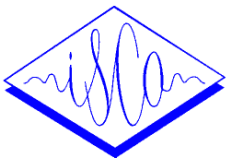

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

THE HMM INTERFACE WITH HYBRID GRAMMAR–BIGRAM LANGUAGE MODELS FOR SPEECH RECOGNITION

G.J.F. Jones, J.H. Wright & E.N. Wrigley
Centre for Communications Research,
University of Bristol,
Queens Building, University Walk,
Bristol BS8 1TR, United Kingdom
Tel: +44 272 303727 Fax: +44 272 255265

ABSTRACT

This paper investigates the interface between the HMM pattern matcher and language models for speech recognition. A number of parameters are found to be important to the optimal performance of speech recognition systems. We consider the selection of match factors between these stages in the recognition process and their relationship with the width of the beam search. We develop improvements to a hybrid language model consisting a probabilistic context–free grammar and a bigram able to process both grammatical and non-grammatical speech input. Finally, consideration is given to the development of an algorithm for dynamic variation of the beam search pruning factor based on the ambiguity of the data input.

10.21437/ICSLP.1992-72

1 INTRODUCTION

In recent years much research has been undertaken in the development of acoustic speech models and Markov language models for speech recognition. Increasing attention is now being given to the ideation of alternative language models and their incorporation into speech recognition systems. Typical examples in this category are those based upon probabilistic grammar models and efficient parsers able to process ambiguous input. Comparatively little work has been released regarding the interaction between the stages of the recognition process. The purpose of this paper is to examine several of the issues that emerge in this area, these include the choice of the match factor between the acoustic pattern matcher and the language model(s) and selection of the pruning factor in the beam search.

We examine the interaction between the matching factor and the width of the beam search and extend this to propose application of a data driven width adaptable beam search. The objective of this method is to allow maximum propagation of hypotheses whilst working within the confines of a finite amount of memory. This is particularly important for the probabilistic grammar model where the number of paths is dependent not only on word sequence but also on grammatical derivation.

In this paper these are investigated in the context of a hybrid language model composed of a bigram model and a probabilistic context-free grammar (PCFG) model operating in parallel [1].

2 INTERACTION OF INTERFACE PARAMETERS

In order to achieve optimal performance of a speech recognition system a number of parameters must be appropriately selected. Examples of these parameters include the pattern matcher match factor (PMMF) between the pattern matcher and the language model, the pruning factor (PF) controlling the width of the beam search and word insertion and deletion penalties. Whilst these latter examples are important they are not relevant in our experimental set-up and will not be considered further here. Direct analytical estimation of the parameters under examination is not possible which means that it is necessary to tune them recursively using a section of training data for optimal system performance. A method to achieve this for a number of these parameters using an N-Best approach is described in [2]. The method to be described here is similar and adjusts the parameters to achieve optimal performance of an experimental recognition system incorporating our hybrid language model.

The objective in speech recognition is to find the word sequence w which maximises the posterior probability $P(w/a)$ of the word sequence given the observed acoustic features. Equation 1 shows the relationship between $P(w)$, the *a priori* probability of the word sequence w from the language model and $P(a/w)$, the conditional probability of the acoustic channel output string a given the word sequence w.

$$P(w/a) = P(w)P(a/w) \tag{1}$$

The PMMF as described in [3] balances the relative weights given to the language model and pattern matcher in determining the most likely sentence output. This is described by equation 2 where h is the PMMF chosen to optimise recognition results.

$$P(w/a) = P(w)(P(a/w))^{1/h} \tag{2}$$

It should be clear from consideration of equation 2 that the variation in h will change the dynamic range of the values of $(P(a/w))^{1/h}$ for different word sequences w. In consequence the dynamic range of the different partial paths propagated by the system during recognition will also be changed. In order to preserve identical sets of partial paths for different PMMFs the appropriate PF would also have to be selected. However, the variation in dynamic range will clearly be non-linear and so variation in the linear PF will not be able to compensate for this and we will potentially produce different sets of partial paths for different PMMF and PF combinations. The objective of the training must be to select a pairing which optimises performance for a particular language model given the practical constraints of the system such as memory and real-time processing power.

3 CHOICE OF OPTIMAL PARAMETER VALUES

The objective of the speech recognition system is to correctly identify the intended input word string. Selection of optimal system parameters will maximise the number of correctly identified words and as such is an important aspect of system specification. The approach taken here to parameter tuning is to use a gradient method. The percentages of words and sentences correctly identified from a training set by the system are used recursively as utilities to drive the system towards maximal performance by tuning the values of the PMMF and the PF.

4 HYBRID LANGUAGE MODEL

4.1 Language models as knowledge sources

Language models can be viewed as knowledge sources (KSs) to assist with the recognition process. The different representations of observed spoken language production of the different models means that incorporation of more than one language model in a single recognition system can lead to improved performance. At present there are two approaches to the incorporation of multiple language models into a recognition system. The first operates the models in parallel coupling their outputs to choose the most likely sentence hypothesis, the hybrid model used here is of this type. The second approach is to apply the language models in successive stages beginning with computationally less expensive ones, such as a bigram, to reduce the search space and continuing to more expensive ones, for example a probabilistic parser. The latter is best demonstrated by BBN BYBLOS speech recognition system [2].

4.2 Justification of the hybrid approach

The successive application of the KSs will for an equivalent task be computationally less expensive than simultaneous application of the different KSs. However, we are considering the development of more advanced interactive language models where the search process is guided by all available KSs. In this approach the syntactic derivation is used to guide the process as well as the likelihood of word sequence. The hybrid model as described here is the first stage of the process. An extension of this to incorporate a first-order dependence model is described in [4].

The hybrid works by switching automatically between the most likely hypothesis generated by the PCFG and the bigram on the grounds of the highest score. The parameters of these language models must be separately tuned for optimal performance. Moreover, since these models depend on fundamentally different algorithms there is no reason why the optimal parameters for the two systems should be the same. In the case of the PF this does not matter but for the PMMF this presents a problem for the hypothesis selection switch in the hybrid model. The hypothesis score is dependent on the PMMF equation 2 and thus use of different PMMFs for the language models will unbalance the score based automatic switch. A solution to this problem is described in the next section along with an additional score balancing technique.

4.3 Balancing of different optimal PMMFs

The hypothesis score for each language model is composed of a product of the factored pattern matcher scores and those derived from the language model. The optimal path chosen in each case is dependent on use of the appropriate combination of PMMF and PF. In order to balance the hypothesis scores we rescore one of the paths as if it operated with the same PMMF as the other one. This is a simple procedure based on the relationship shown in equation 3. h_{pcfg} and h_{bi} are the PMMFs associated with the PCFG and bigram models respectively. $1/h_{bal}$ is the difference between the reciprocal PMMFs and can be used to rescore the path hypothesis associated with the bigram model.

$$(P(a/w))^{1/h_{pcfg}} = (P(a/w))^{1/h_{bi}} * (P(a/w))^{1/h_{bal}} \qquad (3)$$

Before application to the hypothesis selection switch the score of the most likely hypothesis sequence from the bigram model is multiplied by the product of the pattern matcher scores for each word of this sequence raised to $1/hbal$, $(P(a/w))^{1/h_{bal}}$.

This technique is applicable to any situation where a combination of language models is used with different PMMFs. A further modification to our simple automatic switch involves additionally factoring the score to take account of differences, between the hypotheses scores for the two systems, associated with the effect of different PMMFs on the operation of the language model component. For both language models the optimal path is theoretically dependent on the dynamic range of the scores associated with the pattern matcher. For example, consider two parallel bigram models operating with different PMMFs at their front end. For the system with the PMMF leading to a larger dynamic range in pattern matcher scores a path associated with lower bigram transition scores may dominant those associated with larger bigram transition scores but comparatively lower pattern matcher scores. One of these other paths may score best in the system associated with the lower dynamic range in the pattern matcher scores.

The second balance phase is to additionally multiply the score of the sequence hypothesis associated with the optimised bigram system by the product of the difference between the bigram word transition scores of this path and that observed for an identical system operated with the same PMMF as used for the PCFG language model. This second factor is calculated as the product of the factors shown in equation 4.

$$RescoreFactor(t) = a_{ij}^{parse}(t)/a_{ij}^{bigram}(t) \qquad (4)$$

This is computationally unattractive since it necessitates running a second bigram algorithm in parallel with the first to know what the relevant transitions would have been. Despite this it does seem to lead to improved results for non-grammatical input. Of course, if the two paths are identical there will of course be no modification of path score.

5 DYNAMIC CONTROL OF BEAM SEARCH WIDTH

Data driven organisation search space of the path hypotheses is described elsewhere [5]. This approach allows maximum use of available computer memory particularly when coupled with reuse of nodes made redundant when paths are pruned out. These factors are particularly important for a more complex language models such as a PCFG where the complexity of syntactic derivation leads to a much increased number of hypotheses. In this case it is possible to use node sharing between hypotheses to save space [6] but there will necessarily be a finite limit to the number of nodes that can be used as any point. In order to use this space most effectively it is important to select the correct combination of PMMF and PF.

The number of paths propagated can be increased to a finite limit by appropriately modifying the PF. Increasing the number of paths means that the probability of the correct path being pruned out is reduced and as such is an attractive option. However, the greater number of paths increases computational expense and more importantly we risk overflowing the stack of path hypotheses. A recovery procedure can of course be developed for this situation but this is not an ideal solution to the problem.

In the case of the PCFG component of the hybrid language model it is observed that the rate of node expansion is much higher for non-grammatical speech input ie that for which the correct decoding falls outside the scope of the PCFG. The non-grammatical structure of these sentences is shown by greater ambiguity in the hypotheses stack and hence the number of partial paths propagated is higher in such cases. In order to run a system capable of processing both grammatical and non-

grammatical input it is necessary to use a low beam width to facilitate processing of the non-grammatical input. However, optimal performance over the grammar sentences is likely to be observed with a wider beam search. The performance for grammatical input may thus be impaired by failure to fully utilise the potential search space. Use of a wide beam search with a recovery procedure to prevent overflow (which would be invoked by the non–grammatical input) is a possibility but this would be undesirable.

It would seem reasonable to try to let the data ambiguity encountered in the form of the rate of expansion of the hypothesis stack control the tightness of the beam search. If a proportionally low number of hypotheses are generated we can proceed with a wide, and possibly increasing, beam width maximising our chance of propagating the correct path. If the stack is increasing quickly and we are likely to overflow the stack the width of beam can be decreased possibly increasing later in the sentence if the rate of node expansion slows sufficiently, the implementation of node reuse means that the number of free nodes may naturally increase when certain paths are pruned out. In this way we should be able to obtain near optimal performance for all input without the necessity for expensive recovery procedures and hopefully achieve better performance in highly ambiguous sections of input than would otherwise be possible. Factors which might be considered in such an algorithm are the length of sentence hypotheses, maximum permitted sentence length, the number of nodes used at present, the maximum number of nodes available and the rate of expansion of number of nodes.

In addition to these factors it should be possible to estimate the maximum expected expansion for the next word using the number of currently active paths, the number of likely next words associated with them and a look–ahead of the likelihood of these words having occurred using a technique similar to that described in [7]. Further, using the data from the patten matcher a coarse beam search of the most likely hypothesis associated with each next word can be used to determine the maximum number of paths that will survive the beam search. This gives a more accurate measure of the maximum number of paths that might actually be propagated assuming that no path merging is possible. If this figure is not appropriate the coarse beam search can be repeated with a different PF. It might also be useful when the number of free nodes becomes small to prune back over several previous words with a tighter PF to release some more nodes since we know that if the stack is nearly full then the beam search has been too generous for earlier partial hypotheses.

The most appropriate methods of combination of these factors is not obvious. Probably the most intuitive approach is to use a scaled combination of them, however, this is unattractive since, once trained, we will have a fixed system and ideally the ambiguity could be used more powerfully to control events.

6 EXPERIMENTAL RESULTS

The experimental work was carried out on a PC based DSP32C Telephony board. The example considered is simple, we have a vocabulary of 100 words and the speech is isolated, however, it is possible to observe the effects of the algorithms.

A 7-state HMM of each word was enroled on the speech board. The training was limited to only 5 utterances of each word and the system is speaker dependent. The probabilistic grammar used for the experiments consisted of 230 rewrite rules governing the 100-word vocabulary.

A bigram table was obtained from a corpus of text generated at random from the PCFG, using the rule probabilities. The bigram model was smoothed using the standard Good-Turing technique [8]. The Markov model is thus representative of the language governed by the grammar and is able to compete effectively with the grammar in the hybrid.

Two sets of sentences were generated. The first was generated at random from the PCFG, again using the rule probabilities. These were not part of the set used to train the bigram. The second set was generated at random from a bigram model trained using the Held-Out method [8], but with all grammatically correct sentences filtered out. This second set is therefore intended to activate the bigram model. Both sets contained nearly 100 sentences.

The first experiment investigated the relationship between selection of the PMMF and the choice of PF for the PCFG. Table 1 shows the for 3 sets of grammar compatible sentences. In all results a test sentence which produces no output hypothesis is treated as incorrect regardless of pattern matcher performance.

PMMF	Target	Pruning Factor			
		0.1	0.05	0.01	0.005
60	Words	82.84	86.70	89.61	93.33
	Sentences	69.64	71.39	74.84	78.26
70	Words	85.24	89.16	92.83	95.47
	Sentences	69.83	71.99	78.26	81.03
80	Words	85.88	87.74	95.19	–
	Sentences	70.80	71.99	81.03	–
100	Words	88.03	90.49	95.52	–
	Sentences	70.43	73.35	79.06	–

Table 1: Relationship between PMMf and PF (% correct)

From these results it is observed that the best results can apparently be observed with any correctly matched set of PMMF and PF. In effect we are optimising the PMMF to match a particular PF. This will be given further consideration in the next section.

For each PMMF the apparent improvement in performance for increasing beam width is predominantly due to increased number of test sentences for which a parsed output was generated.

The second experiment shows the effect on the performance of the hybrid system for variation of PF for a chosen pairing of PMMFs. The PMMF for the bigram was selected to give optimal performance and that for the PCFG selected to be different to demonstrate the effectiveness of the matching parameters. Again these results, shown in table 2 are observed over three sets of grammar compatible and non–compatible sentences.

Performance of the system over the grammar compatible sentences is generally observed to improve for increasing beam width. In the cases where the parser fails to produce an output the bigram hypothesis is of course invoked. As the parser provides more outputs for increased beam width it is selected more often and performance improves. The reverse is observed for the grammar incompatible sentences, we know that if the parser is invoked for these sentences its output will be wrong. Therefore as the parser output is invoked more often for increased beam width performance declines. However, if the grammar is representative of the recognition application the bigram model is only operating as a back-up and we are justified in optimising for performance over the grammar sentences.

Table 3 shows the results for three different PMMF and rescaling set-ups, both language models with the same PMMF and optimised separately first with rescaling of pattern matcher scores and second additional rescaling of bigram scores. These tech-

	Target	Pruning Factor		
		0.1	0.05	0.01
Pattern matcher	Words	72.29	72.29	72.29
	Sentences	10.47	10.47	10.47
PCFG model	Words	88.03	90.50	95.52
	Sentences	70.43	73.35	79.05
Bigram model	Words	92.65	92.65	92.65
	Sentences	61.69	61.69	61.69
Hybrid model	Words	95.18	95.47	96.27
	Sentences	74.43	76.22	79.06

(a) Grammar compatible sentences

	Target	Pruning Factor	
		0.1	0.05
Pattern matcher	Words	73.12	73.12
	Sentences	24.70	24.70
PCFG model	Words	23.73	27.72
	Sentences	00.00	00.00
Bigram model	Words	91.74	91.74
	Sentences	65.87	65.87
Hybrid model	Words	91.03	90.70
	Sentences	63.63	63.08

(b) Grammar incompatible sentences

Table 2: System performance for different PFs (% correct)

Test Set	Target	Hybrid balance		
		PMMFs equal	PM rescaling	PM & bi rescaling
PCFG model	Words	96.27	96.27	96.27
	Sentences	79.06	79.06	79.06
Bigram model	Words	90.00	90.60	90.78
	Sentences	60.40	63.08	63.08

Table 3: Hybrid model balancing (% correct)

niques are shown to be effective but in this example only marginally.

Table 4 show the average maximum number of stack graph and parse forest [6] nodes used for PF 0.05 for PMMF 70 and 100 for both grammar compatible and incompatible sentences.

As expected the number of nodes needed in each case for the grammar incompatible sentences is significantly larger than for the grammar compatible sentences. This means that for the grammar sentences the system must operate with a large memory redundancy for a fixed PF.

Ave Max No of nodes		PMMF	
		70	100
Grammar sentences	stack graph	18.02	23.11
	parse forest	43.21	53.06
Non-grammar sentences	stack graph	24.77	38.34
	parse forest	56.62	84.58

Table 4: Average max no of Stack Graph and Parse Forest nodes

7 LANGUAGE MODEL MATCH FACTOR

The apparent similar performance for the pairings of PMMF and PF means that we have no means of fixing the dynamic range of the pattern matcher scores. However, with reference to equation 2 the dynamic range of $(P(a/w)^{1/h})$ is dependent on h which affects the relative weighting of the pattern matcher and language model. It is reasonable to consider that a further optimisation might be obtained by completing a balance between the two stages by varying the dynamic range of the language model. The score of the language model component would then be $(P(a)^{1/k})$ where the parameter k is referred to as the Language Model Match Factor (LMMF). The posterior probability $P(w/a)$ is given by equation 5.

$$P(w/a) = (P(w))^{1/k}(P(a/w))^{1/h} \quad (5)$$

8 SUMMARY

The optimisation of interface parameters for an automatic speech recognition system comprising an HMM pattern matcher and a hybrid language model has been examined. In particular, consideration has been given to the balance of the influence of the pattern matcher and the language model in derivation of sentence hypotheses.

9 ACKNOWLEDGEMENTS

This work is financially supported by the U.K. Science and Engineering Research Council and by Ensigma Limited, Chepstow, U.K. Thanks are extended to the staff of Ensigma Limited for help with HMM software and DSP hardware and also to the Centre for Communications Research for provision of computing facilities.

REFERENCES

[1] G.J.F. Jones, J.H. Wright, E.N. Wrigley, M.J. Carey and E.S. Parris, *"Isolated-word sentence recognition using probabilistic context-free grammar"* , Proceedings of EUROSPEECH-91, pp487-489 (1991)

[2] R. Schwartz, et al, *"New uses for the N-best sentence hypotheses within the BYBLOS speech recognition system"* , Proceedings of ICASSP-92, pp(1-(1-4)) (1992)

[3] L.R. Bahl, R. Bakis, F. Jelinek and F. Mercer, *"Language-model/Acoustic channel balance mechanism"* , IBM Technical Disclosure Bulletin 23(7B), pp3464-6465 (December 1980)

[4] J.H. Wright, G.J.F. Jones and E.N. Wrigley, *"Hybrid grammar-bigram speech recognition system with first-order dependence model"* , Proceedings of ICASSP-92, pp(1-(169-172)) (1992)

[5] H. Ney, D. Mergel, A. Noll and A. Paeseler, *"Data driven organisation for continuous speech recognition"* , IEEE Transactions on Signal Processing, Vol. 40 No. 2, pp272-281 (February 1992)

[6] E.N. Wrigley and J.H. Wright *"Computational requirements of probabilistic LR parsing for speech recognition using a natural language grammar"*, Proceedings of EUROSPEECH-91, pp761-764 (1991)

[7] R. Haeb-Umbach and H. Ney, *"A look-ahead search technique for large vocabulary continuous speech recognition"* , Proceedings of EUROSPEECH-91, pp495-498 (1991)

[8] K.W. Church and W.A. Gale *"A comparison of enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams"* , Computer Speech and Language, vol 5, pp19-54 (1991)