Improving Speech Recognition Performance via Phone-Dependent VQ Codebooks and Adaptive Language Models in SPHINX-II

M. Hwang **R. Rosenfeld** **E. Thayer** **R. Mosur** **L. Chase** **R. Weide**
X. Huang **F. Alleva**

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

This paper presents improvements in acoustic and language modeling for automatic speech recognition. Specifically, semi-continuous HMMs (SCHMMs) with phone-dependent VQ codebooks are presented and incorporated into the SPHINX-II speech recognition system. The phone-dependent VQ codebooks relax the density-tying constraint in SCHMMs in order to obtain more detailed models. A 6% error rate reduction was achieved on the speaker-independent 20,000-word Wall Street Journal (WSJ) task.

Dynamic adaptation of the language model in the context of long documents is also explored. A maximum entropy framework is used to exploit long distance trigrams and trigger effects. A 10% – 15% word error rate reduction is reported on the same WSJ task using the adaptive language modeling technique.

1 INTRODUCTION

For speech recognition, hidden Markov models (HMMs) with continuous observation densities [1, 2, 3] offer direct modeling of the speech cepstra, unlike the HMMs with discrete observation distributions in which vector-quantization (VQ) [4, 5] errors are unavoidable. On the other hand, continuous HMMs (CHMMs) require heavy and expensive computation. The semi-continuous HMM (SCHMM) approach [6, 7] reduces VQ errors by using the top few best-matched VQ codewords instead of only the best one. It also achieves much cheaper computation by tying all Markov states to the same set of VQ densities. To relax the density-tying constraint in order to approach the performance of CHMMs, we incorporate SCHMMs with phone-dependent (PD) VQ codebooks, in which Markov states in all triphones that represent that same phone share the same set of VQ codebooks or densities [8, 9].

The second issue addressed in this paper is the incorporation of long-distance language models into our search algorithm. We report a flexible interface that supports such models, and give experimental results for two examples: a trigram language model and an adaptive language model that is based on the maximum entropy (ME) principle [10, 11]. The ME framework allows inclusion of multiple sources of statistical language information, including conventional bigrams and trigrams, longer distance bigrams and trigrams, and word-pair triggers. The resulting long-distance language model is applied in the final pass of our three-pass decoder [12].

The organization of this paper is as follows. Section 2 explains the PD VQ codebooks in SCHMMs. Section 3 summaries the search algorithm in our SPHINX-II speech recognition system [13]. Section 4 presents the language adaptation algorithm. Section 5 describes our experimental environment and presents our results on the Wall Street Journal task. Finally, future work on PD VQ codebooks is discussed.

2 SCHMMs WITH PD VQ CODEBOOKS

As the density-tying constraint in SCHMMs is relaxed, the degree of freedom in the model is increased and thus more detailed models can be obtained. We relax the tying-constraint such that Markov states from the same phone share the same VQ codebook. Ideally, we would like to relax the constraint further to be senone[14]-dependent VQ codebooks. In the latter case, since there are usually a large number of senones, the VQ size should be reduced to remove redundant fine resolution in the cepstrum space.

However, to make experiments possible in the short term, we clustered the 50 phones for English into 27 classes. A greedy algorithm with cross entropy of the context-independent models as the distortion measure [15] was used for the clustering. Markov states from the same phone class share the same VQ codebook. Sharing densities for different phones is necessary be-

cause we did not reduce the VQ size (256) in our experiments and because phones like BD, GD, and ZH are rare in the training corpus. The 27 phone classes we used for our experiments in Section 5 are:

class	phones	class	phones
0	silence	14	F TH
1	AA AO	15	HH
2	AE EH	16	JH CH
3	AH AW	17	K
4	AX IX IH UH	18	L
5	AXR R	19	M N NG
6	ER	20	OW
7	AY	21	OY
8	SH ZH	22	P
9	B G	23	BD PD TD
10	W		KD DD GD
11	D DX T	24	TS S Z
12	DH	25	UW Y IY
13	EY	26	V

3 SEARCH

The search mechanism in SPHINX-II, described in detail in [12], is a three-pass system. The first two passes generate a word lattice with possible begin and end times, using a bigram language model. The third pass, an A^* search through the word lattice generated by the first two passes, has been extended to flexibly support long distance language models that give values of the form Pr$(w|h)$, where w is a single word extension of the partial solution h. Two such language models are reported here, with results presented in Section 5. The first is a trigram language model and the second is the adaptive scheme described in the following section.

The estimation function [12] required for the A^* search is derived from the results of the second (backward) search pass. During the A^* search, this estimation function is not strictly admissible with respect to the trigram language model that is used since the scores in the word lattice are produced by a bigram. Thus the A^* pass does not produce its results in strict monotonic order with respect to the total path score. It is possible to produce an admissible estimation function based on the trigram language model by rescoring the results of the second pass of the decoder. This, however, increases the necessary decoding computation by at least 50%.

We compared the error rate performance of this more costly estimation function with the performance of selecting the best scored hypothesis among the top 100 solutions generated with the inadmissible estimation function. The two approaches yielded nearly identical error rates. The selection among N-best hypotheses was used in all the experiments reported in Section 5 due to its efficiency.

4 ADAPTIVE LANGUAGE MODELING

The ME principle has recently been demonstrated as a powerful tool for combining statistical estimates from diverse sources [16, 17]. Under the ME scheme, average characteristics of multiple statistical sources constrain the search for a combined language model. An iterative search algorithm, upon convergence, selects the language model with the greatest entropy among all models that satisfy these constraints [18]. We estimate an ME-based language model from conventional bigrams and trigrams, longer distance bigrams and trigrams and word-pair triggers [17]. The resulting language model is then combined with three other sources of information through variable-weight linear interpolation: (1) a static conventional backoff trigram, estimated from 38 million words of WSJ texts, (2) a "rare words only" unigram cache, and (3) a bigram cache. The weights of the four components depend on the length of the part of the document already processed. At the beginning of the document, the static component is dominant. The weight of the adaptive components is then increased gradually as the decoder progresses through the document.

5 PERFORMANCE EVALUATION

The PD VQ codebooks and the A^* search with flexible language models have been incorporated into the SPHINX-II speech recognition system and tested on the 20,000-word open-vocabulary continuous speech speaker-independent WSJ task without punctuations (20o-nvp). The official ARPA trigram language model has a test-set perplexity of 198 [19]. The acoustic training data contain 37,200 utterances from 284 speakers. Two sets of experiments were run separately to explore the improvements of acoustic and language modeling. All the experiments reported in this paper used the decision-tree based senones, in which unseen triphones were also modeled by senones [20].

To explore acoustic improvement, the standard speaker-independent development set (si_dt_20) was tested. It contains 503 contextually independent utterances from five male and five female speakers. To evaluate the advantage of dynamic language adaptation, the November-1993 ARPA evaluation set for language adaptation (si_et_s1) was tested. It has 424 utterances produced in the context of complete long documents by two male and two female speakers. Results for both unsupervised and supervised adaptation are reported in Section 5.2.

5.1 SCHMMs with PD VQ Codebooks

Based on experiences in the November-1992 speaker-independent evaluation set, 10,000 decision-tree based senones were trained using the 37.2K utterances. The gender of each tested speaker was assumed to be known since gender determination is very reliable [21, 22, 23]. We ran the decoder in three-pass mode using the official trigram in the A^* search and the known-sex acoustic models. We generated 100 hypotheses for each utterance and selected the one with the best total score as the recognized output.

Table 1 lists the relative error rate reduction achieved by using the 27 PD VQ codebooks described in Section 2. The acoustic model used in the baseline system is the traditional tied-mixture HMM, in which only one VQ codebook is trained for each feature. The second row shows the word error rate using the 27 PD VQ codebooks. The small improvement (6% in total) is encouraging and leads us to pursue further refinement in the cepstrum space. We believe by increasing the number of codebooks (i.e., relaxing the density-tying constraint further) and decreasing the VQ size (to remove redundant codewords so that the essential ones are well trained), we will be able to observe more improvement. In fact, when the density-tying constraint is completely removed, i.e., when all Markov states have their own set of VQ densities, the HMMs become CHMMs.

5.2 Adaptive Language Modeling

To demonstrate the adaptive language model, we used the Nov93 evaluation set as the testbed. The acoustic model we used for the experiments was the sex-dependent *non*-PD 10K-senone model discussed in the previous subsection. In addition to the 20K words in the lexicon, 178 out-of-vocabulary (OOV) words and their correct phonetic transcriptions were added in order to create closed vocabulary conditions.

acoustic model	male	female
tied-mixture	19.8%	13.9%
27 PD codebooks	18.2%	13.6%

Table 1: Word error rates on the si_dt_20 set of the 20o-nvp WSJ task using 10K decision-tree based senones and the official trigram language model.

To test language model adaptation, we first ran the forward and backward passes of the decoder to create word lattices. Next, three independent A^* passes were performed on the word lattices. The baseline used the official static trigram language model. The second A^* run was for unsupervised word-by-word adaptation in which the decoder output was used to update the language model. The final A^* test used supervised adaptation, in which the decoder output was used for within-sentence adaptation, while the correct sentence transcription was used for across-sentence adaptation. The results are shown in Table 2.

language model	error rate	reduction
static trigram	19.9%	—
unsupervised	17.9%	10%
supervised	17.1%	14%

Table 2: Word error rates on the si_et_s1 set of the 20o-nvp WSJ task using different language models.

6 CONCLUSION AND FUTURE WORK

We have shown that using phone-dependent VQ codebooks and adaptive language models can improve the recognition accuracy significantly.

Our next short-term goal is to relax the density-tying constraint to be senone-dependent VQ codebooks, with each VQ size around 10 – 30. This way, not only is computation reduced compared with CHMMs, but similar states also share parameters so that the parameters are well trained. Because we advocate the use of decision trees to generate the senone

sharing structure, we will continue to use the same technique for the density sharing. In addition, we are encouraged about this line of work because similar efforts in senone-dependent VQ codebooks have been shown to provide improved performance [24, 25].

Acknowledgements

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The authors would like to express their gratitude to Professor Raj Reddy for his encouragement and support, and other members of CMU speech group for their help.

References

[1] Rabiner, L. R., Juang, B. H., Levinson, S. E., and Sondhi, M. M. *Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities*. **AT&T Technical Journal**, vol. 64 (1985), pp. 1211–33.

[2] Richter, A. *Modeling of Continuous Speech Observations*. in: **Advances in Speech Processing Conference, IBM Europe Institute**. 1986.

[3] Ney, H. and Noll, A. *Phoneme Modelling Using Continuous Mixture Densities*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1988, pp. 437–440.

[4] Gray, R. *Vector Quantization*. **IEEE ASSP Magazine**, vol. 1 (1984), pp. 4–29.

[5] Makhoul, J., Roucos, S., and Gish, H. *Vector Quantization in Speech Coding*. **Proceedings of the IEEE**, vol. 73 (1985), pp. 1551–1588.

[6] Huang, X., Ariki, Y., and Jack, M. **Hidden Markov Models for Speech Recognition**. Edinburgh University Press, Edinburgh, U.K., 1990.

[7] Bellegarda, J. and Nahamoo, D. *Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1989, pp. 13–16.

[8] Lee, C., Rabiner, L., Pieraccini, R., and Wilpon, J. *Acoustic Modeling for Large Vocabulary Speech Recognition*. **Computer Speech and Language**, vol. 4 (1990), pp. 127–165.

[9] Aubert, X., Ney, H., and Haeb-Umbach, R. *Philips Research System for Continuous-Speech Recognition Overview and Evaluation on the DARPA RM Task*. in: **DARPA Continuous Speech Recognition Workshop**. DARPA Microelectronics Technology Office, Stanford, CA, 1992.

[10] Jaines, E. *Information Theory and Satistical Mechanics*. **Phys. Rev.**, vol. 106 (1957), pp. 620–630.

[11] Kullback, S. **Information Theory and Statistics**. Dover, New York, 1959.

[12] Alleva, F., Huang, X., and Hwang, M. *An Improved Search Algorithm for Continuous Speech Recognition*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1993.

[13] Hwang, M., Huang, X., and F., A. *Senones, Multi-Pass Search, and Unified Stochastic Modeling in SPHINX-II*. in: **Proceedings of Eurospeech**. 1993.

[14] Hwang, M. and Huang, X. *Subphonetic Modeling with Markov States — Senone*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1992.

[15] Lee, K. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. Computer Science Department, Carnegie Mellon University, April 1988.

[16] Rosenfeld, R. *Adaptive Statistical Language Modeling: a Maximum Entropy Approach*. Ph.D. Thesis Proposal, 1992.

[17] Lau, R., Rosenfeld, R., and Roukos, S. *Trigger-Based Language Models: a Maximum Entropy Approach*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1993.

[18] Darroch, J. and Ratcliff, D. *Generalized Iterative Scaling for Log-Linear Models*. **The Annals of Mathematical Statistics**, vol. 43 (1972), pp. 1470–1480.

[19] 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.

[20] Hwang, M., Huang, X., and Alleva, F. *Predicting Unseen Triphones with Senones*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1993.

[21] Soong, F., Rosenberg, A., Rabiner, L., and Juang, B. *A Vector Quantization Approach to Speaker Recognition*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1985, pp. 387–390.

[22] Huang, X., Lee, K., Hon, H., and Hwang, M. *Improved Acoustic Modeling for the SPHINX Speech Recognition System*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. Toronto, Ontario, CANADA, 1991, pp. 345–348.

[23] Lamel, L. and Gauvain, J. *Cross-Lingual Experiments with Phone Recognition*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1993.

[24] Digalakis, V. and Murveit, H. *An Algorithm for Optimizing the Degree of Mixture-Tying in a Large-Vocabulary HMM-Based Speech Recognizer*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1994.

[25] Young, S. and Woodland, P. *The Use of State Tying in Continuous Speech Recognition*. in: **Proceedings of Eurospeech**. 1993.