BYBLOS SPEECH RECOGNITION BENCHMARK RESULTS

F. Kubala, S. Austin, C. Barry, J. Makhoul, P. Placeway, R. Schwartz

BBN Systems and Technologies
Cambridge MA 02138

ABSTRACT

This paper presents speech recognition test results from the BBN BYBLOS system on the Feb 91 DARPA benchmarks in both the Resource Management (RM) and the Air Travel Information System (ATIS) domains. In the RM test, we report on speaker–independent (SI) recognition performance for the standard training condition using 109 speakers and for our recently proposed SI model made from only 12 training speakers. Surprisingly, the 12–speaker model performs as well as the one made from 109 speakers. Also within the RM domain, we demonstrate that state-of-the-art SI models perform poorly for speakers with strong dialects. But we show that this degradation can be overcome by using speaker adaptation from multiple-reference speakers. For the ATIS benchmarks, we ran a new system configuration which first produced a rank–ordered list of the N–best word–sequence hypotheses. The list of hypotheses was then reordered using more detailed acoustic and language models. In the ATIS benchmarks, we report SI recognition results on two conditions. The first is a baseline condition using only training data available from NIST on CD-ROM and a word-based statistical bi-gram grammar developed at MIT/Lincoln. In the second condition, we added training data from speakers collected at BBN and used a 4-gram class grammar. These changes reduced the word error rate by 25%.

INTRODUCTION

This paper will present new test results for running the BBN BYBLOS system on the speech recognition benchmarks in both the Resource Management (RM) and the Air Travel Information System (ATIS) domains.

During this reporting period we have concentrated on speaker-independent recognition conditions. However, we will also report a new result demonstrating the need and usefulness of speaker adaptation in order to be able to recognize the speech of speakers with different dialects than those found in the training data.

For the RM corpus, we report on three conditions:

1. The common SI-109 training condition that has been widely reported in the past.
2. The new SI-12 training paradigm that we introduced at the previous DARPA workshop.
3. Adaptation to the dialect of the speaker

The ATIS domain presents a new type of speech recognition problem in several respects. First of all, and most importantly, the speech was collected during simulations of actual use of the ATIS system. The speakers were completely uncoached, and therefore, the range of speech phenomena goes far beyond that of the carefully controlled read-speech conditions that exist in the RM corpus. We will describe our recent efforts to deal with these new problems.

Since *understanding* is the ultimate goal of the ATIS domain, we use a rank ordered list of the N–best speech recognition hypotheses as the interface to the natural language component of BBN's spoken language system. Below, we describe a new procedure which allows the system to use powerful but computationally prohibitive acoustic models and statistical grammars to reorder the hypotheses in the N–best list.

For the ATIS corpus, we report on two conditions:

1. A baseline control condition using a standard training set, lexicon, and bi-gram grammar.
2. An augmented condition using additional training, acoustic models for non-speech phenomena, and a 4–gram class grammar.

In the next section, we describe the main features of the baseline Byblos system used in both RM and ATIS tests. Next, the RM results are presented. For the ATIS domain, we first describe the speech corpus used. Then we describe the informal baseline training condition which was developed to provide informative *controlled experiments* for this domain. Next, we explain how the Byblos system was modified for this evaluation. Finally, we describe our *augmented* condition and present comparative results.

BYBLOS SYSTEM DESCRIPTIOPN

The BBN BYBLOS system had the following notable characteristics for the Feb 91 evaluation:

- Speech was represented by 45 spectral features: 14 cepstra and their 1st and 2nd differences, plus normalized energy and its 1st and 2nd difference.
- The HMM observation densities were modeled by tied Gaussian mixtures. The mixture components were defined by K-means clustering and remained fixed during training.
- Context-dependent phonetic HMMs were constructed from triphone, left-diphone, right-diphone, and phoneme contexts and included cross-word-boundary contexts. The individual context models were trained jointly in forward-backward.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1991		**00-00-1991 to 00-00-1991**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Byblos Speech Recognition Benchmark Results	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **BBN Technologies,10 Moulton Street,Cambridge,MA,02238**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**6**	

- Cooccurrence smoothing matrices were estimated from the triphone contexts only and then were applied to all HMM observation densities.
- Gender-dependent models were used for SI recognition. Each test sentence was decoded on both models and the final answer was chosen automatically by picking the one with the higher probability.
- For all SI results other than the 109 RM condition, SI models were created by combining multiple, independently-trained and smoothed, speaker-dependent (SD) models. For the SI 109 condition, however, the training data was simply pooled before training.

This baseline system is the same for both RM and ATIS results reported below.

RESOURCE MANAGEMENT RESULTS

SI Recognition with 109 Training Speakers

At the June 90 DARPA workshop, we reported our first result on the standard 109 SI training condition of 6.5% word error on the Feb 89 SI test set using the word–pair grammar. When we retest with the current system the error rate is reduced to 4.2%. For these two tests, the system differed in three ways. First, we augmented the signal representation with vectors of second–order differences for both the cepstral and energy features. Secondly, the discrete observation densities of the phonetic HMMs were generalized to mixtures of Gaussians that were tied across all states of all models, as in [2], [5]. The VQ input to the trainer preserved the 5 most probable mixture components for each speech frame. In the earlier system, only the input to the decoder was modeled as a mixture of Gaussians. To date, we have not found any improvement for re-estimating the parameters of the mixture components within forward–backward. Finally, in the newer system, we trained separate codebooks and HMMs for male and female talkers and selected the appropriate model automatically. We observed a small additive improvement for each of these three changes to the system.

For the current Feb 91 evaluation, we ran our latest system on the standard 109 SI training condition with both *no–grammar* and the *word–pair–grammar*. Results for these runs are shown in the first two rows of table 1 below.

# Training Spkrs	Grammar	% Word Err	% Sent Err
109	None	18.8	69
109	Word–Pair	3.8	21
12	Word–Pair	3.8	23

Table 1: Resource Management SI recognition results— Feb 91 Test Set.

SI Recognition with 12 Training Speakers

Since it is often difficult and expensive to obtain speech from hundreds of speakers for a new domain, we recently proposed [7] creating SI models from a much smaller number of speakers but using more speech from each speaker. To test the proposal, we ran an experiment using 600 utterances from each of the 12 speakers in the SD segment of the RM corpus.

12 speakers could hardly be expected to cover all speaker types in the general population (including both genders), so we anticipated that smoothing would be needed to make the models robust for new speakers. Our usual technique for smoothing across the bins of the discrete densities, triphone cooccurrence smoothing [10], has proven to be an effective method for dealing with the widely varying amounts of training data for the detailed context–dependent models in the Byblos system. This technique estimates the probability of any pair of discrete spectra cooccurring in a density by counting over all the densities of the triphone HMMs. These probabilities are organized into a set of phoneme–dependent confusion matrices which are then used to smooth all the densities in the system.

The data from each training speaker is kept separate through forward–backward (Baum–Welch) training and cooccurrence smoothing. The individually trained SD HMMs are then combined by averaging their parameters. We have found that this approach leads to better SI performance from a small number of training speakers than the usual practice of pooling of all the data prior to training and smoothing.

In table 1, we show that the model made from 12 training speakers performs as well as the standard 109 speakers on the Feb 91 SI test set. This is better than we had expected based on our previous experience with the Feb 89 SI test set. To get a better estimate of the relative performance of the two approaches, we tested the current system on three evaluation test sets (Feb 89, Oct 89, Feb 91). Averaging the results for these 30 test speakers, the SI 109 model achieved 3.9% word error while the SI 12 model got 4.5%. This is a very small degradation for nearly a 10–fold decrease in the number of training speakers.

Adaptation to Dialect

We found that our current state-of-the-art (SI) models perform poorly when a test speaker's characteristics differ markedly from those of the training speakers. The speaker differences which cause poor recognition are not well understood, but outliers of this sort are not a rare phenomenom. Our SI models have difficulty with the RM test speaker RKM, for instance, a native speaker of English with an African-American dialect. Moreover, non-native speakers of American English nearly always suffer significantly degraded SI performance.

The problem was made obvious in a pilot experiment we performed recently. As noted above, our baseline SI performance is currently about 4% word error using the standard 109 training speakers and word-pair grammar. But when we tested four non-native speakers of American English under the same conditions, the word error rates ranged from 22% to 45%, as shown in table 2.

Speaker	Gender	Native Language	Years of English	SI 109 % Wd Err	Adapted % Wd Err
JM	male	Arabic	>25	27.6	5.2
MS	male	Cantonese	5	45.4	10.7
SA	male	British Eng.		31.7	5.4
VS	female	Hebrew	>15	22.2	4.7
		Average		31.7	6.5

Table 2: SI and SA results for non–native speakers of English.

The table also shows the native language of each speaker and the number of years that each has spoken English as their primary language. Even though they vary widely in their experience with

English (and in their subjective intelligibility), each of them suffered severely degraded SI performance. Even native speakers of British English are subject to this degradation, as the result from speaker SA demonstrates. Furthermore, the result from speaker JM shows that this problem does not disappear even after many years of fluency in a second language.

We then tried to adapt the training models to the new dialects by estimating a probabilistic spectral mapping between each of the training speakers and the test speaker as described in [7]. The resulting set of adapted models are combined to obtain a single adapted model for the test speaker. In this experiment, we used the 12 SD speakers from the RM corpus as training speakers. Each test speaker provided 40 utterances for adaptation and 50 additional ones for testing. The word error rates after adaptation are also shown in table 2. The overall average word error rate after speaker adaptation is 5 times better than SI recognition for these speakers!

The success of speaker adaptation in restoring most of the performance degradation is quite surprising given that no examples of these dialects are included in the training data. Furthermore, only spectral and phonetic differences are modeled by our speaker adaptation procedure. No lexical variations were modeled directly; we used a single speaker-independent phonetic dictionary of American English pronunciations. These results show that systematic spectral and phonetic differences between speakers can account for most of the differences in the speech of native and non-native speakers of a language.

THE ATIS CORPUS

Corpus Description

The ATIS corpus consists of several different types of speech data, collected in different ways. First, there are approximately 900 utterances that were collected during a "Wizard" simulation of an actual ATIS system. The subjects were trying to perform a particular task using the system. This data was collected from 31 speakers. The data from five of the speakers was used for the test of the natural language systems prior to the June 1990 meeting. These have since been designated as the development test speech. Thus, there remained 774 spontaneous training sentences from 26 speakers.

In addition to the spontaneous sentences, several of the subjects read cleaned up versions of their spontaneous queries. Finally, 10 subjects each read 300 sentences during a single 1-hour session. The first 100 sentences were read by all the subjects. The next 200 were chosen at random from a list of 2800 sentences constructed by BBN and SRI, by generalizing from previously recorded sentences from several sites. The average total duration of the 300 sentences was about 18.5 minutes per speaker (counting only the parts of the utterances containing speech).

The 774 sentences from a total of about 30 speakers is clearly not sufficient for creating a powerful speaker-independent speech model. Collecting speech from an additional 70 speakers would require a large additional effort. Therefore, the additional 3000 sentences read by the 10 speakers provided the most efficient source of speech for estimating a speaker-independent model.

ATIS Training Speech Characteristics

The subjects were instructed to push a button (push-to-talk) before speaking. However, frequently, they pushed the button quite a while before they began speaking. In many cases, they breathed directly on the microphone windscreen, which was apparently placed directly in front of the mouth and nose. Therefore, many files contain long periods of silence with interspersed noises. In fact, only 55% of the total duration of the training data contains speech (for both the read and spontaneous data). In addition, some subjects paused in the middle of a sentence for several seconds while thinking about what to say next. Others made side comments to themselves or others while the microphone was live. All of these effects are included in the speech data, thus making it much more difficult to recognize than previously distributed data.

In the RM corpus, there was a concerted effort to use subjects from several dialectal regions. In addition, since the speakers were reading, they tended toward standard General American. Thus, the models generated from this speech were reasonably robust for native speakers of American English. In contrast, the ATIS corpus consisted primarily of speakers from the South (26 of 31 speakers were labeled South Midland or Southern).

In order to estimate speech models, we need an accurate transcription of what is contained in each speech file. This transcription usually is in the form of the string of words contained in each sentence. However, since this was spontaneous speech, there were often nonspeech sounds and long periods of silence included among the words. Most of these effects were marked for the spontaneous speech. Unfortunately, the transcriptions distributed with the read speech did not follow the usual conventions for the string of words. A significant amount of work was required to correct these inconsistencies. This work was undertaken by BBN and NIST, and was thoroughly checked by Lincoln. When all the corrections had been made, they were redistributed to the community.

Definition of Common Baseline

The new ATIS task presents several new problems for speech recognition. Therefore, it will be essential to try many new algorithms for dealing with it. These experiments will deal with a wide variety of topics, including the makeup of the training data, the vocabulary, and the grammar. However, it is just as important with this domain, as it was with the RM domain, that we use well-founded controlled experiments across the different sites. Without a baseline, meaningful comparisons between techniques cannot be made. It is necessary in order for researchers at other sites to be able to determine whether a new technique has actually made a significant improvment over previous techniques.

Since no common evaluation condition has been specified by the speech performance evaluation committee, BBN and MIT/Lincoln have defined, promoted, and distributed an ATIS control condition to provide a common reference baseline. This baseline condition consists of a common lexicon, training set, and statistical grammar. In order to provide a useful baseline condition, all of the standardized data should represent a reasonable approximation to current state-of-the-art conditions. BBN and Lincoln undertook to define such a baseline condition, under the severe constraints of limited available data and time.

Training Set We defined as the baseline training set, all of the speech that had been distributed by NIST on CD-ROM excepting the spontaneous speech spoken by the speakers used for the June 1990 test of ATIS. This test consisted of 138 sentences spoken by a total of 5 speakers, (bd, bf, bm, bp, bw). While an additional 435 sentences

that been recorded at SRI were made available on tapes at a later date, we felt that the small amount of additional speech and the late date did not warrant including the speech in the baseline conditon. Of course the data was available to all who wanted to use it in any other experimental or evaluation condition.

Vocabulary One of the variables in designing a real speech system is to specify the recognition vocabulary. Given that we do not know what words will be included in the test speech, we have to make our best attempt to include those words that would seem reasonably likely. Of course, if we include too many words, the perplexity of the grammar will increase, and the recognition error rate will increase. We felt that, for a baseline condition, the vocabulary must be kept fixed, since we wanted to avoid random differences between sites due to correctly guessing which words would occur in the test. We decided, at BBN to define a standard vocabulary based on the transcriptions of all of the designated training data. Thus, all of the words included in the Standard Normal Orthographic Representation (SNOR) were included in the dictionary. We made sure that many fixed sets of words, such as the days, months, and numbers were complete. In addition, we filled out many open class word categories based on the specific ATIS database that was being used. This included plurals and possessive forms of words wherever appropriate. This included names of airlines, plane types, fare codes, credit cards, etc. When we did this completely, the result was a vocabulary of over 1700 words, most of which seemed quite unlikely. Therefore, we applied an additional constraint on new open class words. We added to the vocabulary only the names of all of the airlines and plane types, etc., that served the 10 cities whose flights were included the current database. In total, we added about 350 words to the vocabulary actually used in the training speech. This brought the baseline vocabulary up to 1067 words. The result, when measured on the development set, was that the number of words in the test that were not in the vocabulary was decreased from 13 to 4.

Grammar While the definition of the grammar to be used in speech recognition is certainly a topic of research, it is necessary to have a baseline grammar with which any new grammars may be compared. It is also essential that this standard grammar be relatively easy for most sites to implement, in order that this not be an impediment to the use of the baseline condition. Therefore, Lincoln estimated the parameters of a statistical bigram grammar using the back-off technique developed by IBM [6]. The derivation of this grammar is described in more detail in [9]. The transcriptions used to estimate this grammar included those of all of the speech in the designated training set (SNOR transcriptions only) and also used the 435 transcriptions for the new SRI set. The parameters of this model were distributed in simple text format so that all sites could use it easily. The grammar has a test set perplexity of 17 when measured on the development test set. Thus, it provided a low upper bound for comparison with any new language models.

BBN SPEECH TECHNIQUES USED FOR ATIS

In this section we describe the techniques that we used in the ATIS speech recognition evaluation. In particular, we only discuss those techniques that differed from those used for RM. The techniques include:

1. Speech/silence detection.
2. N-Best recognition and rescoring with detailed models.
3. Optimization.

Each of these techniques will be described below.

Speech/Silence Detection

As described in a previous section, both the training and test speech contained large regions of silence mixed with extraneous noises. While the HMM training and recognition algorithms are capable of dealing with a certain amount of silence and background noise, they not very good at dealing with large periods of silence with occasional noise. Therefore, we applied a speech end–point detector as a preprocess to the training and recognition programs. We found that this improved the ability of the training algorithms to concentrate on modeling the speech, and of the recognition programs to recognize sentences.

N-Best Recognition

Since the purpose of the speech recognition is to understand the sentences, we needed to integrate it with the natural language (NL) component of the BBN HARC spoken language understanding system. For this we use the N-Best recognition paradigm [3]. The basic steps are enumerated below:

1. Find N-Best hypotheses using non-cross-word models and bigram grammar
2. For each hypothesis:
   (a) rescore acoustics with cross-word models
   (b) score word string with a more powerful statistical grammar
3. Combine scores and reorder hypotheses
4. Report highest scoring answer as speech recognition result
5. Feed ordered list to NL

For efficiency, we use the Word-Dependent N-Best algorithm [11]. In addition to providing an efficient and convenient interface between speech and NL, the N-Best paradigm also provides an efficient means for applying more expensive speech knowledge sources. For example, while the use of cross-word triphone models reduces the error rate by a substantial factor, it greatly increases the storage and computation of recognition. In addition, a trigram or higher order language model would immensely increase the storage and computation of a recognition algorithm. However, given the N-Best hypotheses obtained using non-cross-word triphone models, and a bigram grammar, each hypothesis can be rescored with any knowledge source desired. Then, the resulting hypotheses can be reordered. The top scoring answer is then the speech recognition result. The entire list is then sent to the NL component, which chooses the highest answer that it can interpret.

By using the N-Best paradigm we have found it efficient to apply more expensive knowledge sources (as a post process) than we could have considered previously. Other examples of such knowledge sources include: Stochastic Segment Models [8] or Segment Neural Networks [1].

Optimization

We usually run the recognition several times on development test data in order to find the optimal values for a few system parameters, such as the insertion penalty, and the relative weight for the grammar and acoustic scores. This is a very slow and inexact process. However, given the N-Best paradigm, it is a simple matter to find

the values that maximize recognition accuracy. Briefly, we generate several hypotheses for each utterance. For each hypothesis, we factor the total score into a weighted combination of the acoustic score(s), the language model score, and the insertion penalty. Then, we search for the values of the the weights that optimize some measure of correctness over a corpus. This technique is described more fully in [8].

ATIS BBN AUGMENTED CONDITION

We decided to consider three different conditions beyond those specified in the common baseline condition. These include:

1. Use of additional training speech
2. Inclusion of explicit nonspeech models
3. More powerful statistical grammars

Additional training speech

One of the easiest ways to improve the accuracy of a recognition system is to train it on a larger amount of speech, from a representative sample of the population that will use it. Since there was clearly not time to record speech from a very large number of speakers, we decided to record a large amount of speech from a smaller number of speakers. We had shown previously [7] that this training paradigm results in similar accuracy, with a smaller data collection effort (since the effort is largely proportional to the number of speakers rather than the total amount of speech.)

We collected over 660 sentences from each of 15 speakers. Five were male and ten were female. Due to the lack of time, most of the speakers were from the northeast. However, we made an effort to include 4 female speakers from the Southeast and South Midland regions. We found that, once started, the subjects were able to collect about 300 sentences per hour comfortably.

Nonspeech Models

One of the new problems in this speech data is that there were nonspeech sounds. Some were vocal sounds (e.g. "UH", "MM", etc.), while some were nonvocal (e.g. laughter, coughing, paper rustling, telephone rings, etc.). The only frequent nonspeech sound was "UH", with 57 occurrences in the training corpus. All the rest occurred only 1 to 5 times. We created a separate "word" for each such event. Each consisted of it's own special phoneme or two phonemes. All of them were included in the same language model class within the statistical language model.

While several of the nonspeech events were correctly detected in the development test speech, we found that the false alarm rate (i.e. typically recognizing a short word like "a" as "UH") was about equal to the detection rate. Thus, there was no real gain for using nonspeech models in our development experiments.

Statistical Language Models

In this section, we discuss the use of statistical language models that have been estimated from very limited amounts of text. We argue that it is clearly necessary to group words into classes to avoid robustness problems. However, the optimal number of classes seems to be higher than expected.

Since there is essentially no additional cost for using complex language models within the N-Best paradigm, we decided to use a 4-gram statistical class grammar. That is, the probability of the next word depends the classes of the three previous words.

Need for Class Grammars An important area of research that has not received much attention is how to create powerful and robust statistical language models from a very limited amount of domain-dependent training data. We would certainly like to be able to use more powerful language models than a simple word-based bigram model.

Currently, the most powerful "fair" grammars used within the program have been statistical bigram class grammars. These grammars, which use padded maximum likelihood estimates of class pairs, allow all words with some probability, and share the statistics for words that are within the same domain-dependent class. One issue of importance in defining a class grammar is the optimal number of classes into which words should be grouped. With more classes we can better distinguish between words, but with fewer classes there is more statistical sharing between words making the grammar more robust. We compared the perplexity with three different grammars for the RM task with 100 classes, 548 classes, and 1000 classes respectively. In the first, words were grouped mainly on syntactic grounds, with additional classes for the short very common words. In the second, we grouped into classes only those words that obviously belonged together. (That is, we had classes for shipnames, months, digits, etc.) Thus, most of the classes contained only one word. In the third grammar, there was a separate class for every word, thus resulting in a word-bigram grammar. We used the backing off algorithm to smooth the probabilities for unseen bigrams. The perplexities of the three grammars measured on training data and on independent sentences are given in the table below.

	Number of Classes		
	100	548	1000
Training	79	20	14
Test	95	42	49

Table 3: Perplexity for three bigram class grammars measured on the training and test set.

As shown in table 3, perplexity on the training set decreases as the number of classes increases, which is to be expected. What is interesting is the perplexity on the test set. Of the three grammars, the 548-class grammar results in the lowest test set perplexity. (Interestingly, the 548-class grammar is easier to specify than the 100-class grammar.) The increased perplexity for the 1000-class grammar is due to insufficient training data.

The effective difference between the 548- and 1000-class grammars was larger than implied by the average perplexity. The standard deviation of the word entropy was one half bit higher, which resulted in a doubling in the standard deviation of the perplexity. To explain, the word bigram grammar frequently has unseen word pairs with very low probability, while this effect is greatly reduced in the class grammar. Thus, as expected, the class grammar is much more robust. Initial recognition experiments also seem to indicate a factor of two difference in error rate between a class bigram grammar and a word bigram grammar of the same perplexity. These effects are likely to be even larger when we use higher order n-gram models.

ATIS RECOGNITION RESULTS

The table below contains the recognition results for the ATIS corpus for both the development test set and the evaluation set. The first line shows the recognition results for the development test set consisting of 138 sentences spoken by five speakers (bd, bf, bm, bp, bw). All speech data from these five speakers was left out of the training. The development results are given for the "augmented" condition only. Next, we give the results for the evaluation test set. The first two results are the baseline condition and our augmented condition. We also give results separately for the subset of 148 sentences that were designated as Class A (unambiguous, context–independent queries) for the NL evaluation.

To review the two basic conditions, the baseline condition used the standard vocabulary, training set, and grammar throughout. The augmented condition used more training data, a 4-gram class grammar, and a nonspeech model.

Condition	Corr	Sub	Del	Ins	Word Err	Sent Err
Augmented; all Dev	92.2	6.1	1.6	1.6	9.4	46.2
Baseline; all 200	80.2	16.2	3.6	6.1	25.8	73.5
Augmented; all 200	84.2	12.6	3.2	4.7	20.5	60.5
Baseline; ClassA	82.5	14.5	3.0	5.3	22.8	67.6
Augmented; ClassA	87.6	9.9	2.6	3.7	16.1	54.5

Table 4: ATIS speech recognition results.

The first clear result is that the error rates for the evaluation test set are more than twice those of the development test set. In addition, the perplexity of the evaluation test set is significantly higher than for the development set (26 instead of 17 for the standard word–based bigram grammar, and 22 instead of 13 for the 4-gram class grammar). Thus, we surmise that the evaluation data is somehow significantly different than both the training data and the development test set.

Next, it is clear that the Class A subset of the sentences presents fewer problems for the recognition. This is also indicated in the perplexities we computed for the two subsets.

Finally, we see that, for both the full set of 200 sentences and the Class A subset of 148, the augmented condition has about 20%-30% fewer word errors than the baseline condition. We are currently attempting to understand the causes of this improvement by more careful comparison to the baseline. The augmented condition was rerun after including the training data from the held–out development test speakers (about 900 utterances), but this made no difference. We suspect, therefore, that very little gain was also derived from the additional training speech collected at BBN (which suffers from both environmental and dialectal differences). We have also retested with a *class* bigram grammar instead of the 4–gram, and again, there was no change in performance. This behavior may be explained by the large difference between the evaluation test and the training. It is interesting, then, that the higher order grammar did not *degrade* in the presence of such a difference. This result also indicates that smoothing a word–based bigram by class definitions is important for training statistical grammars from small training corpora. We have not retested without the nonspeech models, but their contribution appears small from a preliminary review of the recognition errors made. The two worst test speakers were also the ones that tended to produce numerous pause fillers (e.g. "UH", "UM") as well as many other disfluencies. Clearly, better nonspeech modeling will be essential if we continue to evaluate on this kind of data.

CONCLUSIONS

We have reported several new benchmark speech recognition results for both the RM corpus and the new ATIS corpus. On RM, using the standard 109 speaker training set and the word-pair grammar, the word error rate for the BYBLOS system was 3.8%. Surprisingly our new SI paradigm, using only 12 training speakers, achieved the same result! In addition, we have demonstrated that SI performance is generally very bad for speakers with strong dialects. But we have achieved a 5–fold reduction in error rate for these speakers by using speaker adaptation from only 40 training utterances.

For the ATIS corpus we developed several new techniques based on the N-Best paradigm. These have allowed us to use cross-word triphone models and a 4-gram statistical grammar efficiently in the recognition. We have improved performance over a baseline condition by 20%-30% by using additional training, models of nonspeech, and a 4–gram class grammar. Our preliminary conclusion is that most of this gain is due to the smoothing of the grammar by classes. The spontaneous speech effects that appear in this corpus clearly present a new set of difficult problems, since the error rates are about 4 times higher than for the RM corpus.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Office of Naval Research under Contract No. N00014-89-C-0008.

REFERENCES

[1] S. Austin, J. Makhoul, R. Schwartz and G. Zavaliagkos, "Continuous Speech Recognition Using Segmental Neural Nets," this proceedings.

[2] Bellegarda, J., D. Nahamoo, "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Dec. 1990, Vol. 38, No. 12.

[3] Chow, Y-L. and R.M. Schwartz, "The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses," *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Inc., Oct. 1989.

[4] Feng, M., F. Kubala, R. Schwartz, J. Makhoul, "Improved Speaker Adaptation Using Text Dependent Spectral Mappings," *IEEE ICASSP-88*, paper S3.9.

[5] Huang, X., K. Lee, H. Hon, "On Semi-Continuous Hidden Markov Modeling," *IEEE ICASSP-90*, Apr. 1990, paper S13.3.

[6] Katz, S., "Estimation of Probabiliities from Sparse Data for the Language Model Component of a Speech Recognizer", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Mar. 1987, Vol. 35, No. 3.

[7] Kubala, F., R. Schwartz, "A New Paradigm for Speaker-Independent Training and Speaker Adaptation," *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Inc., Jun. 1990, pp. 306–310.

[8] Ostendorf, M., Kannan, A., Austin, S., Kimball, O., Schwartz, R., and J.R. Rohlicek. "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses" this proceedings.

[9] D. B. Paul, "New Results with the Lincoln Tied-Mixture HMM CSR System," this proceedings.

[10] Schwartz, R., O. Kimball, F. Kubala, M. Feng, Y. Chow, C. Barry, J. Makhoul, "Robust Smoothing Methods for Discrete Hidden Markov Models," *IEEE ICASSP-89*, May 1989, paper S10b.9.

[11] Schwartz, R.M., and S.A. Austin, "Efficient, High-Performance Algorithms for N-Best Search," *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Inc., Jun. 1990.