4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

LARGE VOCABULARY MULTILINGUAL SPEECH RECOGNITION USING HTK

D. Pye, P.C. Woodland & S.J. Young

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, England.

ABSTRACT

The HTK large vocabulary speech recognition system has been shown to produce state-of-the-art results on American English data. The system uses decision tree state-clustered mixture-density cross-word triphones and statistical N-gram language modelling. Recently, as part of the EC-funded SQALE project, versions of the system have been developed in several European languages. The paper gives an overview of the HTK speech recognition system with American English baseline results, and then describes the progress made in developing British English, French and German versions. The official SQALE evaluation results are reported for each of these four languages and their relative performance is discussed.

10.21437/Eurospeech.1995-45

1. INTRODUCTION

Research in large vocabulary speech recognition has mainly focused on American English. This has been largely due to the availability of suitable corpora and the high visibility of the annual continuous speech recognition (CSR) evaluations organised by ARPA Human Language Technology programme. In the last two years, the HTK large vocabulary speech recognition system has performed very well in these ARPA CSR evaluations [4, 5]. Currently, as part of the EC-funded SQALE project we have been porting the HTK large vocabulary recogniser to a number of new languages.

The purpose of the SQALE project is to extend the ARPA evaluation paradigm to a multilingual context. The project specifies a well-defined speaker independent large vocabulary CSR task in each of four languages: twenty thousand word recognition tasks in American English, British English and French and, to give a comparable OOV rate, a 64k word task in German. Each task consists of 200 test sentences consisting of ten utterances spoken by each of twenty speakers. To support system development for each language, training databases of acoustic data and text data are required. For the American English training data, the WSJ0 corpus is used; for British English WSJCAM0; for French BREF80 and for German the PHONDAT corpus. These corpora were selected to give roughly similar amounts of acoustic training material for all four languages.

This paper first gives an overview of the HTK large vocabulary speech recognition system, and then describes the particular features of the systems developed for American English, British English, French and German. The recognition results on the SQALE evaluation test data are given for each language and briefly discussed.

2. SYSTEM OVERVIEW

This section gives an overview of the baseline HTK speech recognition system as used in these experiments. The system uses state-clustered, cross-word mixture Gaussian triphone acoustic models and a back-off trigram language model.

Each frame of speech is represented by a 39 dimensional feature vector that consists of 12 mel frequency cepstral coefficients, normalised log energy and the first and second differentials of these values. Cepstral mean normalisation is performed on a sentence by sentence basis.

The HMMs are built in a number of stages. First, using a pronunciation dictionary and sentence orthography a phone level label string is generated by Viterbi alignment to choose the most likely pronunciation variants. These labels are used to generate single Gaussian monophone HMMs, which are then cloned for every triphone context that occurs in the training data, and the resulting single Gaussian cross-word triphone HMMs trained.

To obtain good recognition performance, mixture Gaussian densities are required, but for the majority of triphone contexts there is insufficient data to train a mixture Gaussian. Furthermore many of the cross-word triphones needed during decoding do not occur in the training data ("unseen triphones"). To solve both of these problems a tree-based state clustering procedure is used.

A phonetic decision tree is built for every monophone HMM state position to determine equivalence classes between sets of triphone contexts. The tree-growing procedure uses a language-specific set of questions about the immediate phonetic context to repeatedly divide the triphones seen in training into groups. The final clusters contain triphone contexts that are acoustically similar but also have enough training observations for robust estimation of mixture Gaussians. The single Gaussian state output distributions of the members of each class are then tied to each other. By using the decision trees, the tied-state labels needed to synthesise any unseen triphones can be determined. The number of mixture components in each tied-state distribution is incremented using an iterative mixture-splitting and retraining procedure until the optimal number of mixture components is established. Alternatively, if the initial phonetic alignment used unsuitable models (such as aligning French training utterances with British English monophones), then when medium complexity models are attained, it is necessary to retrain from scratch using these new models for re-alignment. Finally, the gender independent models are cloned and separate gender dependent means vectors are trained [4], whilst retaining the gender independent variances.

The system uses a time-synchronous decoder using a dynamically built tree-structured network. Although this can use cross-word acoustic models and a trigram language model in a single pass, it is more efficient for system development to first produce word lattices [5] which compactly store multiple sentence hypotheses. The lattices contain both language model and acoustic information and can be used to either investigate different language models (or language model weightings) or as a word graph for re-scoring with new acoustic models. In the experiments reported here, this lattice approach was adopted. The bigram lattices are generated in an initial pass with gender independent models followed by the application of a trigram language model. The resulting lattices are re-scored using the gender dependent models. As a final stage, for each individual utterance, the hypothesis of greatest likelihood is selected from either the gender dependent or gender independent results.

3. AMERICAN ENGLISH (WSJ0)

The baseline system for American English used 7,183 sentences from 84 of the short-term training speakers in the WSJ0 database. This dataset was chosen to make it roughly comparable in size to those available for the other languages. A set of HMMs with 3,948 clustered states and 8 component mixture Gaussians was built, using the LIMSI 1993 WSJ lexicon. Recognition tests were run using the ARPA November 1993 20k word list and associated trigram language model trained on 37 million words of Wall Street Journal data by MIT Lincoln Labs. The official recognition results on the SQALE test sets are given in Table 1.

Test Set	% OOV	% Word Error	
	rate	bg	tg
us_dev	1.25	14.1	10.6
us_eval	1.46	16.7	13.2

Table 1: % word error rates for the US English system on the official SQALE development and evaluation test sets.

Table 2 shows results achieved on the SQALE test sets using increasingly sophisticated systems. It should be noted that there is further American WSJ acoustic data available for model training (WSJ1 corpus) that contains around 30,000 training sentences. Combining the WSJ0 and WSJ1 corpora results in sixty six hours of speech and training with this makes a 20-25% reduction in word error rate. This quantity of training data also allows more sophisticated quinphone models to be constructed in which the tree clustering procedure can ask questions about the preceding and following two phones. The word lattices generated during the SQALE evaluation (first line) are re-scored with these more sophisticated acoustic models, whilst retaining the same 20k word list and trigram language model. Despite possible errors incurred by rescoring lattices from a less sophisticated system, the results in the second line still show a significant improvement. The third line features results produced by a system using the quinphone models, a 65k word list and a fourgram language model trained on 227 million words – considerably more material than the original 37 million word trigram LM. The 65k word list reduces the OOV rate from 1.25% to 0.15% for us_dev and from 1.46% to 0.39% for us_eval. A combination of these factors explains the improved performance. Finally, the last line consists of this same system with the addition of unsupervised incremental speaker adaptation. This uses maximum likelihood linear regression to estimate parameters of a set of matrices to transform the Gaussian mean vectors. This provided an additional 5% improvement to the results although with longer sentences and more sentences per speaker it is of greater effect. The final system, which gave the lowest error rate in the Nov 1994 ARPA evaluation [5], halves the original word error rate on the SQALE test set.

Model Set	Vocab Size	Grammar Type	% Word Error	
			us_dev	us_eval
si_84 triph	20k	1993 3-g	10.6	13.2
si_284 quin	20k	1993 3-g	8.6	10.3
si_284 quin	65k	1994 4-g	6.0	6.9
si_284 quin†	65k	1994 4-g	5.7	6.3

Table 2: % word error rates on SQALE US development and evaluation test data for different acoustic model sets, vocabulary size and language models. † denotes the use of incremental speaker adaptation.

4. BRITISH ENGLISH (WSJCAM0)

The system for British English used training speech from the WSJCAM0 corpus [1]. This corpus uses the same text prompts as the American WSJ0 corpus, and therefore permits the use of the same language model, and a detailed comparison of performance. A pronunciation dictionary constructed in Cambridge was used with a phone set chosen to account for pronunciation differences between British and American English.

The training set used 7,861 sentences from 92 speakers and an HMM set with 3,494 clustered states and eight component mixture distributions was built. Table 3 shows the results on the SQALE British English test data using the same 20k word list and trigram language model as used for the American English tests above.

Test Set	% OOV	% Word Error	
	rate	bg	tg
uk_dev	1.34	14.5	12.0
uk_eval	1.69	18.3	14.4

Table 3: % word error rates for the UK English system on the official SQALE development and evaluation test sets.

In comparison with American English, the British English results in these experiments were slightly worse. A major factor may be the greater experience of working with American English since the focus of HTK development has been on American English data. It may also be due to the greater variation of accents inherent in British than American English. The various accents and dialects such as Irish and Scottish are all represented with a pronunciation dictionary constructed with Southern British English in mind.

5. FRENCH (BREF80)

The system for French used speech from the Bref80 corpus [3]. The training set included 5,063 sentences from a total of 76 speakers reading texts from the French newspaper *Le Monde*. The language model was derived from 37 million words of *Le Monde* texts. The pronunciation dictionary for French was supplied by LIMSI, and the original lexicon contains only 35 different phones — 10 less than for American English.

One particular feature of French that is not significant in English is *liaison*. Liaison is where normally silent word final consonants are pronounced when immediately followed by a word initial vowel. This improves the fluency of articulation of natural French speech. To accommodate liaison, pronunciation variants of words are added to the lexicon where liaison may occur. Since the vast majority of these liaison consonants are /z/,/t/ or /n/, a preliminary experiment was performed that modelled each of these independently of the normal occurrences of these consonants. This gave slightly better performance than the standard 35 phone set. The use of these liaison models allowed improvements to the phonetic alignments of the training utterances to be made. Where the alignment procedure produces a phone label string which includes a liaison model followed by a non-vowel, then it is erroneous and the liaison model label is deleted. The model sets used during the final evaluation were built using phone label strings edited in this way. Liaison consonant occurrences can be constrained also during recognition, in an experiment the decoder was made to discontinue paths from liaison models to non-vowels since only vowels may legally succeed them. In practice, this last experiment made no difference to performance. The final system configuration used for the evaluation used the 38 phone set including the liaison consonant models, but did not constrain their use during decoding.

The final system built for French contained 2,638 clustered states, significantly fewer than for the English systems as a result of the smaller phone set and less training data. They were created using decision tree questions that consist of translations of relevant American questions supplemented with further French-specific ones collated from French phonetics books. Despite having the least quantity of training data, the relatively small number of clustered states allowed ten-component Gaussian mixture models to be constructed — thus each state is modelled with more complexity than for both American and British English. However, the final French system still contained rather fewer parameters in total.

Test Set	% OOV	% Word Error	
	rate	bg	tg
fr_dev	1.89	19.1	15.6
fr_eval	1.82	18.9	15.1

Table 4: % word error rates for the French system on the official SQALE development and evaluation test sets.

The performance of the system using a 20k trigram grammar on the SQALE French test data sets is shown in Table 4. The performance of the French system is notably worse than for the British and American English systems. Although relatively little time has been spent working with French and less acoustic training material is available, the difference mainly stems from the inherent difficulties of word-level French recognition. The SQALE development and evaluation test sets both have larger OOV rates than the English tests and this is typical of French. Moreover, each OOV word produced on average 1.7 word errors on the fr_dev test set, with the worst example ("*épanouissait*" as "*est pas nous il s' est*") causing six errors.

A significant problem with French is a large homophone rate [2] and many errors on this data appeared to be of this type (i.e. language modelling errors). This is caused by having the same pronunciation for various inflected forms of a word and by having a high number of average pronunciations per word. Furthermore, many of these homophones are frequently occurring (including monophone) function words whilst other examples are multiple word homophones such as mis-recognising "*théoriser leur pratique*" as "*théories et leurs pratiques*". Often homophones can only be disambiguated by using knowledge of grammatical agreement, which implies a need for more sophisticated language modelling with a longer range than trigrams.

6. GERMAN (PHONDAT)

A system for German was also constructed. This used the PHONDAT corpus of phonetically balanced sentences and railway information queries for training of acoustic models. The language model was trained using texts from the *Frankfurter Rundschau* newspaper of comparable size to the WSJ and *Le Monde* Language Models discussed previously.

A characteristic of German is an unusually high OOV rate due to the effects of word compounding. For this reason, a 64k vocabulary was chosen to provide similar lexical coverage to the 20k tests of English and French. The 64k word lexicon provided by Philips has only a single pronunciation per word. This lexicon contained 51 phones which was reduced to 49 by removing the distinction between short and long duration for the two least frequent vowels. Two silence models supplement this set making a total of 51 phones including the glottal stop. This number is considerably larger than for the two English systems and particularly for French. This means that it is possible to need any one of 122,000 logical triphone models (with gender independent models) as compared to 53,000 for French.

The Phondat corpus is of a less phonetically diverse nature than the other corpora used. The speech is also less clean — it was recorded in various environments with different microphones over a prolonged period of time. Even after discarding isolated word utterances, the Phondat corpus provides the largest quantity of training data. The training set is comprised of 15,200 utterances by 155 speakers with some of the longer ones (up to 240 seconds) split to ensure training robustness. Before gender cloning, a model set of 4,268 clustered states of ten Gaussian mixtures components was produced.

Some preliminary experiments were undertaken both with and without explicit modelling of glottal stops. Where it was modelled, the same three state topology as with all other phones was used. Although the difference in performance was not significant, the system configuration used for the evaluation featured glottal stop modelling.

Test Set	% OOV	% Word Error	
	rate	bg	tg
de_dev	2.35	27.5	23.7
de_eval	1.97	21.6	18.7

Table 5: % word error rates for the German system on the official SQALE development and evaluation test sets.

The results of the German system on the SQALE evaluation are featured in Table 5. The performance for German is significantly worse than for the other languages. This is for several reasons. A major factor is the variation in the recording environment of the training data and the resulting mis-match between this and the relatively clean test data. Another reason due to the database rather than a fundamental problem of the language itself is the relative lack of phonetic diversity. This reduces the effectiveness of context dependent modelling. In addition, the 64k vocabulary required to bring the OOV rate down to a similar level to the 20k French and English tasks causes an increase in substitution errors. This effect would be negligible with good language modelling, but the language model used for German is trained on far less data than is desirable given the effects of word compounding. Along with the high OOV rate, word compounding causes an additional problem in that a compound word may be mis-recognised as its constituents or vice versa. Since in German words generally are merged where possible, a relatively high word insertion penalty is necessary to model this behaviour. It is also worth noting that the English recognition tasks follow tradition in being case insensitive tasks, whereas the French and German tasks introduce additional errors due to case sensitivity.

7. CONCLUSIONS

This paper has described versions of the HTK large vocabulary recogniser that have been developed for a number of European languages as well as American English. New languages have language-specific features that need to be incorporated and these were discussed. The relative performance of these systems was reported and some suggestions have been made as to why these differences in performance arise. The US English version has given state-of-the-art performance, and we have shown that using the same techniques good performance can also be obtained on other languages. In particular, the performance on the English and French tasks is fairly similar although German is more difficult to interpret given the nature of the training material.

It was also shown that for American English, a more sophisticated system built using additional acoustic and language model training data can reduce the error rate by ≈50%. It is anticipated that such improvements would be equally beneficial to the other languages considered.

ACKNOWLEDGMENT

This work is in part supported by the EC LRE project 62-058 SQALE. Julian Odell assisted in extending the HTK Large Vocabulary Recognition System to handle the French and German languages.

8. REFERENCES

[1] J. Fransen, D. Pye, T. Robinson, P.C. Woodland & S.J. Young (1994). *WSJCAM0 Corpus and Recording Description.* Technical Report CUED/F-INFENG/TR.192, Cambridge University Engineering Department.

[2] J.L. Gauvain, L.F. Lamel, G. Adda & M. Adda-Decker (1994). Speaker Independent Continuous Speech Dictation. *Speech Communication*, **15**, pp. 21-37.

[3] L.F. Lamel, J.L. Gauvain & M. Eskenazi, (1991). BREF: A Large Vocabulary Spoken Corpus for French, *Proc. Eurospeech'91*, Genoa, Italy.

[4] P.C. Woodland, J.J. Odell, V. Valtchev & S.J. Young (1994). Large Vocabulary Speech Recognition Using HTK. *Proc. ICASSP'94*, Adelaide.

[5] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev & S.J. Young (1995). The 1994 HTK Large Vocabulary Speech Recognition System. *Proc. ICASSP'95*, Detroit.