. Further normalization was achieved by using common word lists and language models, by ensuring that comparable amounts of training data were used for each language, and by attempting to select test material in each language of similar difficulty.

2.2. Training materials

Table II summarizes the main characteristics of the training data. The 84 speaker subset called *SI-84* of the Phase 0 portion of the *Wall Street Journal Database* (*WSJ0*) (Paul & Baker, 1992) was used to train the baseline American English systems and a British English version based on the same text source called *WSJCAM0* was used for the British English systems (Fransen *et al.*, 1994). *BREF-80* based on the *Le Monde* text source was used for the French systems (Lamel, Gauvain & Eskanazi, 1991). All of these three databases were recorded using similar close-talking microphones and all

Table III. Characteristics of lexica and language model training text

	English	French	German
Source	*Wall Street Journal*	*Le Monde*	*Frankfurter Rundschau*
Corpus size	37·2 MW	37·7 MW	31·5 MW
No. of distinct words	165k	280k	500k
Vocabulary size	20k	20k	64k
2-gram perplexity	198	178	430
3-gram perplexity	135	119	336
Ave. phones/word	4.16	3.53	5.09
Text homophone rate	18%	57%	—
Monophone words	3%	19%	0·5%

three come from the same domain as the corresponding language models and evaluation data. In contrast, the German systems were built using the *PHONDAT* database which consists of a mixture of phonetically balanced sentences and train information queries recorded using a variety of different microphones at multiple locations. Thus, PHONDAT was not matched to the evaluation data either acoustically or linguistically. The use of PHONDAT was in many ways regrettable but it was the only German data available at the time. Its use does, however, highlight the consequences of mis-match between the training and testing data.

2.2. Vocabularies and language models

A common trigram language model was used for each system in each language and newspaper text was used for training data in each case. Table III summarizes the main characteristics of each language model. The standard MIT-Lincoln Labs 20k-open *Wall Street Journal* trigram language model was used for the American and British English systems[1]. Trigram language models using similar-sized training sets were constructed for French and German using the *Le Monde* and *Frankfurter Rundschau* newspapers, respectively. Note that the increased number of distinct words for French and German is due in part to the fact that upper and lower case were distinguished in these languages whereas case was ignored in English.

The word lists used to build recognizers for each of the test languages were determined by sorting the unigram frequencies within each training corpus. Note that a word in this context is defined as being any distinct sequence of letters delimited by white space or punctuation marks[2]. Table IV shows the lexical coverage of each language for different vocabulary sizes. As can be seen, a much larger vocabulary is needed for German in order to achieve similar coverage. Hence, a 64k word vocabulary was used for German as compared to 20k for each of English and French.

Table III also shows some properties of the lexicon for each language. French has a high number of monophone words and a correspondingly high homophone rate. In contrast, German has relatively few monophone words and a relatively low homophone

[1]Used in the 1993 US ARPA CSR Evaluation.

[2]In practice, the definition of what constitutes punctuation is language specific. For example, "I'ami" in French is split into two words whereas "friend's" is a single word.

TABLE IV. Lexical coverage as a function of vocabulary size

No. of words	English	French	German
5k	90·6%	85·2%	82·9%
10k	94·9%	90·6%	86·1%
20k	97·5%	94·6%	90·0%
40k	99·0%	97·3%	93·9%
60k	99·6%	98·3%	95·1%
80k	99·7%	98·9%	95·7%

TABLE V. Distribution of sentence length and perplexity for the 10 sentences uttered by each speaker

	Sentence length			With OOV
	Short	Normal	Long	
Low PP	1		1	1
Normal		4		
High PP	1		1	1

rate. Thus, although the perplexity of the German language model is high compared to English and German, the confusability of the vocabulary is lower. These factors, which were expected to influence the relative recognition rates on the three languages, are discussed further below.

2.3. Selection of test data

The evaluation data was assembled from existing recordings augmented by new recordings undertaken by TNO. For each language there were 200 test sentences from 20 speakers plus a set of approximately 60 diagnostic sentences. The test sentences were chosen to give a reasonable spread of difficulty as determined by sentence length and perplexity. Table V shows the general distribution for these measures over the 10 sentences uttered by each speaker.

The diagnostic sentences were designed to allow cross-speaker and within-speaker variability to be studied. They consisted of a set of 10 different speakers uttering the same three *common* sentences plus a set of six speakers uttering the same *replica* sentence five times. The diagnostic sentences were not part of the official evaluation and they were processed by the same systems after the evaluation had finished.

2.4. Evaluation protocols

The evaluation was conducted along similar lines to the official ARPA CSR evaluations. The test data along with reference transcriptions was delivered to each site on CD-ROM with instructions it not open it before 18th April 1995 at 9am. The deadline for submission of recognition output for American English, British English and French was 8th May 1995 and for the German it was 25th May 1995. There was then an

TABLE VI. Comparison of front-end processing for each of the recognition systems

	CU-CON	CU-HTK	LIMSI	Philips
Front-End	PLP + MEL +	MFCC	MFCC	FBANK
Time Dep.	Rec NN	$\Delta + \Delta^2$	$\Delta + \Delta^2$	LDA
Dimension	23 + 13	39	48	35

TABLE VII. Comparison of the Three HMM recognition systems

Feature	CU-HTK	LIMSI	Philips
Emission probabilities	Gaussian	Gaussian	Laplacian
Training method	Baum–Welch	MAP	Viterbi
Triphone type	Full cross-word	Cross-word	Word-internal
State clustering method*	Decision tree	Agglom. data clustering	

*The LIMSI French system used 779 context-dependent models and no state-tying.

adjudication period followed by an announcement of the official results on 8th June 1995.

All recognition output was scored using the standard NIST scoring software. The output for American and British English was case insensitive and for French and German it was case sensitive. A standard was defined for mapping accented characters to plain ASCII.

3. Systems

This section briefly describes and compares the four recognition systems used in the SQALE evaluations. It then discusses some of the language specific issues encountered.

3.1. The recognition systems

3.1.1. Front-end processing

The front-end processing used by each of the four systems is summarized in Table VI. The CU-HTK and LIMSI systems use standard MFCC-based front-ends (Davis & Mermelstein, 1980) augmented by 1st and 2nd order derivatives. The Philips system uses a two stage process. Firstly, a 63-dimension acoustic vector is computed from a 30 channel filter bank plus total energy augmented by the first 16 first and second differences. Linear discriminant analysis is then applied to a window of three successive acoustic vectors to generate a 35-dimension feature vector (Haeb-Umbach, Geller & Ney, 1993). The CU-CON system uses two separate front ends. One consists of 12 PLP coefficients (Hermansky, 1990) plus energy and the other consists of 20 mel-filter bank amplitudes augmented by pitch, voicing and energy (MEL +).

3.1.2. Acoustic modelling

The CU-HTK, LIMSI and Philips systems are all tied-state continuous density HMM-based systems (Dugast, Aubert & Kneser, 1995; Lamel, Adda-Decker & Gauvain, 1995; Pye, Woodland & Young, 1995). Table VII summarizes the main characteristics of each

TABLE VIII. Number of states and number of Gaussians per state used in each of the three HMM recognition systems

Language	No. of parameters	CU-HTK	LIMSI	Philips
American	No. of comps/state	9	~32	~30
	No. of states	3950	2814	~2500
British	No. of comps/state	8	~32	
	No. of states	3494	2582	
French	No. of comps/state	10	~32	~30
	No. of states	2638	2337*	~2100
German	No. of comps/state	10	~32	~30
	No. of states	4268	2481	~3000

*The LIMSI French system used 779 context-dependent models and no state-tying.

recognizer and Table VIII lists the number of parameters used by each system for each language. The CU-HTK system uses phonetic decision trees to perform state clustering (Young, Odell & Woodland, 1994). This allows it to synthesize models for contexts which do not occur in the training data and it can thereby use full cross-word triphones. The LIMSI system uses cross-word triphones for which there are sufficient training examples and backs-off to diphones for unseen and infrequent contexts. The Philips system uses word-internal triphones only. The acoustic modelling in the Philips system is further simplified by using Laplacians with a single global deviation vector instead of the more conventional state-specific diagonal variance Gaussians used by the CU-HTK and LIMSI systems. All three systems use gender dependent model sets.

Acoustic modelling in the CU-CON system uses four recurrent neural networks, one for each input parameterization (MEL+ and PLP) and one for each time direction (forward and backward) (Hochberg, Renals & Robinson, 1995). Each network consists of a single layer and the output at each time frame is a vector of phone probability estimates augmented by a 256-dimension state vector which is fed-back to the input. The phone probability estimates from each of the four networks are merged to form a single posterior probability of each phone for each input frame.

In addition to these phone probability estimation networks, four sets of feed-forward networks are trained to estimate context-classes for each phone based on the state feedback vector of the corresponding recurrent network. The outputs of these are merged and then multiplied by the context-independent phone probabilities to give posterior context-dependent phone probabilities. The contexts are chosen using a decision tree clustering procedure to give 527 context-dependent phones for American English and 465 for British English (Kershaw, Hochberg & Robinson, 1995).

3.1.3. Dictionaries

The requirement to use a common language model for each of the languages and the limited availability of lexicons meant that there was little variation across systems. For American English all systems used the LIMSI Nov '93 20k pronunciation dictionary except for Philips who used the Dragon dictionary. For British English, all systems used the CUED BEEP dictionary with little modification. For French, the LIMSI French pronouncing dictionary was used. However both CU-HTK and Philips modified this to handle liaisons (see below). For German, all systems used a dictionary supplied by

Philips. However, LIMSI made extensive modifications in the form of corrections, a reduction in the number of vowels and the addition of alternative pronunciations.

3.1.4. Decoding

The three HMM systems use time-synchronous Viterbi-decoders in a two-pass scheme. The details vary but the broad outline of operation for all of them is as follows. In a first pass, a bigram language model is used with gender independent models to create word level lattices. The lattice for each sentence is then expanded using a trigram language model. In a second pass, the word-level trigram lattices are re-scored using gender dependent models (Odell *et al.*, 1994; Aubert & Ney, 1995).

The CU-CON system uses a stack decoder and operates in a single pass with a loosely coupled language model. The search space is reduced through the usual likelihood-based pruning and also through posterior-based phone deactivation pruning in which phones with a low estimated posterior probability are pruned (Renals & Hochberg, 1995).

3.2. Language specific issues

Only the three HMM-based systems were applied across different languages. Given the very limited time for development, the approach of each site was to apply existing techniques wherever possible and to minimize the amount of language-specific engineering put into each system. The language-specific problems which were encountered are discussed here.

3.2.1. Liaison in French

Liaison in French can be regarded as an optional pronunciation variant in which normally silent word final consonants are pronounced when immediately followed by a word initial vowel. In the LIMSI and Philips systems, all words in the dictionary which might give rise to a liaison are marked and rules are applied during both training and recognition to ensure that liaisons are only allowed when the following word starts with a vowel or, for some words, starts with a "h". These rules ensured that only the liaison consonants appropriate to the context were inserted (Gauvain *et al.*, 1994; Aubert & Ney, 1995).

In the CU-HTK system, liaison was handled by adding pronunciation variants to the dictionary for all cases where liaison could occur. This simple approach has the disadvantage that it also allows the recognizer to accept liaisons in many situations where they would not actually occur. It nevertheless gave $\sim 10\%$ reduction in errors when compared to a baseline system using the original LIMSI dictionary. As a further improvement to the CU-HTK system, additional liaison-specific consonants for /z/, /t/ and /n/ were added so that cross-word triphone context could be used to minimize the probability of inadmissable liaison pronunciation variants being used during recognition. This was found to give a further small improvement and when it was compared to a recognizer which had been modified in a similar way to the LIMSI and Philips systems, there was no significant difference.

3.2.2. Compounding in German

Compounding in German results in a much reduced coverage for a similar sized vocabulary compared to French or English. An effective solution to compounding

TABLE IX. Word error rates using a trigram grammar

System	American	British	French	German
CU-CON	12·9%	13·8%		
CU-HTK	13·2%	14·4%	15·1%	18·7%
LIMSI	13·5%	15·4%	15·3%	16·1%
Philips	14·7%		16·1%	19·7%

TABLE X. Statistically significant differences between systems using a trigram grammar. MP: Matched pair sentence segment; SI: signed pair comparison; WI: Wilcoxon signed rank; MN: McNemar sentence error test

Difference		Language	Tests
CU-CON	PHILIPS	American	MP,WI
CU-HTK	Philips	American	MP,WI
CU-CON	LIMSI	British	MP,WI
LIMSI	CU-HTK	German	MP,SI,WI,MN
LIMSI	Philips	German	MP,SI,WI,MN

requires a morphologically motivated decomposition procedure for source training texts and a similar inverse procedure for reforming compounds in the recognizer output (Geutner, 1995). However, since compounding can also give rise to pronunciation changes this is not at all straightforward.

The compounding problem and the difficulty of dealing with it was noted at the outset of the project and the simple solution adopted was to increase the German vocabulary size to 64k words compared to the 20k word vocabularies used for French and English. A consequence of this is that the language model is much larger and there are more distinct triphone contexts, especially for word-internal triphone systems. Thus, effective smoothing techniques for both acoustic and language modelling were particularly important for German.

3.2.3. Glottal stops in German

The glottal stop in German has no distinctive role within isolated words, but it does frequently occur at word and morpheme boundaries in continuous speech. Both CU-HTK and LIMSI performed experiments to determine whether or not to include the glottal stop in the phone set for German. CU-HTK found little significant difference whereas LIMSI found a slight improvement. In the evaluation, LIMSI and CU-HTK included the glottal stop in their acoustic modelling for German and Philips excluded it.

Results

4.1. Baseline results

The results of the SQALE evaluation using trigram language models are summarized in Table IX and the corresponding statistical significance tests are shown in Table X

TABLE XI. Word error rates using a bigram grammar

System	American	British	French	German
CU-CON	17·0%	17·2%		
CU-HTK	16·7%	18·3%	18·9%	21·6%
LIMSI	17·2%	18·8%	17·7%	18·4%
Philips	20·3%		20·3%	22·4%

TABLE XII. Word error rates on the American English evaluation data using the 1994 state-of-the-art HTK recognition system

Training data	Acoustic model	Vocabulary size	Language model	Word error	Error reduction
SI84	triphone	20k	1993 3-gram	13.2%	
SI284	quinphone	20k	1993 3-gram	10.3%	22%
SI284	quinphone	65k	1994 4-gram	6·9%	33%
SI284	quinphone*	65k	1994 4-gram	6·3%	8%

*System incorporating incremental speaker adaptation.

(Gillick & Cox, 1989; Martin, 1995). As can be seen, the CU-CON system had the lowest error rate on the two English tests, the CU-HTK system had the lowest error rate on French and the LIMSI system had the lowest error rate on German. However, all the systems were very similar and there was no statistically significant difference between the 1st and 2nd ranked system in the English and French tests. Only the LIMSI system on German was significantly better than any other.

Table XI shows the corresponding results using a bigram language model where the general pattern is similar.

4.2. Comparison with state-of-the-art

In order to allow cross-language comparisons using existing training material, the data allowed for the American English systems was limited to the *WSJ*0 corpus which contains around 14 h of acoustic training data S184 and 37M words of text. These were the conditions in force for the 1993 U.S. ARPA CSR evaluations. However, the subsequent addition of the *WSJ*1 corpus has greatly extended the amount of training material available for American English to around 66 h of speech (SI284) and 227M words of text.

Table XII shows the further improvements that can be gained with the CU-HTK system when more training data is available (Woodland *et al.*, 1995). The first line shows the standard CU-HTK SQALE evaluation result on the American English test. Increasing the amount of acoustic training data from 12 h to 66 h allows more robust models to be constructed using wider context in the phonetic decision trees. This results in a 22% reduction in error rate. The third line shows the effect of increasing the vocabulary size to 65k words and increasing the amount of language model training data from 37M to 227M words. The increased vocabulary size reduces the OOV rate from 1·46% to 0·39% and the increased training material allows a 4-gram language

TABLE XIII. Word and sentence error rates of human listeners compared to the SQALE recognizers on American and British English test data

Recognizer	Word error	Std. Dev.
Native listeners	2·63%	0·93%
Non-native listeners	7·40%	1·67%
SQALE recognizers	12·60%	1·92%

model to be built. This results in a further 33% reduction in error rate. Finally, the last line shows the effect of using incremental speaker adaptation. This uses maximum likelihood linear regression to estimate the parameters of a set of matrices which are used to transform the Gaussian mean vectors (Leggetter & Woodland, 1995). This provided an addition 8% reduction in error and with more sentences per speaker, this adaptation would have had greater effect. Overall, this state-of-the-art CU-HTK system showed a 51% reduction in error rate compared to the best SQALE evaluation system. In a separate study on the performance of their French recognition system, LIMSI reported comparable error reductions when using additional training data (Gauvain *et al.*, 1994).

4.4. Comparison with human performance

In order to compare the recognition performance of the automatic systems with humans, a set of 80 sentences used in a dry-run of the evaluation were presented to 30 listeners. The test sentences consisted of 40 American and 40 British English. The sentences were selected to have an even distribution of sentence length, perplexity and speaking rate. The listeners consisted of 20 native speakers and 10 non-native Dutch speakers who had lived in the U.S. or U.K. for some time (van Leeuwen, van den Berg & Steeneken, 1995).

The sentences were presented to the subjects using headphones in a quiet room. First the whole sentence was played and it was then replayed in segments of four to seven words, split wherever possible at natural phrase boundaries. The subject then typed the words that had been perceived. Subjects could replay segments but were encouraged to make a decision as soon as they could. The human generated transcriptions were manually corrected for spelling errors and then scored against the reference transcriptions in the usual way. The same sentences were also processed by the CU-CON, CU-HTK and LIMSI recognition systems set up exactly as in the main evaluation.

Table XIII shows the word error rates and standard deviations of the humans compared to the average performance of the three recognition systems. As can be seen, the native listeners easily outperform the machines. However, while the non-native performance is better than the SQALE systems, it is worse than that of the state-of-the-art CU-HTK system described earlier which recorded 5.2% errors on this test set.

An analysis of variance was performed on these results to determine the main

TABLE XIV. The significant parameters influencing the word error rates of human listeners and machines recognizing American and British English test data

Group	Parameter	Word error
Native listener	Sentence length (short/long)	1·7%/2·9%
Non-native listener	Sentence length (short/long)	5·2%/8·0%
Non-native listener	Perplexity (low/high)	5·7%/7·5%
Machine	Perplexity (low/high)	5·1%/20·1%

influences on recognition accuracy. Table XIV shows the most significant distinctions where for sentence length, short is ~eight words and long is ~23 words; for perplexity, low is ~50 and high is ~500. As can be seen, native listeners are mostly influenced by sentence length but machines are mostly influenced by perplexity. Non-native speakers are influenced by both. Speaking rate was measured in terms of words per min, surprisingly perhaps, this was not found to be significant.

4.4 Speaker variability

As mentioned in Section 2.3, the official SQALE evaluation data was augmented by a small set of diagnostic sentences designed to investigate issues of within and between-speaker variability. This diagnostic set consisted of approximately 30 *common* sentences spoken by 10 speakers and a set of 30 *replica* sentences consisting of five repetitions by each speaker.

The mean word error rate and variance was first computed for each replica set of five sentences for each recognition system and language. The number of replica sentences is rather small and hence the spread of variance as a function of the mean is rather high. Nevertheless, it was found that the assumption of a binomial distribution for word error rate was consistent with the experimental data and provides a reasonable estimate of variance.

Given this assumption, the within-speaker variance on the per speaker word error rate σ_w^2 can be approximated by

$$\sigma_W^2 = \frac{\sum_{j=1}^{J} w_j(1-w_j)N_j}{\sum_{j=1}^{J}(N_j-1)} \tag{1}$$

where w_j is the mean word error rate for speaker j expressed as a fraction, N_j is the number of words spoken by the jth speaker and J is the total number of speakers. The between-speaker variance σ_B^2 is computed from the usual sample average, that is

$$\sigma_B^2 = \frac{\sum_{j=1}^{J}(w_j-\bar{w})^2 N_j}{J-1} \tag{2}$$

where $\bar{w}$ is the global mean error rate.

From these estimates of variance, the F-ratio σ_B^2/σ_w^2 was calculated for each language

TABLE XV. The *F*-ratios of between and within-speaker variability

Language	$F_{95\%}$	CU-CON	CU-HTK	LIMSI
American	1·88	4·09	5·22	2·34
British	1·88	2·31	1·99	1·08
French	2·21		1·30	1·50
German	1·88		1·05	1·61

TABLE XVI. The number of errors caused by each OOV word for each language

American	British	French	German
1·5	1·6	1·8	1·3

and system and the results are shown in Table XV (Hays, 1963). Also shown in this table is the *F*-ratio required for the difference in means between speakers to be significant at the 95% confidence level. As can be seen this analysis indicates that only the American English data contains significant speaker variability for all systems.

4.5. Errors caused by out of vocabulary (OOV) words

To measure the effects of OOV words on error rate in the differing languages, each test sentence containing an OOV word was paired with a sentence of similar perplexity which had no OOV words. The error rates of the OOV set and the non-OOV set could then be compared. Scaling the number of words in the OOV set by the error rate on the non-OOV set gives an estimate of the number of errors that would have occurred in the OOV set even if there had been no OOVs. This allows the ratio of actual errors incurred for each OOV to be estimated and the results are shown in Table XVI. As can be seen French has the highest ratio and German the lowest. The ratios for American are somewhat lower than those calculated in a comparable analysis made for the ARPA 1994 evaluation results where values in the range 1·7–2·1 were found (Pallet *et al.*, 1995).

5. Discussion

The results of the main SQALE evaluation showed broadly comparable performance across all of the systems and languages tested. The connectionist system CU-CON ranked first on both the American and British tests and although there was no statistically significant difference between it and the nearest HMM-based system, it is still interesting to speculate on the reasons for its relatively good performance. The recurrent neural network used as a probability estimator in the CU-CON system is trained to discriminate between phones. It thus uses the training data to estimate decision boundaries, unlike the HMM systems which use the training data to estimate probability distributions. The net result is that the CU-CON system uses fewer parameters and as a consequence, it has relatively more training data per parameter. As shown in Section 4.2, a HMM-

based system can give much better performance when there is more training data and although comparable results for the CU-CON system were not produced, the results of recent ARPA evaluations (Pallett *et al.*, 1994, 1995) suggest that the CU-CON system would not be better in this case. Thus, there is some evidence that in these SQALE evaluations, the HMM-based systems suffered more from the limited training data than did the CU-CON system.

Although all of the HMM-based systems tested had broadly similar performance, the Philips system had the highest error rate on all of the languages that it was tested on. Thus, it appears that the simplifications adopted within the Philips system of using a single global rather than state-specific deviation (variance) vector, and limiting triphone contexts to within words did have an adverse effect on performance.

The performance of the German systems was around 30% worse than for the English systems and it is believed that the mismatch between the acoustic training data and the test data played a significant part in causing this degradation.

The experience of each of the sites taking part in SQALE was that a reasonably competitive recognition system can be constructed without paying any special attention to the specific target language. Once a phone set has been chosen and a dictionary obtained, most of the processing is routine and language independent. However, once the basic system has been constructed subsequent refinement is typically language dependent. For example, both CU-HTK and Philips made significant improvements to their initial French systems by taking specific actions to deal with liaison. Similarly, LIMSI expended considerable effort in refining their version of the German dictionary and this resulted in their system being significantly better than any of the others.

The human benchmark tests show that whilst there is still some way to go before machines can match the competence of native speakers, performance is approaching that of non-native speakers. The factors which affect performance also seem to be different for humans who prefer short sentences and machines which prefer low perplexity sentences. These differences may be attributed to the limited short-term memory capabilities of humans and the rather crude language model available to the machines. This suggests that improvements in language modelling will be essential if the performance gap between humans and machines is to be bridged. Unlike other studies performed during the ARPA evaluations, speaking rate appears to have had no significant influence on the results achieved here. However, the speaking rate was measured in terms of words per unit time and a phone-based measurement of rate might have led to a more meaningful analysis.

Part of the motivation for the SQALE project was to discover what new problems would arise when adapting the U.S. ARPA evaluation paradigm to European languages. One of the first issues that arose was that of case sensitivity in the output transcriptions. Like the ARPA tests, the English reference comparisons were case insensitive, whereas case was significant for French and German. In the event, this turned out to be a minor issue since scoring the French and German systems with case disregarded resulted in only a 0·1% and 0·3% drop in error rate, respectively.

Perhaps a more important issue in comparing performance across languages concerns homophones. French has a very high homophone rate and most phones can correspond to one or more graphemic forms. For example, /ɛ/ can stand for *ai*, *aie*, *aies*, *ait*, *aient*, *hais*, *hait*, *haie*, *haies*, *es*, and *est*. This ambiguity can extend to phrase length sequences. For example, one OOV induced error resulted in "*épanouissait*" being recognized as "*est pas nous il s'est*". Overall, homophone confusions in French represent about 20%

of all recognition errors. Again the long term solution to these problems will require more powerful language models. In the meantime, however, the problem that remains is how to account for them when making cross-language comparisons.

In designing the test material and in the subsequent analysis, TNO attempted to normalize for the differences in difficulty between the test sets. The results showed that perplexity has a strong influence on recognition performance. Given that recognition systems explicitly include language model log likelihoods in their hypothesis scoring, this is perhaps not surprising. However, the relationship between perplexity and recognition error is complex. TNO have studied this relationship and they have developed a model relating word error to perplexity based on an *arcsin* transform (Steeneken & van Leeuwen, 1995).

In addition, word-level perplexity clearly has limitations when making cross language comparisons. Although the perplexity of the German test data was much higher than the other languages, much of this increase was due to compounding. Compounding, however, tends to make words longer and hence less confusable. Thus, taking into account the training data mismatch discussed above, the error rate for German was lower than a simple perplexity measure would suggest. In future, alternatives to word-based perplexity need to be studied, for example by using morphemes as the basic unit.

6. Conclusions

The evaluation framework imported from the U.S. ARPA evaluations appears to be highly portable and little difficulty was encountered in making the necessary performance analyses in each language. However, finding an effective method for determining task difficulty and equalizing across languages remains an unsolved problem.

The overall conclusions of the SQALE project were therefore that the same general approach to recognition system design is applicable to all the languages studies. Typically there will be some language specific problems to solve such as liaison in French and compounding in German. However, the basic modelling approach appears to be straightforward to port across languages.

Finally, a major benefit of extending the ARPA evaluation paradigm for use in the European context was that the consequent sharing amongst the sites of training data, test materials and language-specific expertise was highly beneficial. Each site made rapid progress in developing systems for languages which it might not otherwise have had the opportunity to work on. Overall this created a very strong technology pull which if continued would make a significant difference to the development of large vocabulary recognition systems in Europe.

The SQALE project was sponsored by the European Commission, DG XIII, as part of the linguistic Research and Engineering Programme. In addition to the authors of this paper, many other people have contributed to the various systems and the evaluations including: at CUED, Gary Cook, Jeroen Fransen, and Julian Odell; at LIMSI, Gilles Adda; and at Philips, Reinhard Kneser.

The project partners would also like to thank Lidia Pola at the EC and the project reviewers, Isabel Trancoso and Melvyn Hunt.

References

Aubert, X. & Dugast, C. (1995). Improved acoustic-phonetic modeling in the Philips' dictation system by handling liaisons and multiple pronunciations. *Proceedings of Eurospeech*, pp. 767–770, Madrid, Spain.

Aubert, X. & Ney, H. (1995). Large vocabulary continuous speech recognition using word graphs. *Proceedings of ICASSP*, Vol 1, pp. 49–52, Detroit, MI, U.S.A.

Bahl, L. R., Balakrishnan-Aiyer, S., Bellegarda, J. R., Franz, M., Gopalakrisnan, P. S., Nahamoo, D., Novak, M., Padmanabhan, M., Picheny, M. A. & Roukos, S. (1995). Performance of the IBM large vocabulary continuous speech recognition system on the ARPA. *Wall Street Journal task. Proceedings of ICASSP*, Vol 1, pp. 41–44, Detroit, MI, U.S.A.

Davis, S. B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions of ASSP*, **28**, 357–366.

Dugast, C., Aubert, X. & Kneser, R. (1995). The Philips large-vocabulary recognition system for American English, French and German. *Proceedings of Eurospeech*, pp. 197–200, Madrid, Spain.

Dugast, C., Kneser, R., Aubert, X., Ortmanns, S., Beulen, K. & Ney, H. (1995). Continuous speech recognition tests and results for the NAB'94 corpus. *Proceedings of the Spoken Language Technology Workshop*, pp. 156–161, Austin, TX, U.S.A.

Eskenazi, M., Mariani, J. & Bornerand, S. (1991). Report on the ICSLP satellite workshop on assessment in Kobe (Japan) and visits to several Japanese laboratories working on speech communication, 19–30 November 1990. *Speech Communication*, **10**.

Fransen, J., Pye, D., Robinson, A. J., Woodland, P. C. & Young, S. J. (1994). *WSJCAM0 corpus and recording description.* CUED Technical Report, CUED/F-INFENG/TR.192.

Gauvain, J.-L., Lamel, L. & Adda-Decker, M. (1995). Developments in large vocabulary dictation: the Limsi Nov 94 NAB system. *Proceedings of the Spoken Language Technology Workshop*, pp. 131–138, Austin, TX, U.S.A.

Gauvain, J.-L., Lamel, L., Adda, G. & Adda-Decker, M. (1994). Continuous speech dictation in French. *Proceedings of International Conference Spoken Language Processing, ICSLP*, Yokohama, Japan.

Geutner, P. (1995). Using morphology toward better large vocabulary speech recognition systems. *Proceedings of the ICASSP*, Vol 1, pp. 445–448, Detroit, MI, U.S.A.

Gillick, L. & Cox, S. (1989). Statistical significance tests for speech recognition algorithms. *Proceedings of ICASSP*, pp. 532–535, Glasgow, U.K.

Haeb-Umbach, R., Geller, D. & Ney, H. (1993) Improvements in connected digit recognition using linear discriminant analysis and mixture densities. Proc ICASSP, Vol II, pp. 239–242, Minneapolis.

Hays, W. L. (1963). *Statistics*. London: Holt, Rinehart and Winston.

Hermansky, H. (1990). *Perceptual Linear Predictive (PLP) Analysis of Speech. Journal of the Acoustical Society of America,* **87**, 1783–1752.

Hochberg, M., Renals, S. & Robinson A. (1995). ABBOT: the CUED hybrid connectionist-HMM large vocabulary recognition system. *Proceedings of the Spoken Language Technology Workshop*, pp. 170–178, Austin, TX, U.S.A.

Kershaw, D. J., Hochberg, M. M. & Robinson, A. J. *Context dependent classes in a hybrid recurrent network–HMM speech recognition system.* Cambridge University Engineering Department, Technical Report, CUED/F-INFENG/TR.217.

Kubala, F., Bellegarda, J., Cohen, J., Pallett, D., Paul, D., Phillips, M., Rajesekaran., R. Richardson, F., Riley, M., Rosenfeld, R., Roth, R. & Weintraub, M. (1994). The hub and spoke paradigm for CSR evaluation. *Proceedings of the Spoken Language Technology Workshop*, Plainsboro, NJ, U.S.A.

Lamel, L. F., Gauvain, J.-L. & Eskanazi, M. (1991). BREF, a large vocabulary spoken corpus for French. *Proceedings of Eurospeech* '91, Genoa, Italy.

Lamel, L., Adda-Decker, M. & Gauvain, J.-L. (1995). Issues in large vocabulary multilingual speech recognition. *Proceedings of Eurospeech*, pp. 185–188, Madrid, Spain.

Leggetter, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, **9**, 171–185.

Martin, A. (1995). *Statistical significance tests for speech recognition benchmark tests*. Technical Report, National Institute of Standards and Technology, Gaithersburg, U.S.A.

Moore, R. K. (1988). *Connected digit recognition in a multilingual environment*. RSRE Memo No 4134, Malvern, U.K.

Odell, J. J., Valtchev, V., Woodland, P. C. & Young, S. J. (1994). A one-pass decoder design for large vocabulary recognition. *Proceedings of the Human Language Technology Workshop*, pp. 405–410, Plainsboro, NJ, U.S.A.

Pallett, D. S., Fiscus, J. G., Fisher, W. M., Garofolo, J. S., Lund, B. S., Martin, A., Przybocki, M. (1994). The 1993 benchmark tests for the ARPA spoken language program. *Proceedings of the Spoken Language Technology Workshop*, pp. 15–40, Plainsboro, NJ, U.S.A.

Pallett, D. S., Fiscus, J. G., Fisher, W. M., Garofolo, J. S., Lund, B. S., Martin, A., Przybocki, M. (1995). The 1994 benchmark tests for the ARPA spoken language program. *Proceedings of the Spoken Language Technology Workshop*, pp. 5–38, Austin, TX, U.S.A.

Paul, D. B. & Baker, J. M. (1992). The Design for the *Wall Street Journal*-based CSR corpus. *Proceedings of the 5th DARPA Speech and Natural Language Workshop*, pp. 357–362, San Mateo, CA: Morgan Kaufmann.

Pye, D., Woodland, P. & Young, S. J. (1995). Large vocabulary multilingual speech recognition using HTK. *Proceedings of Eurospeech*, pp. 181–184, Madrid, Spain.
Renals, S. & Hochberg, M. M. (1995). Efficient evaluation of the LVCSR search space using the NOWAY decoder. *Proceedings of ICASSP*, Vol 1, pp. 149–152.
Steeneken, H. J. M. & van Velden, J. G. (1989). Objective and diagnostic assessment of (isolated) word recognizers. *Proceedings of ICASSP*, pp. 540–543, Glasgow, U.K.
Steeneken, H. J. M. & van Leeuwen, D. A. (1995). Multilingual assessment of speaker independent large vocabulary speech recognition systems: the SQALE project. *Proceedings of Eurospeech*, pp. 1271–1274, Madrid, Spain.
van Leeuwen, D. A., van den Berg, L.-G. & Steeneken, H. J. M. (1995). Human benchmarks for speaker independent large vocabulary recognition performance. *Proceedings of Eurospeech*, pp. 1461–1464, Madrid, Spain.
Woodland, P.C., Leggetter, C. J., Odell, J., Valtchev, V. & Young, S. J. (1995). The 1994 HTK large vocabulary speech recognition system. *Proceedings of ICASSP*, Vol 1, pp. 73–76, Detroit, MI, U.S.A.
Young, S. J., Odell, J. J. & Woodland, P. C. (1994). Tree-based state tying for high accuracy acoustic modelling. *Proceedings of the Human Language Technology Workshop*, pp. 307–312, Plainsboro, NJ, U.S.A.

(*Received 13 August 1996 and accepted for publication 3 December 1996*)