LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION OF WALL STREET JOURNAL DATA

X. Aubert [1] *C. Dugast* [1] *H. Ney* [2] *V. Steinbiss* [1]

[1] Philips GmbH Research Laboratories Aachen, P.O. Box 1980, D-52021 Aachen, Germany
[2] Lehrstuhl für Informatik VI, RWTH Aachen, D-52056 Aachen, Germany

ABSTRACT

We report on recent developments of the Philips large vocabulary speech recognition system and on our experiments with the Wall Street Journal (WSJ) corpus. A two-pass decoding has been devised that allows an easy integration of more complex language models. First, a word lattice is produced using a time synchronous beam search with a bigram language model. Next, a higher-order language model is applied to the lattice at the phrase level. The conditions insuring the validity of this approach are explained and practical results for trigram demonstrate its usefulness. The main system development stages on WSJ data are presented and our final recognizers are evaluated on Nov'92 and Nov'93 test-data for both 5K and 20K vocabularies.

1. INTRODUCTION

In this paper, we report on recent developments of the Philips large-vocabulary continuous-speech recognition system that have been triggered off by working on Wall Street Journal (WSJ) data.

The Philips dictation system for phoneme-based large-vocabulary continuous-speech recognition relies on an integrated statistical framework and was recently described in details [1]. It has been successfully applied to speaker-dependent (SD) dictation tasks in German with 10K to 20K words. Last year, a closely related version of this system has been thoroughly evaluated on the 1000-word DARPA RM task [2], for speaker-independent (SI) American English. Characteristic features of our approach include continuous mixture densities, linear discriminant analysis, Viterbi-training, one pronunciation per word and within-word triphones.

The ARPA WSJ corpus [3] consists (a.o.) of samples of read texts drawn from the Wall Street Journal publications and provides training and test material for SI continuous speech recognition in American English. Vocabulary sizes are typically ranging from 5K (closed) to 20K (open, i.e. out of lexicon words do appear in the test sentences). In addition, standard bigram and trigram language models have been supplied by D. Paul from MIT Lincoln Lab.

The paper is organized as follows. After a brief presentation of the acoustic-phonetic modeling in section 2, we focus on the decoding procedure. The search has been extended to fulfill the WSJ-task requirements that appear drastically increased with respect to RM. A two-pass decoding strategy has been devised. In the first pass, a word lattice is produced using a time synchronous beam search with a tree-organized lexicon and a bigram language model. This step simultaneously provides the best bigram-scored sentence hypothesis. In the second pass, the trigram language model is applied to the lattice at the phrase level for extracting the best trigram-scored sentence hypothesis. This new search strategy is explained in section 3 and the conditions insuring the validity of this approach are also clarified. Besides, the efficiency of the beam search has been greatly improved by distributing the language model scores across the lexical tree, using both unigram and bigram probabilities as shown in section 3.4. In the last two sections, we present a broad sample of results obtained on the WSJ task. First, the main development stages are explained on a step-by-step manner and next, our final systems are evaluated on the Nov'92 and Nov'93 test-data for both 5K and 20K vocabularies. In particular, the influence of the training size and of the language models on the error rate is discussed.

2. ACOUSTIC-PHONETIC MODELING

Our standard acoustic analysis [1] is applied at a 10 ms frame-rate. The log-spectrum energies are normalized for each sentence by spectral mean subtraction. Linear discriminant analysis is performed at the HMM state level [4] from automatically segmented training data, the output vector being reduced to its 35 "first" components. Mixtures of continuous Laplacian densities are estimated state-specifically with a pooled absolute deviation vector using Viterbi approximation and data-driven splitting [2]. One single pronunciation is taken for each vocabulary word and contextual dependencies are captured with word-internal triphones only. These were selected based on their occurrence frequency in the training script. Neither multiple pronunciations nor across-word models have been considered.

3. SEARCH STRATEGY

3.1. Bigram Decoding with Tree Lexicon

Time-synchronous beam search has been successfully used for several years at Philips [5] to handle 10,000 and more vocabulary words. A significant reduction of the computational effort has been achieved by structuring the pronunciation lexicon into a tree, the active search space being dynamically constructed with a list organization [6].

Still, resorting to a tree-organized lexicon has some rather contrasting effects for time-synchronous breadth-first

0-7803-1775-0/94 $3.00 © 1994 IEEE

strategies. On one hand, it is observed that the bulk of the decoding effort occurs in the first two phonemes of each word and this explains the highly beneficial impact of sharing the initial phoneme sequences that are common across all lexical entries. On the other hand, as the word identities are only known at the tree leaves, it is no longer possible to perform an early language model (LM) recombination [9], and even for a bigram LM, separate tree copies have to be held depending on the predecessor word. Accordingly, the inclusion of the bigram probabilities has to be postponed until the word-end which prevents from taking full advantage of the LM predictive properties to promptly prune unpromising word candidates (see 3.4. for an efficient remedy).

In spite of this adverse side-effect, the lexical tree organization has proven to be very advantageous in practice for large vocabulary tasks. A detailed discussion with experiments is given in [6]. So far however, this algorithm was used with a bigram LM and only applied to speaker-dependent recognition based on context-independent phoneme models, the latter being particularly beneficial to the tree compression effect.

Hence, two questions arose when tackling WSJ data: first, how could the trigram language model be best integrated into the decoding process and second, how would the search cost increase when the lexical tree is made of context-dependent subword units like triphones (a must for accurate SI recognition) ? As explained in the following sections, our solution consists of a two-pass decoding, the interface being an acoustically-scored word lattice. This choice offers the advantage that a complex language model can be exploited in a post-processing step without increasing the complexity of the acoustic search.

3.2. Generation of Word Lattice

A word lattice can be efficiently generated with only minor modifications of our time-synchronous beam search algorithm based on a tree lexicon. It essentially amounts to collecting the information about word-endings as they occur in the course of the left to right decoding process. This first pass simultaneously provides the best bigram-scored sentence hypothesis, the lattice overhead being virtually negligible in terms of CPU time.

As opposed to the word-graph generation technique presented in [7], here we take full advantage of the bigram LM to constrain the lattice, without requiring any further optimization or pruning stage. More precisely, our analysis relies on the assumption that the position of a word boundary depends only on the word pair under consideration and not on further predecessor words. This simplification has been successfully used by BBN in their Word-Dependent N-Best algorithm [8] and is also known as the "word-pair approximation" [9].

Therefore, in the present study the lattice is defined as a time-structured list of word hypotheses consisting of word identity, start- and end-time, acoustic score and predecessor word identity. It has to be stressed that the collection of word-ending information is done before the bigram LM recombination takes place, to preserve as much as possible different word sequences for subsequent use with a higher-order LM.

The computational complexity of this first pass is nearly identical to that of our bigram beam search, the efficiency of which having been further improved by a new handling of the LM probabilities.

Concerning the use of context-dependent models in this acoustic pass, it appears that the influence of triphones on the lexical tree structure is largely compensated by a more focused search consecutive to more detailed phone models (cf. section 4).

3.3. Trigram Rescoring in Lattice

In this second pass, the trigram language model is applied to the lattice at the phrase level. More precisely, the acoustic probabilities of the word hypotheses are combined with the trigram probabilities taking account of the predecessor-word as computed in the first pass. Searching for the optimal rescoring still proceeds time-synchronously and requires a Dynamic Programming (DP) recursion taking account of all time and predecessor constraints contained in the lattice [9]. The final output is the best trigram-scored sentence hypothesis under the lattice restrictions.

The optimality of this procedure (in the Viterbi sense) is preserved only under the following two conditions: the word-pair approximation for the position of a word boundary has to be valid and next, the beam used for generating the lattice must be wide enough to keep enough phrase hypotheses for subsequent trigram rescoring.

In practice, this algorithm appears to work well with relatively modest lattice densities as shown by the results included in sections 4 and 5. The computational costs are quite small since this second pass does not require any further acoustic scoring at the state level. This follows from the word-pair assumption which implies that the word boundaries have already been optimized in the first pass. Moreover, a careful list organization allows to achieve great efficiency (without requiring the cashing of the LM scores) to such an extent that the trigram rescoring represents only a few percent of the main bigram decoding CPU time.

3.4. Improved LM-based Beam Pruning

As explained in 3.1, the tree lexicon delays the application of the language model by one word as compared to the situation of a linear lexicon. Word identities are only known at the tree leaves so that the true bigram probabilities can only be incorporated at word-endings. On the other hand, when the word identity is known from its first model state, the LM scores can be immediately added leading to a more focused search space. Indeed, significantly more partial hypotheses can be safely eliminated from the beam and the search costs are reduced in proportion.

To alleviate the delaying effect of the tree lexicon, the following solution has been conceived [10]. The general idea is to use at each tree node some conservative estimate of the LM score relevant to all possible word continuations and to perform this "smearing" process on an incremental manner from the tree root to the leaves.

The application of this principle is complicated by the existence of tree copies that depend on the predecessor word and are essential for the DP optimality. Therefore, only the unigram LM information has been incrementally distributed across the lexical tree, i.e. the prior probabilities of each word, these values being always available as part of an N-gram LM. This operation can be easily done in a pre-processing step the details of which are given in [10] and leads to a negligible overhead during search. Last, a correction step is necessary when reaching the word-ends: the cumulated partial unigram scores have to be withdrawn while the true bigram log-probabilities are added. Although based on the unigram LM part only, this scheme appears to be remarquably efficient by strongly reducing the active number of hypotheses and moreover, by allowing the use of a smaller beam width.

4. SYSTEM DEVELOPMENT

We now present some intermediate results illustrating the main development stages of our "WSJ systems". Unless specified, all (non-stressed) pronunciations were taken from the original Dragon lexicon, training was performed on the 84-speaker corpus, and recognition was done with a bigram LM. Experiments have been run on several WSJ0 development sets with non-verbalized punctuation and for various vocabularies (5K closed, 20K open and closed).

The first step shows the reduction of the error rate when using phone models that capture gradually more contextual dependencies.

Table 1. Influence of Contextual Units (Dev-5K)
WER=Word Error Rate (Del+Ins+Sub)

#Units	Type	#Densit.	WER %	Progress
43	Monophones	33 K	18.5	Ref.
772	Left-Diphones	37 K	15.2	-18 %
"	"	115 K	14.0	-24 %
736	Triphones	73 K	13.1	-30 %

Our interest for left-diphones stemmed from the fact that they preserve the lexical tree structure of monophone transcriptions as opposed to triphones. However triphones clearly lead to more accurate models. Therefore, next stages have been running with the set of 736 triphones occurring more than 150 times in the WSJ0 training script.

The second step concerns the effect of Linear Discriminant Analysis (LDA) and of Gender-Dependent (GD) estimation that has been used both for the LDA transform and for the mixture parameters as well [2]. Our experiments are summarized in the following table.

Table 2. LDA and Mixture GD Estimation (Dev-5K)
GI=Gender-Independent, over M & F Speakers

LDA?	Gender	#Densit.	Gender	WER %	Progress
NO	-	2 * 65 K	M/F	12.4	Ref.
YES	M/F	2 * 65 K	M/F	11.3	-9%
NO	-	1 * 133 K	GI	12.4	Ref.
YES	GI	1 * 139 K	GI	10.6	-15%
YES	GI	2 * 123 K	M/F	9.7	-22%
YES	GI	" + 139 K	M/F/A	9.4	-24%

The best configuration is achieved with a single GI LDA transform followed by GD mixture estimation. Adding the uni-sex models to the male- and female-specific models only brings a further marginal improvement.

The next table shows that slightly but consistently better results are achieved with the LIMSI lexicon by comparison with the Dragon lexicon.

Table 3. LIMSI versus DRAGON Lexicon (Dev-5K)
Gender-Dep. : GD2=M/F, GD3=M/F/All

Lexicon	GI WER	GD2 WER	GD3 WER
Dragon	10.6%	9.7%	9.4%
LIMSI	9.9%	9.4%	9.0%

The third step involves the LM rescoring technique in word lattices generated with a bigram LM and shows the error reduction when going from bigram to trigram.

Table 4. From Bigram to Trigram Language Model
WER=Word Error Rate (Del+Ins+Sub)

Corpus (Mode)	BIGRAM		TRIGRAM		Relat.
	WER %	Perp.	WER %	Perp.	Reduct.
Dev 5K Closed[1]	10.6	110	7.9	62	-25%
Dev 5K Closed[2]	9.7	110	7.3	62	-24%
Dev20K Closed[1]	18.8	242	15.1	155	-20%
Dev20K Open[1]	19.9	205	16.4	136	-16%

[1] and [2] refer resp. to GI and GD (M/F) mixture modeling both after GI LDA (cf. Table 2).

When switching from bigram to trigram, both the test-set perplexity and the error rate are significantly reduced. It is interesting to observe that the error rate decreases like the square root of the bigram-to-trigram perplexity ratio, at least for our closed-vocabulary experiments. Note however that this is nothing but a rule of thumb deduced from limited experimental data. For open vocabulary, the interpretation is complicated by the presence of out-of-vocabulary words that constitute about 2% ot the test words and give rise to additional insertion errors. Moreover, in this case the perplexity measures are no longer that reliable.

Last, we give a few figures concerning some characteristic properties of the bigram search cost:

- When using triphones, the number of arcs in the first two generations of the tree lexicon are multiplied by resp. 6 and 2 with respect to the monophone tree: However, due to the improved precision of the triphone models, the average number of state hypotheses in the beam search is actually smaller !
- The improved LM-based pruning reduces the average number of hypotheses by a factor of 3 to 5 compared to the original handling of bigram scores.
- When the vocabulary grows from 5K to 20K words, the average number of hypotheses increases by not more than 50% owing to the lexical tree.

5. EVALUATION RESULTS

5.1. System Description

Two systems have been set-up differing mainly in the number of triphones and the amount of training data. In each case, the mixture density parameters have been estimated gender-dependently with respectively male, female and unisex models. During decoding, the word sequence achieving the highest cumulated probability has been taken for the recognized sentence. The table below gives the main system characteristics.

Table 5. Main Characteristics of Evaluation Systems

NAME	TRN-data	TRN-time	Lexicon	#Mono+#Tripho.
SI-84	WSJ0	≈15 hours	LIMSI	45 + 740 Tri
SI-284	WSJ0+1	≈80 hours	Dragon	43 + 1864 Tri

As indicated, the LIMSI lexicon has been used in the first system (trained over 84 speakers) while the Dragon lexicon has been used in the second one. LDA has been applied gender-independently based on 84 speakers (WSJ0 TRN-data). The average number of Laplacian densities per state is about 45. The official bigram and trigram LM have been employed without any modification. These systems have been tested on the evaluation sets of Nov'92 and Nov'93 containing recordings from resp. 8 and 10 new out-of-training speakers.

5.2. Closed 5K Lexicon

The two systems have been running on each 5K evaluation set with standard benchmark conditions, i.e. not using any side information about the utterances. Results for bigram and trigram LM are summarized in the following table.

Table 6. Evaluation Results for 5K Closed Lexicon
P = Test Perplexity for Bigram and Trigram
WER % given as: (Del/Ins) Tot=Del+Ins+Sub

SYS	5K NOV'92 WER %		5K NOV'93 WER %	
	Bi (P=111)	Tri (P=57)	Bi (P=106)	Tri (P=62)
SI-84	(.7/1.) 7.4	(.4/.7) 5.0	(3.2/1.) 12.7	(2.5/1.) 9.4
SI-284	(.5/.8) 6.1	(.4/.5) 4.3	(2.8/.8) 9.2	(1.8/.6) 6.5

Concerning the LM influence, it might again be observed that the error rates are approximately reduced like the square root of the perplexity ratio when going from bigram to trigram. This represents a recovery of 30% of the errors.

A clear improvement follows when more acoustic models are estimated using more training data. With respect to system SI-84, SI-284 achieves an improvement of about 20% on NOV'92 and 30% on NOV'93. This is attributed to the acoustically more difficult recordings of last evaluation data as might be inferred when considering the perplexities and the error rates of both sets.

5.3. Open 20K Lexicon

Here we present the 20K results obtained with the second system SI-284 trained on 284 speakers.

Table 7. Evaluation Results for 20K Open Lexicon
P=Test Perplexity for Bigram and Trigram
WER given as: (Del/Ins) Tot=Del+Ins+Sub

20K NOV'92 WER %		20K NOV'93 WER %	
Bi (P=205)	Tri (P=139)	Bi (P=221)	Tri (P=143)
(1.1/2.4) 14.	(1./2.1) 11.9	(2.8/1.7) 17.3	(2.5/1.3) 14.9

When going from bigram to trigram, the 20K errors are now reduced by about 15% i.e. somewhat less that could have been expected from the "square root of perplexity ratio" rule of thumb. However, the presence of about 2% of out-of-vocabulary words makes the analysis somewhat difficult as they give rise to additional insertion errors having some "cascade" effect on the trigram scores.

6. CONCLUSION

In spite of a relative simplicity, our algorithms achieve performances that are comparable to those obtained by most advanced systems during the last WSJ Nov'93 evaluation. We expect further progress by implementing multiple pronunciations and across-word triphone models.

REFERENCES

[1] Steinbiss V., Ney H., Haeb-Umbach R., Tran B.-H., Essen U., Kneser R., Oerder M., Meier H.G., Aubert X., Dugast C., Geller D., "The Philips Research System for Large-Vocabulary Continuous-Speech Recognition", *Proc. EuroSpeech*, pp 2125-2128, Berlin 1993.

[2] Aubert X., Haeb-Umbach R. and Ney H., "Continuous Mixture Densities and Linear Discriminant Analysis for Improved Context-Dependent Acoustic Models", *Proc. ICASSP'93*, Minneapolis, MN, pp 648-651, 1993.

[3] Paul D. and Baker J., "The Design for the Wall Street Journal-based CSR Corpus", in DARPA Speech and Language Workshop, Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[4] Haeb-Umbach R., Ney H., "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition", *Proc. ICASSP'92*, San Francisco, CA, pp I(13-16), 1992.

[5] Ney H., Mergel D., Noll A., Paeseler A., "A Data Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition", *Proc. ICASSP'87*, Dallas, 20.10.1-4, 1987.

[6] Ney H., Haeb-Umbach R., Tran B.-H., Oerder M., "Improvements in Beam Search for 10000-Word Continuous Speech Recognition", *Proc. ICASSP'92*, San Francisco, CA, pp 9-12, 1992.

[7] Oerder M., Ney H., "Word Graphs: An Efficient Interface Between Continuous-Speech Recognition and Language Understanding", *Proc. ICASSP'93*, Minneapolis, MN, pp 119-122, 1993.

[8] Schwartz R. and Austin S., "A Comparison of Several Approximate Algorithms for Finding Multiple (N-BEST) Sentence Hypotheses", *Proc. ICASSP'91*, Toronto, CANADA, pp 701-704, 1991.

[9] Ney H., "Search Strategies for Large-Vocabulary Continuous-Speech Recognition", to appear in *Proc. of NATO Advanced Study Institute on Speech Recognition and Understanding*, Bubion, Spain, 1993.

[10] Steinbiss V., "Improvements in Beam Search", to be published.