PENNSTATE

TransactionNumber: 1649297

ILL Number: 107099899	**Call #: TK7895.S65D37 1995**
MaxCost: 35.00IFM	**Branch Location: k2**
	9000000061523
Billing Category: Exempt	

Penn State Interlibrary Loan – Article Request

Article Information

Journal Title: Proceedings of the ARPA Spoken Language Systems Technology Workshop.

Volume: **Issue:**
Month/Year: 1994 **Pages:** ?

Article Author: R. Rosenfeld, E. Thayer, R. Mosur, L. Chase, R. Weide, M. Hwang,X. Huang and F. Alleva

Article Title: Improved Acoustic and Adaptive Language Models for Continuous Speech Recognition

Loan Information

Loan Title:

Loan Author:

Publisher:
Place:
Date:
Imprint:

Borrower Information

E & S Library
Carnegie Mellon University / IDS #187A01
Wean Hall
5000 Forbes Ave
Pittsburgh, PA 15213-3890

Odyssey: 128.2.20.146
Ariel: 128.2.21.4/128.2.114.135
Email: es82@andrew.cmu.edu

NOTES

/ 7/17/2013 4:28:27 PM (System) Borrowing Notes; ARIEL 128.2.21.4/128.2.114.135 (maxCost; $35.00) / PMC is a SHARES member.
7/17/2013 4:28:28 PM (System) Billing Notes; Please put IL# on invoice.

RETURN LOANS TO: Penn State University Libraries / Interlibrary Loan / 127 Paterno Library, Curtin Rd. / University Park, PA 16802

IMPROVED ACOUSTIC AND ADAPTIVE LANGUAGE MODELS FOR CONTINUOUS SPEECH RECOGNITION

R. Rosenfeld, E. Thayer, R. Mosur, L. Chase, R. Weide
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

M.Hwang, X. Huang, F. Alleva
Microsoft Research
Redmond, Washington

ABSTRACT

We present improvements in acoustic and language modeling for large vocabulary continuous speech recognition. Two refinements to our acoustic models and a new approach to adaptive language modeling are discussed.

Sphinx-II's semi-continuous HMMs (SCHMMs) have been extended to incorporate phone-dependent VQ codebooks (PDVQ). Phone-dependent VQ codebooks relax the density-tying constraint in SCHMMs, enabling the construction of more detailed models. Using this method a 6% error rate reduction was achieved on the speaker-independent 20,000-word Wall Street Journal (WSJ) task.

More detailed speaker clustering has also been added to the system. Based on a linear warping of mel-scale frequency coefficients at the triphone level we cluster our speakers into four groups (two male and two female). This extension yields a small additional improvement of 3% beyond the PDVQ result, yielding a total of 9% improvement from improvements in acoustic modeling.

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. At the other extreme, continuous HMMs (CHMMs) require expensive computation, precluding real-time implementations for large vocabulary dictation systems on currently available hardware. 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. In order to push the structure of our SCHMMs closer to a truly continuous representation, we relax the density-tying constraint. This is accomplished by incorporating SCHMMs with phone-dependent (PD) VQ codebooks, in which Markov states of all triphones that represent the same phone share the same set of VQ codebooks or densities [8, 9]. Thus a modification of our SCHMM system provides performance closer to those of CHMMs without incurring the speed costs of CHMMs.

A second modification to our acoustic models is also addressed here. Because clustering speakers by gender provided significant performance improvement in the past, we pushed this direction a bit farther and developed a more detailed clustering technique. The end result was an automatic clustering of speakers into four classes, two male and two female. We report the clustering technique and commensurate performance improvements.

The third 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 supports inclusion of multiple sources of statistical language information, including conventional bigrams and trigrams, longer distance bigrams and trigrams, and word-pair triggers. Because these sources of information will typically be overlapping in the information they provide, simple interpolation is inappropriate for their combination. The ME framework provides a theoretically sound approach to correct combination. 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 summarizes the search algorithm in our SPHINX-II speech recognition system [12, 13, 14]. Section 3 explains the PD VQ codebooks in SCHMMs. Section 4 describes the added detail in our speaker clustering work. Section 5 presents the language model adaptation algorithm and some related discussion. Experimental results can be found throughout the paper in the appropriate sections.

2. 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 Section 5.

The estimation function 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 a 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. Therefore, the selection among N-best hypotheses was used in all the experiments reported in this paper due to its efficiency.

3. 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[15]-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 [16] was used for the clustering. Markov states from the same phone class share the same VQ codebook. Sharing densities for different phones is necessary because 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 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

All of the techniques reported in this paper 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). We used the official ARPA trigram language model, which has a test-set perplexity of 198 [17]. The acoustic training data includes 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 use decision-tree based senones, in which unseen triphones are also modeled by senones [18].

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. 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 [19, 20, 21]. We ran the decoder in three-pass mode using bigram and the known-sex acoustic models in the first two passes of search, followed by an A^* search with the official trigram. We generated 100 hypotheses for each utterance and selected the one with the best path score as the recognized output.

Table 1 lists word error rates on si_dt_20. The acoustic model used in the baseline system (1st row) is the traditional tied-mixture SCHMM, 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. This refinement will be pursued 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). In the extreme, when the density-tying constraint is completely removed, i.e., when all Markov states have their own set of VQ densities, SCHMMs become CHMMs.

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.

4. MULTIPLE SPEAKER CLUSTERS

The 37,000 plus training utterances available to us are ample for training multiple sets of models. We explored several approaches to capturing speaker variations with the aim of settling on a good set of speaker clusters. The greatest success we found was in looking for additional clusters within gender, using a linear warping of mel-scale frequency coefficients at the triphone level to further partition the speakers.

The two most diverse speakers in each gender group were first identified. For the remaining speakers of the same gender, the closer of the two initially identified speakers is used to decide the cluster membership of the new speaker in question. This clustering resulted in two male and two female speaker clusters.

Using this technique, we found that the two most distinct male speakers were speaker 4b0 and 4az; the two females were 47s and 47n. It is interesting to note that the two most distinct speakers in each gender group under our measure are also very distinctive speakers upon listening: male speaker 4az has a speech impediment, while 4b0 has a strident and very nasal voice. The two female extremes were also interesting: 47n is an Indian with a very strong Indian English accent, while 47s is a very low-pitched female. Due to their unusual acoustics, the clusters formed around speakers 4az and 47n are much smaller than the other two clusters. The acoustic model for each speaker cluster was obtained by interpolating the Baum-Welch counts of one cluster with the other, with a higher weight on the target cluster. The results shown in the last row of Table 2 seem to indicate that additional speaker clusters are always helpful when

there are enough training data to generate the models. Even though the improvement is small, it makes the system more robust against outlier speakers. For example, cluster 47n may prove to be very helpful for a testing speaker who has a strong Indian accent.

	si_dt_20-M	si_dt_20-F	% change
10k senones	19.8%	13.9%	
+ 27 PD codebooks	18.2%	13.6%	-6%
+2F,2M	17.7%	13.1%	-3%
TOTAL			-9%

Table 2: Word error rates on si_dt_20 with the official trigram.

5. ADAPTIVE LANGUAGE MODELING

The ME principle has recently been demonstrated as a powerful tool for combining statistical estimates from diverse sources [22, 23]. 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 [24]. We estimate an ME-based language model from conventional bigrams and trigrams, longer distance bigrams and trigrams and word-pair triggers [23]. As discussed in [22], 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 conditional 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.

As described above in the section on search, SPHINX-II uses a multi-pass decoder. The first two passes ("forward" and "backward") use a bigram language model and generate a word lattice. The third pass is an A^* search of that lattice. It is during this pass that the adaptive language model calculations were introduced. (During the contrastive, or baseline run, a conventional trigram was used instead of the adaptive model.) Every time the search process needed to expand a node, it called on the adaptive language model, providing it with the path to that node. The language model also had access to the previous sentences in the same document. Upon return, it produced an array of probabilities, one for each word in the vocabulary.

The Spoke 1 Adaptive Language modeling test consisted of 424 utterances produced in the context of complete long documents by two male and two female speakers. The version of SPHINX-II used for the adaptive language modeling experiments did not include all of the acoustic modeling techniques described above. Instead the system was simply configured with sex-dependent (1F,1M cluster) non-PD 10K senone acoustic models for efficiency.

In addition to the ~20,000 words in the standard WSJ lexicon, 178 out-of-vocabulary words and their correct phonetic transcriptions were added manually before search started in order to create closed vocabulary conditions. As described above, the forward and backward passes of SPHINX-II search were first run to create word lattices, which were then used by three independent best-first runs. The first such run used the 38M word static trigram language model, and served as the baseline. The other two runs used the interpolated adaptive language model, which was based on the same 38 million words of training data. The first of these two adaptive runs was for unsupervised word-by-word adaptation, in which the recognizer's output was used to update the language model. The other run used supervised adaptation, in which the recognizer's output was used for within-sentence adaptation, while the correct sentence transcriptions of previous utterances were used for cross-sentence adaptation. Results are summarized in Table 3[1].

language model	word error rate	% change
trigram (baseline)	19.9%	—
unsupervised adaptation	17.9%	-10%
supervised adaptation	17.1%	-14%

Table 3: Word error rate reduction of the adaptive language model over a conventional trigram model.

The OOV words in the test were given special treatment: during the forward-backward passes, they were assigned uniform probabilities. During the third pass, they were assigned their average unigram probabilities. This was done for practical purposes, to avoid the need to train a special ME model for the new, extended vocabulary, while maintaining equitable comparison between the baseline and adaptive models.

In order to better calibrate this approach, a similar experiment was run in which the baseline trigram was retrained to include the OOV's, such that the latter were treated as any other vocabulary word. This reduced the baseline word error rate. The ME model was *not* retrained, which penalized it somewhat. And yet, after interpolation, the relative improvement of the adaptive model was not significantly affected, as is shown in Table 4.

language model	word error rate	% change
trigram (baseline)	18.4%	—
unsupervised adaptation	16.7%	-9%
supervised adaptation	16.1%	-13%

Table 4: Word error rate reduction of the adaptive language model over a conventional trigram model, where the latter was retrained to include the OOVs.

The adaptive language model developed in this work uses various sources of information:

1. Information from the current sentence (conventional trigram, distance-2 trigram, triggers).
2. Information from previous hypothesized sentences (triggers and caches, unsupervised adaptation).
3. Information from previous correct sentences (triggers and caches, supervised adaptation).

[1] The error rates in this experiment were higher than those achieved under comparable conditions in other evaluation runs. Upon analysis, this turned out to be due to two of the four speakers being "goats" (speakers which the recognizer performs poorly on).

Another experiment was run to determine how the different sources contributed to the error rate reduction. An unsupervised adaptation system, identical to that reported in Table 4 was run on the same data, but the context (history) was flushed after every sentence. This created the condition in which every sentence was processed as if it were the first one in a document, namely without the benefit of adaptation from previous sentences. Results are presented in Table 5 (the other results are reproduced from Table 4). It seems that all three sources contributed significantly to error rate reduction.

	word error rate	% change
baseline	18.4	—
isolated sents. adapt.	17.3	-6%
continuous unsupervised adapt.	16.7	-9%
continuous supervised adapt.	16.1	-13%

Table 5: Word error rate reduction broken down by source of information.

The ME-based adaptive language model that was trained on the full WSJ corpus (38 million words) reduced perplexity by 32% over the baseline trigram. The associated reduction in recognition word error rate was 14% under the most favorable circumstances. The word error rate experienced under the adaptative model is not as large as the commensurate perplexity reduction. For a discussion of the sometimes surprising relationship between perplexity and recognition error rate, see [22].

6. CONCLUSIONS

Improvements due to better acoustic modeling yielded a total incremental reduction of 9% in the recognition error rate. This was achieved via a combination of the use of phone-dependent VQ codebooks and more detailed speaker clustering. Independent development of adaptive language models under the maximum entropy framework achieved a 10% to 15% reduction in error rate for utterances in the context of short documents.

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. Huang, X., Alleva, F., Hon, H., Hwang, M., Lee, K., and Rosenfeld, R. *The SPHINX-II Speech Recognition System: An Overview*. **Computer Speech and Language**, vol. 2 (1993), pp. 137–148.
14. Hwang, M., Huang, X., and F., A. *Senones, Multi-Pass Search, and Unified Stochastic Modeling in SPHINX-II*. in: **Proceedings of Eurospeech**. 1993.
15. Hwang, M. and Huang, X. *Subphonetic Modeling with Markov States — Senone*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1992.
16. Lee, K. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. Computer Science Department, Carnegie Mellon University, April 1988.
17. 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.
18. Hwang, M., Huang, X., and Alleva, F. *Predicting Unseen Triphones with Senones*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1993.
19. 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.
20. 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.
21. Lamel, L. and Gauvain, J. *Cross-Lingual Experiments with Phone Recognition*. in: **IEEE International Conference on Acoustics, Speech, and Signal Processing**. 1993.
22. Rosenfeld, R. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. School of Computer Science, Carnegie Mellon University, 1994.
23. 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.
24. Darroch, J. and Ratcliff, D. *Generalized Iterative Scaling for Log-Linear Models*. **The Annals of Mathematical Statistics**, vol. 43 (1972), pp. 1470–1480.