# BBN BYBLOS and HARC February 1992 ATIS Benchmark Results

*Francis Kubala, Chris Barry, Madeleine Bates, Robert Bobrow, Pascale Fung, Robert Ingria, John Makhoul, Long Nguyen, Richard Schwartz, David Stallard*

BBN Systems and Technologies
Cambridge MA 02138

## ABSTRACT

We present results from the February '92 evaluation on the ATIS travel planning domain for HARC, the BBN spoken language system (SLS). In addition, we discuss in detail the individual performance of BYBLOS, the speech recognition (SPREC) component.

In the official scoring, conducted by NIST, BBN's HARC system produced a weighted SLS score of 43.7 on all 687 evaluable utterances in the test set. This was the lowest error achieved by any of the 7 systems evaluated.

For the SPREC evaluation BBN's BYBLOS system achieved a word error rate of 6.2% on the same 687 utterances and 9.4% on the entire test set of 971 utterances. These results were significantly better than any other speech system evaluated.

## 1. OVERVIEW

The BBN HARC spoken language system consists of BYBLOS, the speech recognition component, and DELPHI, the natural language processing component. In this paper, we concentrate on BYBLOS and its interaction with DELPHI through the N-best interface. Results are presented for speech recognition alone and for the overall spoken language system. A detailed discussion of DELPHI is presented in [2,3] elsewhere in these proceedings.

## 2. BYBLOS – SPEECH RECOGNITION

The BYBLOS speech recognition system produces an ordered list of the N top-scoring hypotheses which is then reordered by several detailed knowledge sources. The N-best strategy [4,8] permits the use of computationally prohibitive models by greatly reducing the search space to a few dozen word sequences. It has enabled us to use cross-word-boundary triphone models and trigram language models with ease. The N-best list is also a robust interface between speech and natural language that provides a way to recover from speech errors in the top choice word sequence.

The overall system architecture for this evaluation is similar to that used in the February '91 tests [6]. Specifically, we use a 4-pass approach to produce the N-best lists for natural language processing.

1. A forward pass with a bigram grammar and discrete HMM phonetic models saves the top word-ending scores and times.
2. A backward pass with a bigram produces an inital N-best list.
3. Rescoring each of the N sentence hypotheses with cross-word-boundary triphones and semi-continuous density HMMs reorders the N-best list.
4. Rescoring with a trigram grammar reorders the N-best list again.

Each utterance is quantized and decoded twice, once with each gender-dependent codebook and model. For each utterance, the N-best list with the higher top-1 hypothesis score is chosen. Then they are passed to DELPHI for further reordering and interpretation while the top choices in the lists constitute the SPREC results reported here.

### 2.1 Training and Development Test Data

We used speech data from the ATIS2 subcorpus exclusively to train the parameters of the acoustic model. This subcorpus consists of 10411 spontaneous utterances from 286 subjects. The data originated from 5 collection sites using a variety of strategies for eliciting and capturing spontaneous queries from the subjects [7]. The training data was not balanced across the five sites, however. MIT was represented by 3–4 times as much data as any other site. Overall, MIT data accounted for nearly half of the ATIS2 subcorpus (4600 utterances).

The evaluation test data was drawn from this same pool of data so we decided to ignore the earlier batches of ATIS data that were collected under still different circumstances (most of it was read speech) and would not be represented in the new test (dialects were predominantly southern in the ATIS0 subcorpus).

We filtered the training data for quality in several ways. All utterances that were marked as truncated in the SRO (speech recognition output) transcription were ignored. Similarly, we omitted from the training all utterances that contained a word fragment. We also ignored any utterances that contained rare nonspeech events. Finally, our forward-backward training program rejected any input that failed to align properly. These steps removed about 1200 utterances from consideration.

Another 600 utterances were removed due to name conflicts between a number of subjects from AT&T and MIT that were given identical speaker codes, thus making it difficult to match the speech utterances with the transcriptions.

We held another 890 utterances out of the training as a development test set. We included 2 male and 2 female subjects from each of the 5 collection sites in this set. Each speaker had roughly 40 utterances.

This left a total of 7670 utterances from 237 speakers for training the HMMs. Since we train gender-dependent models, the training was further divided into 3317 utterances for the female speakers and 4349 for the males.

For statistical language model training we used all available sentence texts from ATIS0, ATIS1, and ATIS2. During the development phase, we excluded the 890 sentences from the held-out test set. For the final evaluation this data was included, resulting in a total of 14500 sentences for training the language models.

## 2.2 Recognition Lexicon and Grammars

The lexicon used for recognition was initialized by including all words observed in the complete grammar training texts. This had the side-effect of including the entire development test set within the vocabulary. Common closed-classes of words such as days of the week, months, numbers, plane types, etc. were completed by hand. Similarly, we included derivations (mostly plurals and possessives) of many open-class words in the domain. We also added about 400 concatenated word tokens for the commonly occurring sequences such as WASHINGTON_D_C, SAN_FRANCISCO, or D_C_TEN. The final size of the lexicon was 1881 words. For the February '92 evaluation test set only 35 words, occurring 42 times, were out-of-vocabulary (OOV) for this lexicon. This is only a 0.4% OOV word occurrence rate over the whole test set.

We estimated the parameters of our statistical bigram and trigram grammars using a backing-off procedure similar to Katz [5]. The N-grams were computed on word classes in order to share the very sparse training. A total of 1054 semantic classes were defined (most words remained singletons in their class). The perplexity of these grammars as measured on the evaluation test set (ignoring out-of-vocabulary words) is summarized in Table 1. The perplexities have been measured separately on each of the three sentence classes in the test. The trigram language model consistently, but rather modestly, reduced perplexity across all three classes. We did observe that recognition performance consistently improved with the trigram model.

| Sentence Class | Bigram Perplexity | Trigram Perplexity |
|---|---|---|
| A+D | 17 | 12 |
| A+D+X | 20 | 15 |
| A | 15 | 10 |
| D | 20 | 14 |
| X | 35 | 28 |

Table 1: N-gram perplexities on the February '92 test set.

More striking are the differences between the perplexities of the three sentence classes and the large values for the Class X sentences (those which are unevaluable with respect to the database). We observed that our recognition performance was well correlated with these measured perplexities.

## 2.3 Automatic Endpointing

We estimated that 35% of the entire corpus is ambient noise. Part of the high ambient-to-speech ratio is due to the various ways in which the data was collected.

Several different strategies were employed by the collecting sites for endpointing the waveforms, including subject-initiated push-to-talk and push-and-hold, as well as automatic endpointing, and wizard-initiated manual endpointing. This led to highly variable and often very long leaders of ambient noise on both ends of the waveforms. In addition, these segments frequently contained a variety of nonspeech events.

We employed the speech-detector that we use in our real-time recognizer front-end to remove most of the long duration intervals of ambient noise from the input. This step makes all subsequent processing faster and avoids many spurious word insertions in the results. The parameters of the detector were set on a small sample of ATIS data from 4 collection sites in October '91. These same parameters were used for all data processed thereafter, including all of the data from AT&T, none of which was included in the parameter tuning sample. Although we have carefully verified that the detector is working properly on only very small samples of data, we believe it is quite accurate since we do not observe many errors at the ends of utterances.

## 2.4 Automatic Location of Silence

Another reason for the prevalence of ambient noise is due to the subjects having difficulty satisfying the simulated travel planning problem given them to elicit spontaneous utterances. Hence, the ATIS2 corpus is marked with a great number of hesitations and pauses in the speech – many of which are quite long in duration.

We observed that the marking of such pauses in the SRO transcriptions was highly inconsistent. Some transcribers simply neglected to make detailed markings altogether, while others marked only the extremely long pauses. Since we do not allow silence to appear optionally between words in our training algorithm, we believed that the large number of unmarked pauses could degrade our phonetic models. Therefore, we devised a procedure to automatically locate the remaining pauses in the training data and thereby correct the corresponding transcriptions.

1. Train an initial model as usual, ignoring unmarked pauses.
2. Run the recognizer constrained to the correct answer on the training data, but allow optional silence between every word.
3. Retrain using the recognized output from (2) as the corrected transcriptions.

We found that a large positive bias was needed in step 2 above to induce the recognizer to hypothesize inter-word silences. Since the initial model was trained disallowing many real pauses, the corrupted phonetic models easily absorb many frames of silence. We adjusted the bias by comparing the recognizer's output on a sample of speech with known pause locations. Although the hypothesized pause locations never matched the truth exactly, we did observe a 15% reduction in word error rate on a very early development condition. Specifically, it improved the performance of the cross-word-boundary rescoring stage, whereas the non-cross-word N-best stage did not change. This is entirely consistent with the correction we made; only the cross-word-boundary models were corrupted by unlabeled inter-word pauses.

## 2.5 Nonspeech Events

This corpus is also notable for its large number and variety of nonspeech events audible in the waveforms. The phenomena run the gamut from common filled pauses (um's and uh's), throat clearings, coughing, and laughter, to unintelligible mutterings of 20 seconds duration. There were over 175 different markings for nonspeech events in the SRO transcriptions. While these events are typical of casual conversational speech between people, their high frequency and severity in this corpus are likely a consequence of the fact that nearly all subjects were completely new to speech and natural language technologies and had little or no training in how to speak or specific feedback about their speech quality from the system.

To handle the nonspeech events, we first identified those that appeared to have enough training samples to make a robust estimate of the HMM parameters. We then mapped a wide variety of marked events into a set of generic nonspeech tokens: [AMB_NOISE], [BREATH_NOISE], [MIC_NOISE], and [MOUTH_NOISE]. In all, we attempted to model only 11 unique nonspeech events. All nonspeech tokens were assigned to the same class in the grammar.

We tried 3 different ways to use nonspeech models in the system:

1. Treat nonspeech as normal words, including estimating N-gram probabilities for them.
2. Treat nonspeech as normal words in acoustic training, but do not include them in the grammar training, thereby making them very unlikely.
3. Treat nonspeech like silence, allowing them optionally between any words with fixed grammar transition probabilities.

Although method 3 was intuitively more appealing and known to work rather well for silence, it was the least effective of the three approaches for nonspeech. As was the case last year [6], when we try to recognize nonspeech events accurately the false alarm rate is high enough to offset any potential gain. The most successful was method 2 which effectively disallowed nonspeech in the decoder output.

Modeling nonspeech events carefully may not be important for another reason – there are not enough errors due to nonspeech events. There are only about 120 actual nonspeech events in the test SROs. There are 184 marked, but 66 of them are [POP]s at the beginnings of utterances that aren't really in the data! Apparently the transcriber was marking pops caused by the D/A system during playback.

If these markings do not greatly underestimate the true frequency of occurrence, then there is relatively little to be gained by modeling nonspeech accurately. Moreover, of the remaining 118 nonspeech events, half are breath noise or ambient noise at levels that do not interfere with our recognition.

We have indeed noticed that we make errors around most of the long or loud filled-pauses. But there are only 55 filled pauses (as indicated by the SRO) in the 971 utterance evaluation test set. These are primarily from 2–3 speakers. Given that we have nearly 1000 word errors across the entire test set, modeling filled-pauses well will have a very small impact on overall performance at this point.

## 2.6 February '92 Evaluation Conditions

The February '92 evaluation test set has data from 37 speakers. 20 subjects were female and 17 were male. The number of utterances per speaker varied from 5 to 64, but the number of utterances from each of the 5 data-collection sites was carefully balanced. All results given are for the Sennheiser channel (same as the training data). The recognition mode was speaker-independent – the test speakers were not in the training set.

By committee decision there was no common baseline control condition for the February '92 ATIS tests. The only constraint was that the single common test set must be used. Under these circumstances, there is a strong temptation to try to train on as much material as one can. We have resisted this temptation for two reasons. First, we feel that the simple addition of training data for incremental improvements is scientifically uninteresting. Secondly, in our experience with the current definition of ATIS, we have seen very little improvement for increasing the training data beyond about 4000 utterances.

Last year, we attempted to improve on the February '91 common baseline by augmenting the 3700 common acoustic training utterances with 9000 more collected from 15 speakers at BBN. The resulting performance improvement was statistically insignificant. Since the test subjects were collected at TI and had predominantly southern dialects, our conclusion was that the additional training we collected at BBN did not match the test data sufficiently to be of much use. When the training data *does* match the test, we normally expect a quadrupling of the training data to yield a halving of the error rate. This result made it clear to us that simply increasing the amount of training has limited scientific and practical value.

Recently we had another demonstration of the rather weak contribution of additional training data in the ATIS domain. As the ATIS2 data became available, we moved from a pilot development training set of 4100 utterances to our final set of 7700 utts. On a common development test set, we observed no significant gain for nearly doubling the training data, even though the additional data matched the test conditions! Moreover, we observed that data originating from a particular site primarily improved performance only on test data from the same site.

## 2.7 Speech Recognition Results

Official results for BYBLOS on this evaluation are given in Table 2. The performance is shown for two composite results and as a function of utterance class type. The 6.2% word error rate on class A+D sentences and the combined A+D+X error rate of 9.4% were significantly better than any other speech system reporting on this data. The individual speaker results varied widely from 0.0% word error to 30% error with the median at about 7.5% The female speakers got significantly better results than the male speakers.

| Sentence Class | # Sentences | # Word Errors | % Word Errors |
|---|---|---|---|
| A+D | 687 | 501 | 6.2 |
| A+D+X | 971 | 1015 | 9.4 |
| A | 402 | 305 | 5.8 |
| D | 285 | 196 | 7.0 |
| X | 284 | 514 | 17.2 |

Table 2: BYBLOS Official SPREC results on the February '92 test set.

Performance on the class X utterances is markedly worse than either class A or D utterances. In fact, more than half of the speech errors occur on these utterances. The ratio of the error rate for class X utterances to other utterances is higher than we have ever seen. Since these utterances are not evaluable by the natural language component, it does not seem profitable to try to improve the speech performance on these utterances for a spoken language system.

In Table 3 we observe a large variation in overall performance on the class A + D utterances for each segment of the test data originating at a given collection site, as shown in the rightmost column. We believe that most of this variation can be explained by two easily measured factors – amount of training data from the matching site, and the number of errors due to all spontaneous speech effects. The actual number of training utterances that we used from each site is shown in Table 3. The next column shows the word error rate that we attribute to general modeling deficiencies after removing those errors that we judged were due to spontaneous speech effects. The variation due to modeling seems well correlated to the amount of training data available from

|  |  | % Word Error Due To: |  | Overall |
|---|---|---|---|---|
|  | # Utts | Modeling | Spontaneous | % Word |
| Site | Training | Deficiency | Effects | Error |
| MIT | 3700 | 2.7 | 0.5 | 3.2 |
| BBN | 1400 | 4.5 | 0.8 | 5.3 |
| CMU | 1000 | 5.3 | 0.5 | 5.8 |
| SRI | 800 | 5.7 | 2.0 | 7.7 |
| AT&T | 800 | 6.4 | 4.0 | 10.4 |

Table 3: BYBLOS performance on February '92 test as a function of originating site (class A + D).

each site. The numbers show the expected halving of the error rate for a quadrupling of the training data. In particular, we feel that the higher performance on the MIT data can be explained entirely by the increased amount of data from that particular site.

The errors due to spontaneous speech effects in Table 3 were counted by matching the output of BYBLOS against the SRO transcriptions. The SROs contain specific markings for many spontaneous speech effects including: nonspeech events, word fragments, mispronunciations, emphatic stress, lengthening, and verbal deletions. Any error that occurred in the immediate vicinity of such a marking was counted as an error due to spontaneous speech. The table shows that the noticably worse performance on data from SRI and AT&T can be explained by the larger proportion of errors due to spontaneous speech effects. It also shows that errors due to spontaneous speech effects account for only about 22% of the total.

In order to calibrate our recent improvements, we retested on the October '91 dry-run test set. The current system gives a word error rate of 7.8% whereas our unofficial result in October was 14.4% word error. (Note that we did not use the ATIS2 speech to train the system for the October '91 dry-run test.) We attribute our improvement to several important factors:

1. More appropriate training material – ATIS2 multi-site spontaneous data instead of read ATIS0 data from TI and BBN,
2. A trigram language model – versus a bigram in October,
3. Automatic location of silences in the training data.

Note that the quantity of ATIS2 training data used (7700 utts) is only half the amount used to estimate the model used for the October '91 dry-run (about 13,500 utts). Clearly the quality of the training material is an important factor in performance.

## 2. HARC – SPOKEN LANGUAGE UNDERSTANDING

HARC, BBN's spoken language system, utilizes BYBLOS as its speech recognition component, and DELPHI as its natural language understanding component. DELPHI uses a definite clause grammar formalism, augmented by the use of constraint nodes [9] and a labelled argument formalism [3]. The parsing algorithm uses a statistically trained agenda to produce the single best parse for an input utterance [1].

We experimented with several conditions to optimize the connection of BYBLOS with DELPHI. The basic interface between speech and natural language in HARC is the N-best list. Previously, we had allowed the natural language component to search arbitrarily far down the N-best list until it either found a hypothesis that produced a database retrieval or reached the end of the N-best list. For this evaluation, we explored the nature of this connection in more detail. The parameters we varied were:

- the depth of the search that NL performed on the N-best output of speech
- the processing strategy used by NL on the speech output

In our earlier work in integrating speech and natural language, we had noticed that while it was beneficial for NL to look beyond the first hypothesis in an N-best list, the answers obtained by NL from speech output tended to degrade the further down in the N-best list they were obtained. During this last period, we performed a number of experiments to determine the break-even point for NL search. We used an N of 1, 5, 10, and 20 in our experiments.

During our recent development work, we utilized a number of fall-back strategies for NL text processing [2]. In applying these fall-back strategies to speech output, we examined the trade-off between processing speech output with a more restrictive scheme, and thereby potentially discarding meangingful utterances vs. processing speech output with a more forgiving strategy, and thereby potentially allowing in meaningless or misleading utterances. We experimented with three processing strategies:

- fallback processing turned off
- fallback processing turned on
- a combined strategy, in which an initial pass with made with fallback processing turned off. If no hypothesis produced a database retrieval, a second pass was made, with the fallback strategy engaged.

We show the results of one such experiment, utilizing the October '91 dry-run corpus as development test in Table 4. The results of our experiments indicated that an N of 5 was optimal, and that the two-pass processing strategy was slightly better than either of the others. This was the configuration we used on the February '92 evaluation data.

In Table 5 we show our Weighted Error on the February '92 evaluation data for Combined Class A+D, and Classes A and D separately, as calculated by NIST. During the test run, we had neglected to include the date information provided for individual scenarios. We include the results of a re-run with the same system as ran the February '92 test set, with

| Condition | N | WE |
|---|---|---|
| Text | (1) | 47.9 |
| Fallback on | 1 | 64.6 |
| ” | 5 | 58.0 |
| ” | 20 | 60.1 |
| Fallback off | 1 | 64.2 |
| ” | 5 | 56.9 |
| ” | 20 | 59.0 |
| Two Pass | 5 | 56.6 |

Table 4: SLS weighted error (WE) on the October '91 dry-run test set with varying N-best list length (N).

the only change being the inclusion of the date information. Interestingly, the lack of date information only affected 3 utterances, which were given False answers without the date, and True answers with it.

| Corpus | Official WE | WE with date |
|---|---|---|
| A+D | 43.7 | 42.8 |
| A | 35.8 | 34.8 |
| D | 54.7 | 54.0 |

Table 5: SLS weighted error (WE) on the February '92 test set.

## 4. SUMMARY

We have shown superior speech recognition performance with only a modest amount of training speech by aggressively handling the idiosyncrasies of this corpus. All utterances that are degraded due to severe disfluencies or problems with data-capture are eliminated from the training set. The excessively long and numerous segments of ambient noise in the data are removed from consideration by a good speech detector in the front-end. The very numerous hesitation phenomena are automatically located and then explicitly modeled where they occur in the training. Nonspeech events, such as filled-pauses, are made very unlikely in the grammar to clamp the false alarm rate.

In addition, the trigram language model on word classes significantly improved recognition performance compared to a bigram model.

With these improvements, the official BYBLOS speech recognition results for the February '92 DARPA evaluation were 6.2% word error for the Class A+D subset of the test and 9.4% overall. Both of these results were significantly better than any other speech system tested.

Finally, we have shown how the N-best interface between the speech and natural components reduces the error rate compared to considering the top choice only. This was shown to be true whether a robust fragment processor was used as a fall-back or not.

The official SLS result for HARC was a weighted error of 43.7. This was the best overall result for a spoken language system in the February '92 DARPA evaluation.

## 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. Bobrow, R. "Statistical Agenda Parsing", in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19–22, 1991*, Morgan Kaufmann Publishers, Inc., San Mateo, California, pp. 222–224.
2. Bobrow R., D. Stallard, "Fragment Processing in the DELPHI System", *elsewhere in these proceedings*.
3. Bobrow, R., R. Ingria, and D. Stallard "Syntactic/Semantic Coupling in the DELPHI System", *elsewhere in these proceedings*.
4. 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.
5. 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.
6. Kubala, F., S. Austin, C. Barry, J. Makhoul, P. Placeway, R. Schwartz, "BYBLOS Speech Recognition Benchmark Results", *Proceedings of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Inc., Feb. 1991.
7. MADCOW, "Multi-Site Data Collection for a Spoken Language Corpus", *elsewhere in these proceedings*.
8. 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.
9. Stallard, D. "Unification-Based Semantic Interpretation in the BBN Spoken Language System", in *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15–18, 1989*, Morgan Kaufmann, Publishers, Inc., San Mateo, California, pp. 39–46.