Improvements in Language, Lexical, and Phonetic Modeling in Sphinx-II

L. Chase, R. Rosenfeld, A. Hauptmann, M. Ravishankar,
E. Thayer, P. Placeway, R. Weide, C. Lu

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

We studied the effect of the various types and amounts of training data on the quality of the derived vocabulary, and used our findings to derive an improved ranking of the words, using only 19% of the LM training data. We then studied the conflicting effects of increased vocabulary size on the system's accuracy, and used the result to pick an optimal vocabulary size. A similar analysis of ngram coverage yielded a very different outcome, with the best system being the one based on the most data. A new implementation of the cache language model was tested which yielded approximately 4% improvement on a development test. We also studied a phrase grammar for common acronyms, which had a small but consistently positive effect, yielding an approximate gain of 0.2% (absolute) on the evaluation test set. A change was made in the evaluation of right acoustic contexts for single phone words. This yielded a consistent 3% relative improvement across multiple development tests. A very simple class grammar was implemented to capture variations in verbalized pronunciation. It, too, had a small but consistently positive effect, delivering an improvement of 0.1% (absolute) on the final evaluation test.

1. Vocabulary Optimization

1.1. OOV curve minimization

Since Out-Of-Vocabulary (OOV) rate directly affects Word Error Rate, with every OOV word in the test data resulting in at least one (and often more) recognition errors, we set out to minimize the expected OOV rate of the test data. More generally, our goal was to understand how availability of various types and amounts of training data, from various time periods, affects the quality of the derived vocabulary[1]. Given a collection of training data, we sought to create an ordered word list with the lowest possible OOV curve, such that, for any desired vocabulary size V, a minimum-OOV-rate vocabulary could be derived by taking the first V words in that list. Viewed this way, the problem becomes one of estimating unigram probabilities of the test distribution, and then ordering the words by these estimates.

In the 1994 ARPA CSR task, The test-set was sampled from 5 different North American Business news sources (DJIS, RNAB, NYT, WP, LAT), in equal parts, all from the period 6/16/94–7/15/94. Development data was similarly drawn from 4/1/94–6/15/94. The training data consisted of WSJ(87-92, 69M words), DJIS(92-94, 42MW), AP(88-90, 106MW) and SJM(91, 11MW). See [1] for details.

We started by trying to minimize OOV-rate for the DJIS source, using the DJIS portion of the LM development set (680K words). We split the latter in two, using one half in the controlled studies reported below and the other half for validation. In all such studies, except where otherwise noted, the word list was ordered by decreasing frequency in the appropriate subset of the training data.

We first set out to measure the effect on OOV rate of the *seasonality* of the training data, namely the time of year from which it is drawn. For each month of the year, we created a word list based on some 9MW of training data from that month, using the AP(1988-1990) data. The DJIS development data was drawn from 4/94, so a seasonal effect might reduce the OOV rate of training data from this or adjacent months. As Figure 1 shows, no such effect was found.

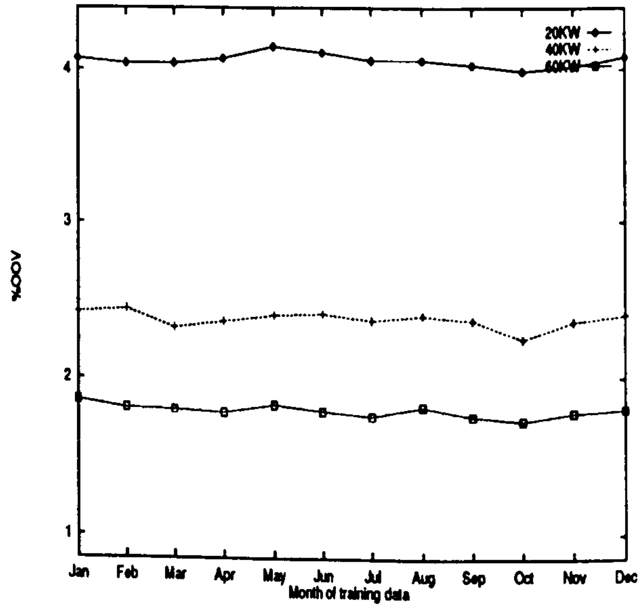

Figure 1: Month from which training data is drawn has no effect on OOV rates (test data is from April).

Next we measured the correlation of OOV rate with the *amount* of training data. Using AP88-90 data, we added training data in increments of 5MW, and measured the impact on OOV rate. We added data in decreased order of recency, so as not to confound the effect of the amount of data with that of its recency. Figure 2 shows our findings. As expected, more training data results in lower OOV rates. But improvement slows down considerably after 30MW–50MW.

Next, we studied the effect of *recency* of the training data. Figure 3 shows OOV rates based on similar amounts of DJIS training data (about 5MW), but from different time periods. Time indeed makes a difference, albeit slowly. Over a period of 2 years, the 20KW (60KW)

[1] The vocabulary thus derived is *static*. It can serve as the initial vocabulary, to be optionally extended at runtime based on the words encountered in the test data.

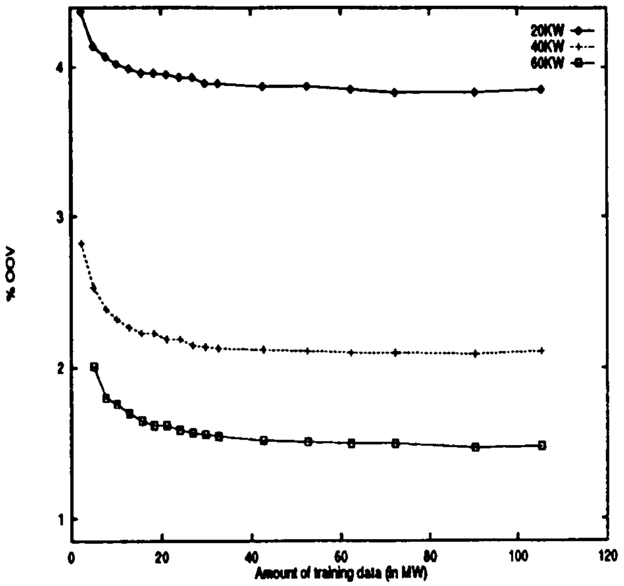

Figure 2: More training data results in lower OOV rates, but mostly up to 30MW–50MW

OOV rate degraded by 5% (15%). Over 4.5 years, it degraded by 11% (24%).

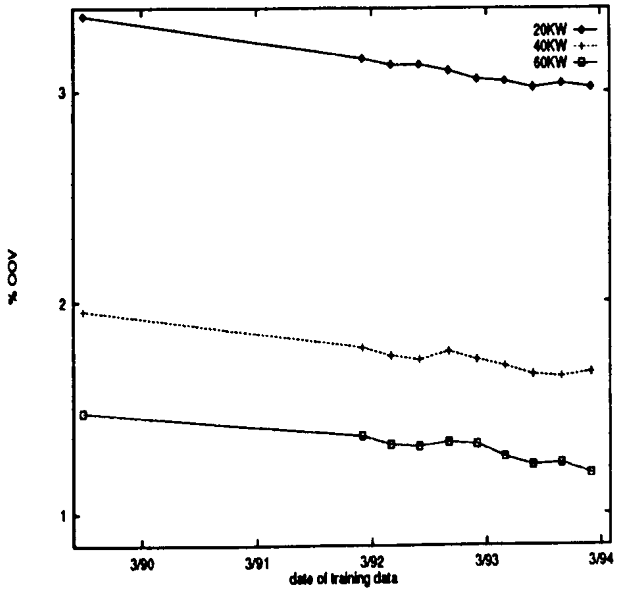

Figure 3: More recent training data results in lower OOV rates.

The difference that the *source* of the training data can make is evident in Figure 4. An OOV curve based on the WSJ90 part of the data (10MW), is lower than that based on the SJM91 part (11MW), even though the latter is larger and more recent.

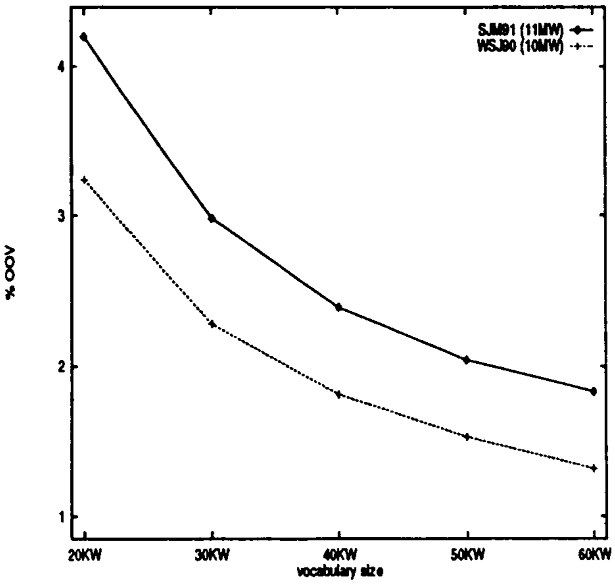

Figure 4: The source of the training data makes a big difference in OOV rates.

Next, we accumulated DJIS data starting from the most recent 'chunk' (DJIS94) and going backwards in time. Given the inherent tradeoff between the amount of data and its recency and source, we hypothesized a U-shape OOV curve, which was indeed achieved as can be seen in Figure 5 (the last datapoint is based on the entire 227MW NAB training corpus). The peak was achieved at about 40MW. It is interesting that the best overall coverage was obtained using only 19% of the available training data!

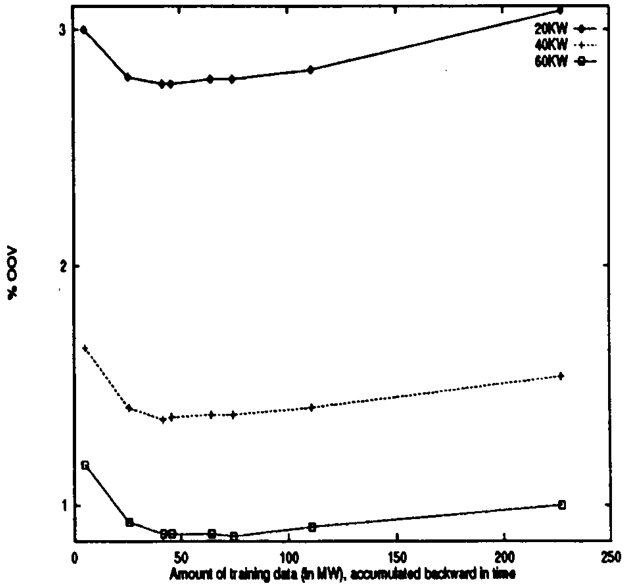

Figure 5: Best OOV rates are achieved with only 19% of the training data!

If recent data is more useful, can we benefit from emphasizing it? Several such attempts failed. The only one that was mildly successful was based on a "leaky capacitor" model of word probabilities. Discounting the word counts by 1% every week reduced the OOV rate very slightly for vocabulary sizes in the range 20KW–50KW, but not at the 60KW level.

In an attempt to further lower the OOV curve, we constructed a 'profile' for each word, consisting of its token counts broken down by week of occurrence. We then browsed the profiles of words that were OOVs with regards to our best 60KW vocabulary. We were hoping to find some patterns that will allow us to predict such words in advance and assign them a better position in the word list. Unfortunately, no such patterns emerged. The vast majority of OOV words had a very mundane profile: 1-2 token occurrences in each of several sporadic weeks — nothing that can predict their upcoming occurrence. This finding, combined with the fact that 90%–95% of the OOV words are proper nouns or their possessives, leads us to conclude that at the current rate of training data accumulation further reduction of the OOV curve is unlikely.

In running similar tests on the other four test-set sources (RNAB, NYT, LAT, WP), we found their qualitative behavior identical to that of the DJIS data, allowing us to use one optimized word list for all sources.

1.2. Vocabulary size optimization

Increasing the vocabulary of a speech recognition system has two conflicting effects. On one hand, it reduces the OOV rate, thereby helping to recover OOV related recognition errors. On the other hand, the added lexical entries increase the average acoustic confusability of words, resulting in new recognition errors.

To quantify these two effects, we ran two controlled experiments on the CSR 1994 acoustic development test set. In the first, we compared two systems that differed only in their lexicon. The first system had a lexicon of 58K words. The second was based on the top 20K words of the 58K lexicon, supplemented with all the test set words that were in the 58K lexicon. Thus the two lexicons had identical coverage of the test set, but very different overall sizes. The 58KW system resulted in 0.6 points higher WER. We interpreted the difference as resulting solely from the increased acoustic confusability. Assuming that acoustic confusability grows roughly linearly with vocabulary size, we arrived at a slope of +0.16 WER points per 10KW increase in the vocabulary. Alternatively, assuming that acoustic confusability grows logarithmically with vocabulary size, we arrived at a slope of +0.39 WER points per doubling of the vocabulary size.

In the second experiment, we again compared two systems differing only in their lexicon. One system used the 58K lexicon; the other used the same 20K lexicon as above (unsupplemented), and had a 1.79% higher OOV rate. Thus, this time the lexicons differed in both size and test-set coverage. The 20KW system had a 1.55 points higher WER. Factoring in the 0.6 points WER reduction due to the reduced confusability, we corrected the effective difference to 2.15 WER points. Assuming that OOV-related errors are linear with the OOV rate, we arrived at a slope of –1.2 WER points per OOV-point, or an average of 1.2 word recognition errors per OOV word[2].

As we increase vocabulary size, OOV rate decreases at an ever slower rate. For any OOV curve, there is a point at which the savings due to reduced OOV rate are exactly offset by the additional errors due to acoustic confusability. That point is the optimal vocabulary size. Figure 6 combines the slopes estimated above to arrive at a projected WER as a function of vocabulary size, for this particular task. Assuming acoustic confusability grows linearly, optimal vocabulary size is about 66K words. But the slope is very mild in the range 55KW–80KW. Assuming acoustic confusability grows logarithmically, optimal vocabulary size is in the range 80KW–110KW, but the slope is very mild starting at 70KW.

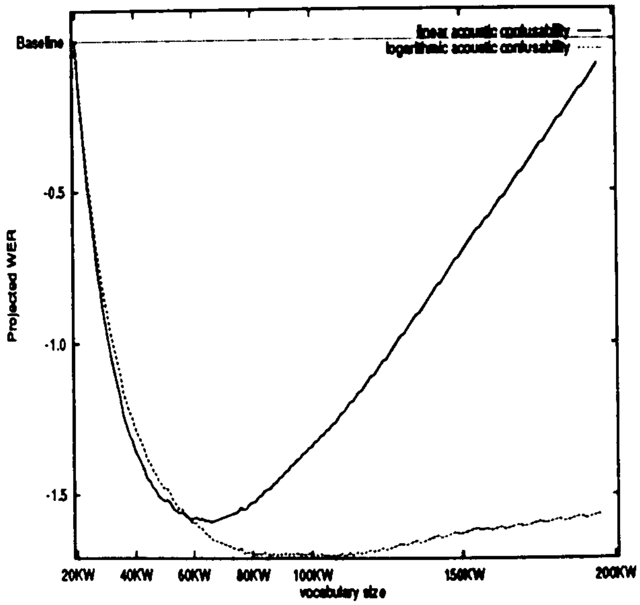

Figure 6: Projected WER based on estimated slopes for OOV errors and acoustic confusabilty. Increasing the vocabulary beyond 64KW is likely to yield negligible improvement at best.

Note though that our estimates are not very accurate. Furthermore, we do not know which of the two assumption is more correct, although it is reasonable to assume that the true answer lies somewhere in between them. We can only conclude that, for this task and with our current recognition system, increasing the vocabulary beyond the 64KW point is likely to yield negligible improvement at best.

1.3. Lexical coverage: summary

From the studies reported above we conclude that, at least in this domain:

- Lexical coverage is strongly affected by the amount of training data used to construct the lexicon, but the effect attenuates around 30MW–50MW.
- Month from which the data is drawn is insignificant.
- Source of the data (SJM vs. AP vs. WSJ vs. DJIS) is very important.
- Recency is also important: over 2 years there is a 5%–15% degradation in OOV rate. Over 4.5 years, a 11%–24% degradation..
- Best lexical coverage of the CSR development test distribution is achieved with only the DJIS data (4/92-3/94, 42MW), with mild emphasis of recency within that period. Note that only 19% of the available training data is being used.
- With an optimized 60KW lexicon, the occurrence profile of the remaining OOV words is very unremarkable, leaving little hope for further improvement.
- Every OOV word results on average in some 1.2 word recognition errors.

[2] This result agrees with similar numbers reported by [2] and [3].

- As the vocabulary grows, increased acoustic confusability is a non-negligible source of recognition errors. Since OOV rate declines at a slowing rate, there is a point of optimal vocabulary size. For this task and our current system, that point is in the range 55KW–80KW (assuming acoustic confusability grows linearly) or 80KW–110KW (assuming it grows logarithmically).

Given the last conclusion, and the limit of 64K pronunciations in our current implementation of the decoder, we settled on a vocabulary of some 59,000 words. These resulted in 64,500 pronunciations. The OOV rate of the 1994 eval test set with regard to this vocabulary was 0.5% (42/8186), compared to 2.4% (194/8186) relative to the official 20KW vocabulary used in the C1 run. Using the slope of 1.2 WER points for every point in OOV rate reduction, we arrive at an estimated WER reduction of 2.2% on the eval set due to the expanded and optimized vocabulary.

The extent of reduction in OOV rate due to word-order optimization depends on the vocabulary size. The larger the vocabulary, the smaller the difference, since OOV rates themselves decline rapidly. With our vocabulary of 59KW, the reduction in OOV rate over the baseline (a simple top-frequency list based on the entire training corpus) was moderate (12%). But more importantly, the OOV studies revealed the dependenceof lexical coverage on various aspects of the training data. This will help us determine how much (and what kind of) data we need in order to get sufficient coverage in other tasks. Moreover, the same technique can be used to study (and subsequently optimize) coverage of bigrams and trigrams. See Section 2 for the beginning of such an investigation.

1.4. Lexical coverage: analysis

In North American Business English (as defined by the 1994 NAB corpus), the least frequent among the most frequent 60K words have a frequency of about 1:7M. In optimizing a 60KW vocabulary we are thus trying to distinguish words with frequency of 1:7M from those that are slightly less frequent. To differentiate somewhat reliably between a 1:7MW word and, say, a 1:8MW word, we need to observe them enough times for the difference in their counts to be statistically reliable, i.e. we must have at least 100MW–200MW of training data. Fortunately, for constructing a decent vocabulary, it is enough that *most* such words are ranked correctly. For this, 50MW–100MW might be sufficient (since the expected difference between the counts will be 1–2). This agrees with the empirical results reported above, according to which the OOV curve improves rapidly as more training data is used up to 50MW, and then continues to improve more slowly beyond that point.

To optimize the vocabulary for coverage of a specific time period, we must use training data from that period, or as close to it as possible. But for, say, 70MW of training data, at the DJIS wire feed data rate, we need 4 years, during which the language shifts considerably, and 60KW OOV-rate degrades by some 22% (see study of recency above). Thus we are inherently unable to fully optimize the vocabulary.

We can further generalize the last observation. Viewing language as a non-stationary stochastic source, and generalizing the word probabilities to any time-dependent linguistic phenomenon (e.g., a rise in the probability of an ngram above its static level), we arrive at the following principle:

> One can never determine accurately both the extent and the time frame of a linguistic phenomenon.

There is an inherent tradeoff between the accuracy of an estimate and the time period it is based on. More precisely, it is not the time period but the amount of training data that is the limiting factor. But since there is only a limited amount of data from each time period, the two are related by a constant.

Thus if a phenomenon is both transient and rare, we are inherently incapable of detecting it. Note that rare phenomena are not necessarily unimportant, since there may be many of them. Estimating an event as having Probability 10^{-7} rather than 10^{-6} can have a devastating effect on the log-probability, and hence recognition, of a sentence. Yet such events are commonly modelled in most existing language models and commonly encountered in test data.

The amount of LM training data available until recently was small enough that the benefit from acquiring more data dominated over the disadvantage due to language shift. But with the larger amounts of data made available recently, this is changing. With the 1994 NAB corpus of 227M words, we have already found that better vocabularies are constructed by using only a fifth of the available data. As will be seen in the next section, similar results do not yet apply to ngram lists. But with several billion words of training data, we believe they will. Language modeling is close to the point where the time-honored maxim "there's no data like more data" no longer holds.

2. Ngram Coverage and Language Model Size

In a recent work ([4]) we found that recognition errors are much more likely to occur within trigrams and (especially) bigrams which have not been observed in the training data. In these cases, the language model typically relies on lower order statistics. If the bigram is missing, predictions are made based on unigram statistics, which are notoriously unreliable. Thus increased ngram coverage may translate directly into improved recognition accuracy.

But increased ngram coverage usually comes at the cost of increased memory requirements. To study the tradeoffs involved, we compared several systems on the 1994 development test. All systems used a 58KW vocabulary (different than the optimized vocabulary reported in Section 1) and conventional trigram backoff language models. The models differed in the amount of data they were trained on, and in their bigram and trigram cutoffs. Table 1 summarizes our results, in decreasing order of Word Error Rate[3]. 't94' refers to the entire official 1994 NAB training corpus (227MW). '-m-n' means that bigrams occurring m or fewer times and trigrams occurring n or fewer time were excluded. The 'coverage' columns reports the rate at which the backoff language model relied on its trigram, bigram, and unigram components to produce scores for the transcripts (1.4% of the words were OOVs).

A few observations:

- Given the same training data, adding bigrams or trigrams (by lowering their respective cutoffs) improves both perplexity and

[3] The last WER result is approximate, since it involves corrections to account for other system components.

system	# of (M)		coverage(%)			PP	WER
	2g	3g	3g	2g	1g		
wsj93-0-0	3	7.5	57	31	11	197	
t94-1-2	6	10	63	29	6.8	156	14.7
wsj91-94-0-1	6	5	59	32	7.3	163	14.55
wsj87-94-0-1	9	8.5	63	29	6.4	153	14.35
t94-0-2	14	10	63	30	5.2	153	14.3
t94-1-1	6	18	67	25	6.8	152	14.25
t94-0-1	14	18	67	27	5.2	150	14.1*

Table 1: Ngram coverage, perplexity and Word Error Rate for LMs based on various amounts of data and different ngram cutoffs. "There's no data like more data" still holds.

recognition. Interestingly, 't94-0-2' and 't94-1-1' performed similarly, even though one had 8M more bigrams while the other had 8M more trigrams.

- In the case of lexical coverage, older and less relevant training data actually hurt performance. But with ngram coverage, this does not seem to be the case. Our hypothesis is that this difference is due to the much lower frequency of the ngrams (as compared to the least frequent words in the vocabulary). See discussion in Section 1.4. The largest system ('t94-0-1') performed best on the dev data, and was consequently used in our evaluaton system.

It is hard to draw further conclusions from comparing models based on different training sets. For example, 'wsj91-94-0-1' has fewer trigrams, worse test-set ngram coverage and worse test-set perplexity than 't94-1-2', and yet it performed better. Perhaps the differences in WER are not large enough to be significant. Clearly, more carefully controlled studies are called for.

3. Pronunciations for Expanded Vocabulary

To generate pronunciations for the new larger lexicon of our system, we used multiple dictionary sources. The pronunciations for the vocabulary words were derived from these multiple sources in order of the reliability of the dictionaries in question:

1. The initial (and most reliable) pronunciations came from the subset of words found in the 20k dictionary we'd been building by hand for use in the 1994 hub test – approximately 75% of these were words we had used in previous years.
2. Words not in this dictionary were then sought in the CMU 100k word dictionary, which is our publicly distributed dictionary [4]. Since this dictionary uses a different phoneset, the lookup hits were mapped by rule into the SPHINX phoneset.
3. Words still not found were then looked up in the UCLA/SHOUP dictionary and the pronunciations from that phoneset converted by rule into the SPHINX phoneset.
4. Remaining words without pronunciations were approached using Mitalk, ORATOR and DECTALK software [5] [6]. Whenever two of the three synthesis programs agreed on a pronunciation, it was included. Appropriate phone set conversions were done by rule.
5. Using a set of suffixes ("S", "'S", "'", "S'S") an attempt was made to find the morphological roots of the remaining words in one of the first three dictionaries. The words were looked up after the suffixes were added/removed and if the modified word was found, the appropriate pronunciation was used. (Only a few hundred words were derived in this manner.)
6. All remaining pronunciations were taken from the ORATOR pronunciation results. (Most of these remaining words were proper names.)

The thousand (1000) most frequent words in the vocabulary were verified by hand, as were the ORATOR-only words. Spot checks were made on the other pronunciations.

4. Cache Language Model

The static language model was linearly interpolated with a selective unigram cache and a conditional bigram cache, identical to those described in Rosenfeld's Ph.D. thesis[7]. Only words with unigram probabilities below 0.001 were included in the unigram cache, whose weight was proportional to its size but saturated at 0.04. The bigram cache had a weight of 0.09 when active, 0 otherwise. Unlike last year's S1 system, the caches were used in the forward (first) pass of the search algorithm in addition to the (third) A* pass [8].

language model	male	female
dev baseline	16.6%	11.9%
with cache LM	16.0%	11.9%

Table 2: The effect of unigram and bigram cache language model. Word error rates are on the Nov 1994 CSRNAB development test set using 10K decision-tree based senones with known gender. The baseline language model is on a non-optimized test vocabulary of 58K words. The language model training data for the baseline language model is all of the standardly available data.

The development baseline for testing the contribution of this version of the cache language model was a language model trained on the 't94' data (see Section 2) with trigram and bigram cutoffs both set at one. The language model was built using the 58K word test vocabulary (also discussed in Section 2). Our standard semi-continuous 10K decision-tree bases senones with known gender were used for the acoustic models. We tested on the 1994 CSRNAB development test set.

As shown in Table 2, an improvement of approximately 4% was obtained for the male speakers. No change was observed for the females.

5. Acronyms as Phrases

We introduced phrase grammars to represent acronyms such as "N.A.S.A." and "G.A.T.T.". A list of acronyms was automatically derived from the language model training data by searching for sequences of individual letters. The frequencies of the resulting acronym phrases were counted and plotted. (The language data we used to calculate the frequencies was the DJIS data from 1991-94.) A distinct knee in the curve was visible (see Figure 7). We selected a cutoff well beyond it, at a frequency of 200. As a result, we included a total of 247 acronyms. The default pronunciation for each acronym was defined as the concatenation of its component letter pronunciations. A small subset of the acronyms were given alternate

[4] Please contact weide@cs.cmu.edu for information.

pronunciations – "G.A.T.T.", for example, was given an alternate "gat" (rhymes with "cat"), as well as its original, "jee ay tee tee". We tokenized the acronyms in all of the DJIS training data.

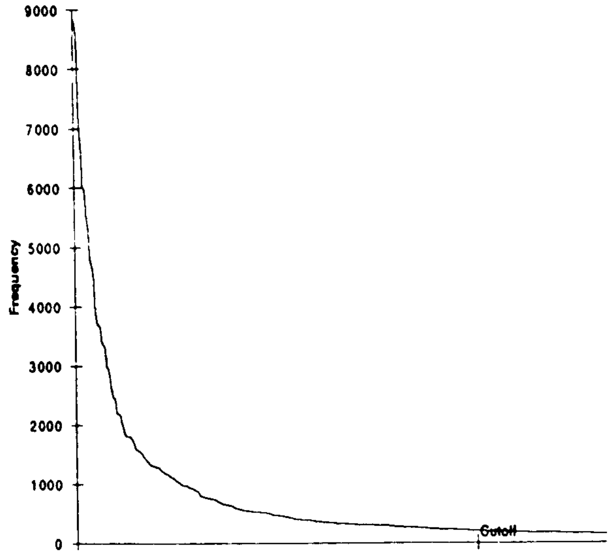

Figure 7: Frequency curve of automatically derived acronyms in the DJIS language training data, 1991-94. The frequency values for the first 350 acronyms are graphed, and the cutoff point for the frequency of 200 is marked. (The first acronym, "U.S.", does not appear on the graph, as its frequency is an order of magnitude greater than the second, which is "G.M.".)

This method consistently corrected a small number of errors without introducing new ones – in the 1994 development test set we corrected 10 errors and introduced one (1).

6. Simple Class Grammar for Verbalized Pronunciation

Two small class grammars were defined for the most common verbalized pronunciation types – double quotes and parentheses. Alternate forms of verbal production of each of these markers were allowed, and a post-processing step in the decoder substituted lexical tokens for the favored verbalized form into the decoder output. For instance, it was possible for the decoder to output "END QUOTE" when the language model score for "QUOTE" was used in decoding. This technique allowed us to correct approximately six errors in the evaluation test set, or about 0.1% absolute WER improvement.

7. Corrected Phonetic Modeling for Single Phone Words

In previous versions of the Sphinx-II decoder, single phone words were modelled with full cross-word phones as the left acoustic context but with only the silence (SIL) phone model as the right context. A correction was made to the search algorithm to support full cross-word phone modeling in the right context for this set of words. Development testing of this change was on a baseline system that used the standard 10K senone known gender acoustic models, the 1994 standard 20k vocabulary, all available language training data, a bigram cutoff of one (1), and a trigram cutoff of three (3). As reported in Table 3, a consistent improvement of approximately 3% relative was found on two development test sets for both genders.

language model	dt-94-m	dt-94-f	dt-93-m	dt-93-f
dev baseline	18.7%	13.6%	18.5%	13.9%
with full RC	18.2%	13.2%	18.0%	13.8%

Table 3: The effect of full right context acoustic modeling for single phone words. Word error rates are on the Nov 1994 CSRNAB development test set and 1993 WSJ development test set using 10K decision-tree based senones with known gender. The baseline language model is on the standard 20K word 1994 vocabulary. The language model training data for the baseline language model is all of the available data.

References

1. Rosenfeld, R. *The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation.* in: **Proc. ARPA Spoken Language Technology Workshop**. 1995.
2. Rudnicky, A. *Personal communication.* 1994.
3. Gauvain, J., L., L., and Adda-Decker, M. *Developments in Continuous Speech Dictation using the ARPA WSJ Task.* in: **ARPA Spoken Language Technology Workshop**. 1994.
4. Chase, L. L., Rosenfeld, R., and Ward, W. *Error-Responsive Modifications to Speech Recognizers: Negative N-grams.* in: **The International Conference on Spoken Language Processing**. 1994.
5. Spiegel, M. *The Orator System User's Manual - Release 10.* January 1992.
6. Corporation, D. E. *DecTalk DTC01: Owner's Manual.* Maynard, MA, 1984.
7. Rosenfeld, R. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach.* Ph.D. thesis, School of Computer Science, Carnegie Mellon University, 1994. *Published as Technical Report CMU-CS-94-138.*
8. 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.