The BBN BYBLOS Continuous Speech Recognition System

Richard Schwartz
Chris Barry, Yen-Lu Chow, Alan Derr, Ming-Whei Feng,
Owen Kimball, Francis Kubala, John Makhoul, Jeffrey Vandegrift

BBN Systems and Technologies
10 Moulton Street
Cambridge, MA 02138

Abstract

In this paper we describe the algorithms used in the BBN BYBLOS Continuous Speech Recognition system. The BYBLOS system uses context-dependent hidden Markov models of phonemes to provide a robust model of phonetic coarticulation. We provide an update of the ongoing research aimed at improving the recognition accuracy. In the first experiment we confirm the large improvement in accuracy that can be derived by using spectral derivative parameters in the recognition. In particular, the word error rate is reduced by a factor of two. Currently the system achieves a word error rate of 2.9% when tested on the speaker-dependent part of the standard 1000-Word DARPA Resource Management Database using the Word-Pair grammar supplied with the database. When no grammar was used, the error rate is 15.3%. Finally, we present a method for smoothing the discrete densities on the states of the HMM, which is intended to alleviate the problem of insufficient training for detailed phonetic models.

1. Introduction

At BBN we have been involved in the development of Spoken Language Systems for almost two decades. As part of DARPA's Speech Understanding Research Program from 1971-1976, we developed a system that integrated continuous speech recognition with natural language understanding in a 1000-word travel management task; we call the system HWIM (Hear What I Mean). As part of another DARPA program, we have been working since 1982 on a more advanced speech recognition system based on using Hidden Markov Models. The result of this work is the BYBLOS Continuous Speech Recognition System.

The basic algorithms used in the BYBLOS Continuous Speech Recognition system have been described in several papers [1, 2, 3]. In Section 2 we give a brief review of the techniques currently used in the BYBLOS system. The two features that have made the largest improvements in recognition accuracy since 1982 were the use of robust context-dependent phonetic models, and the addition of derivative spectral parameters in multiple codebooks. Each of these features used separately reduces the recognition error rate by a factor of two. Taken together, they reduce the error rate by a factor of four. In Section 3 we present the latest recognition results for the BYBLOS system. In particular, we compare the recognition results with and without spectral derivative parameters. We also demonstrate, by testing the system on training data, that the recognition accuracy is likely to improve as more training data is made available. Since several similar systems have provided test results on this database it is possible to determine the benefits of particular algorithms. In particular, we compare the error rate for using discrete densities with that using continuous densities. We also compare the recognition accuracy for speaker-dependent models with that for speaker-independent models derived from a large number of speakers. Finally, in Section 4, we present a method for smoothing the discrete densities on the states of the HMM. The smoothing is intended to alleviate the problem of insufficient training for detailed phonetic models.

2. The BYBLOS system

The BYBLOS system uses context-dependent hidden Markov models (HMM) of phonemes to provide a robust model of coarticulation [1, 2]. Each phoneme is typically modeled as a HMM with three states that correspond roughly to the acoustics of the beginning, middle, and end of the phoneme. To model the acoustic coarticulation between phonemes, we define a separate HMM for each phoneme in each of its possible contexts. Since many of these phonetic contexts do not occur frequently enough to allow robust estimation of model parameters, we interpolate the detailed context-dependent phonetic models with models of the same phoneme that are dependent on less context. In this way we derive the benefit of word-based models for words with sufficient training and the generality of phoneme-based models for the rest. For example, we use triphone models that depend jointly on the preceding and following phonemes, we use diphone models that depend separately on the preceding or following context, and we use context-independent models that are pooled across all instances of the phoneme. We have also experimented with models of the phoneme that depend on the particular word that the phoneme is in [3]. We average the probabilities of the different context-dependent models with weights that depend on the state within the phoneme and on the number of occurrences of each type of context in the training set.

With each state of the HMM we associate a conditional probability density of the spectral features given that state. The basic spectral features are mel-scaled cepstral coefficients (MFCC) [4] and the log of the normalized total energy. We derive the MFCC by warping the log power spectrum of each frame of speech before computing the cepstrum (by inverse Fourier transform). A portion of the training set of MFCC vectors is clustered to produce a codebook of spectral prototypes [5]. We typically use a codebook with 256 prototypes. Then for each frame we find the index of the nearest vector quantizer (VQ) prototype. The discrete probability density is therefore represented as a vector of 256 numbers indicating the probability of each VQ index given the state.

The decoding algorithm used in the BYBLOS system has been described in [2]. The algorithm is a time-synchronous beam search for the most likely sequence of words, given the observed speech parameters. The algorithm is similar to the commonly used Viterbi algorithm with the exception that, when performing the state update within a word, the probability of being in a particular state is derived from the sum of the probabilities at each of the preceding states. This is contrasted with the standard Viterbi algorithm, in which we use the maximum over the preceding states. This algorithm more nearly computes the correct likelihood function for each sequence of words and was found to result in a small but consistent improvement over the standard Viterbi algorithm. As with the Viterbi algorithm, the search can be constrained by any finite-state grammar. It has also been used in a top-down search using context-free grammars.

Multiple Codebooks

As shown by Furui [6], even though the sequence of spectral parameter vectors may be sufficient to reproduce a reasonable facsimile of the original speech, it is beneficial to explicitly include the derivatives of the spectral parameters in the recognition algorithm. To avoid problems associated with trying to estimate probability densities of large dimensional spaces, we use a separate VQ codebook and probability distribution for the steady state and derivative parameter sets. We multiply the probabilities for the different parameter sets as if they were independent [7]. During the past year we modified the BYBLOS system to use multiple sets of features. Currently, the BYBLOS system uses three sets of spectral features: 14 mel-scale cepstral coefficients, the 14 derivatives of these parameters (computed as the derivative of a linear fit to 5 successive frames), and a third set containing the normalized total energy and the derivative of the energy.

3. Results

In this section we present the recognition results for the BYBLOS system under several different conditions. But first, we describe the database and the testing procedure used for all the results in this paper.

DARPA Resource Management Database

Most of the recent research with the system has been performed using the standard 1000-word DARPA Resource Management Database [8]. Tests were performed on the speaker-dependent portion of the database which contains the speech of 12 speakers. The training set for each speaker consists of 600 sentences averaging eight words or three seconds in length, for a total of about 30 minutes of speech. There are two test sets of 100 sentences each. The first test set is designated as "development test" and was distributed by the National Bureau of Standards for formal tests. 25 sentences from 8 of the 12 speakers were distributed in October, 1987; 25 different sentences from all 12 speakers were distributed in May, 1988. After these two formal tests, all 100 of the development test sentences were released for development purposes. The remaining 100 test sentences, which were designated as "evaluation test", were also divided into 25 sentence groups and are being distributed in a similar manner for further formal testing. The first set of 25 was distributed for February, 1989. In the remainder of this paper, we will refer to the different formal test sets by their date of distribution, (e.g. the Oct. '87 test set, etc.).

In addition to the speech data itself, the database also contains a specification of two grammars to be used for testing and testing procedures to be used to assure that results from different research sites can be compared. Recognition runs typically are performed using an artificial "Word-Pair" grammar with perplexity 60 that allows all pairs of word classes that appear in the database, and with no grammar or perplexity 1000. The recognized strings are automatically aligned to the true word strings, and the number of word substitutions, insertions, and deletions are computed. The standard single-number measure of performance is the total word error, which is defined as

$$\%error = 100 \; \frac{\text{substitutions} + \text{deletions} + \text{insertions}}{\text{total words}}$$

When sentence error rate is quoted, (typically only when using a grammar) it is defined as the percentage of sentences with any error at all.

Multiple Codebook Results

We compared the recognition accuracy when the system used 14 steady state cepstral coefficients with that when it used three codebooks (including derivative and energy parameters). The comparison was made on the Oct. '87 test set of 8 speakers under the two grammar conditions. Table 1 below contains the results of this comparison. As can be seen, the use of derivative (and energy) information reduced the error rate by about a factor of two under both grammar conditions.

	Word-Pair	No Grammar
1 codebook	7.5	32.4
3 codebooks	3.6	18.0

Table 1: Recognition error rate with 1 codebook with steady state parameters vs 3 codebooks with added derivative and energy parameters.

The results given above for 3 codebooks were development results, in that the têst set had been used several times. Therefore, we present below in Table 2 the results of testing the system on all 12 speakers on the May '88 and Feb '89 test sets for the first time, using the same phonetic word models as used above.

The average word error rates were 3.4% and 2.9% when the Word-Pair grammar was used, and 16.2% and 15.3% when no grammar was used. The difference between the error rates of the two test sessions is only marginally significant especially given the variation between speakers. Table 3 below shows the detailed results for the February '89 test sets for each of the 12 speakers. The table gives the percent substitution, deletion, and insertion errors, in addition to the total word and sentence error rates.

	Word-Pair	No Grammar
May '88	3.4	16.2
Feb '89	2.9	15.3

Table 2: Recognition Results With and Without Grammar for May '88 and Feb '89 Test Sets

	NO GRAMMAR				WORD PAIR				
	Sub	Del	Ins	Word Err	Sub	Del	Ins	Word Err	Sent Err
BEF	10.8	6.1	0.5	17.4	1.9	0.9	0.5	3.3	24.0
CMR	14.6	3.3	3.3	21.2	2.8	0.9	0.5	4.2	24.0
DAS	2.9	1.0	1.0	4.9	0.5	0.0	0.5	1.0	8.0
DMS	8.6	2.3	1.1	12.0	1.1	1.1	0.0	2.2	12.0
DTB	15.3	2.4	1.0	18.7	2.4	0.5	0.0	2.9	20.0
DTD	13.1	2.3	1.4	16.8	2.3	0.0	0.9	3.2	20.0
ERS	17.7	2.9	2.9	23.5	1.1	1.7	0.6	3.4	16.0
HXS	6.1	1.9	1.9	9.9	0.9	1.4	0.0	2.3	16.0
JWS	7.7	1.7	0.0	9.4	1.7	0.4	0.0	2.1	20.0
PGH	8.8	1.4	2.8	13.0	1.8	0.5	0.0	2.3	20.0
RKM	13.2	4.7	4.3	22.2	3.4	1.3	0.4	5.1	36.0
TAB	10.6	2.3	2.3	15.2	1.8	0.9	0.0	2.7	24.0
Avg	10.8	2.7	1.9	15.3	1.8	0.8	0.3	2.9	20.0

Table 3: Detailed Recognition Results for Each of the Twelve Speakers on the Feb '89 Test Set. Results are given with and without the Word-Pair grammar. For each condition and speaker, the table shows the percent substitution, deletion, and insertion errors, and total word error. Percent sentence error is also given for the Word-Pair grammar.

Test on Training

It is frequently instructive to measure the recognition performance of a system when it is tested on data that was included in the training set. In Table 4 below we compare the word and sentence recognition error rate when the system is tested on the training set versus when it is tested on an independent test set. The same acoustic models were used in both cases. Results are given for the Word-Pair grammar only.

As can be seen, when the system is tested on training data, the error rates are very small. This large difference in performance indicates that there is not enough training data for the number of free parameters we have in our phonetic models. Therefore, we might expect that recognition accuracy would improve considerably as we add more training data.

	Independent Test	Training Data
Word Error	2.9 - 3.4	0.5
Sentence Error	20.0	2.7

Table 4: Comparison of Results on Independent Test vs on Training Data. The Word-Pair grammar was used.

Comparison of Methods

Several other research groups have also reported their recognition results on this same database. Since, in many cases, the algorithms differ in only one or two aspects, it is possible to identify differences in performance with particular aspects of a system. In this section, we attempt to make two such comparisons: discrete vs continuous densities, and speaker-dependent vs speaker-independent models. The comparisons are made on the results provided for the May '88 test set because the different systems were most similar for this test set. We note that each of these systems has evolved since testing on this particular test set, and as a result their results have improved considerably as can be seen in the results presented for those systems elsewhere in this volume.

The continuous speech recognition system developed at MIT Lincoln Labs by Doug Paul uses Gaussian probability densities to represent each of the states of the HMM instead of the discrete densities used in BYBLOS. In most other respects, the two systems are quite similar. The recognition accuracy for the speaker-dependent test on the May '88 test set was 5.5%, as compared with 3.4% for the BYBLOS system. It would appear, then, that continuous HMM densities do not necessarily provide improved results over discrete densities.

Another comparison of interest is the relative performance of speaker-dependent models versus speaker-independent models. While it is clear that, for any given duration of training, a speaker-dependent model (trained for the particular speaker using the system) should always result in much higher recognition accuracy, the practical question remains, "How much more training does a speaker-independent system need to give the same accuracy as a speaker-dependent system?" Three systems that are almost identical to BYBLOS have been used on the speaker-independent portion of the Resource Management Database. Two different training sets have been used in the tests on the May '88 test set: one with 72 different speakers containing 2880 sentences, and a larger one with 105 speakers containing 4200 sentences. The test data used was the same as for the speaker-dependent test described above.

When trained on 72 speakers the word error rate with the Word-Pair grammar was 10.1% for the Sphinx system of Camegie Mellon University, 11.4% for the Decipher system of Stanford Research Institute, and 13.1% for the Lincoln Labs system. The Sphinx system and the Decipher system both use discrete densities similar to those used in BYBLOS. When trained on 105 speakers, the error rates for Sphinx and the Lincoln system were 8.9% and 10.1% respectively. Thus, the BYBLOS system with speaker-dependent training with five to seven times less training data has roughly 1/2 to 1/3 the error rate of the speaker-independent trained systems. It would be interesting to find out how much additional speech is needed for speaker-independent training to result in the same performance as 30-minute speaker-dependent training.

4. Robust Smoothing for Discrete Probability Densities

Much of the research in speech recognition is devoted to improving the structure of the statistical model of speech. Frequently, improving the model involves increasing the complexity or dimensionality of the model. For example, we use context-dependent phonetic models, which increases the number of models. We add features, such as spectral derivatives, which increases the dimensionality of the feature space. We use a non-parametric probability density function (pdf) to have flexibility in the model, but we lose the benefit of the compactness of a parametric model. Each of these improvements comes with an increase in the effective number of degrees of freedom in our model. Unfortunately, more training data is needed to estimate reliably the increased number of free parameters. Conversely, faced with a fixed amount of training data, we must limit the number of free parameters or else our "improvements" will not be realized.

As described above the BBN BYBLOS Continuous Speech Recognition system uses discrete nonparametric pdfs of context-dependent phonetic models. Most of these pdfs are trained with only a few tokens of speech (typically between 1 and 10). These discrete distributions work surprisingly well, given the small amount of training.

However, they are certainly prone to the problem of spectral types that do not appear in the training set for a given model, but are, in fact, likely to occur for that model. The results presented in Table 4 in Section 2 indicate that there is a large difference in recognition rate when the system is tested on the training data and on independent test data. Therefore, we tried to find a smoothing algorithm that would reduce the number of probabilities that are low purely due to a lack of training. Below we describe a general smoothing method based on using a probabilistic smoothing matrix [9].

For each state of a discrete HMM, we have a discrete probability density function (pdf) defined over a fixed set, N, of spectral templates. For example, in the BYBLOS system we typically use a vector quantization (VQ) codebook of size N=256 [5]. The index of the closest template is referred to below as the VQ index or the spectral bin. We can view the discrete pdf for each state s as a probability row vector

$$\mathbf{p}(s) = [p(k_1|s),\ p(k_2|s),\ \ldots,\ p(k_N|s)], \qquad (2)$$

where $p(k_i|s)$ is the probability of spectral template k_i at state s. We can imagine that the probabilities of different spectra are related in that, for each spectrum that has a high probability for a given pdf, there are several other spectra that are also likely to have high probabilities. These might be "nearby" spectra, or they might just be statistically related. We represent this relation by $p(k_j|k_i)$, the probability that if spectrum k_i occurs, the spectrum k_j will occur also. The set of probabilities $p(k_j|k_i)$ for all i and j form an $N \times N$ *smoothing matrix*, $\mathbf{T}$, where $\mathbf{T}_{ij} = p(k_j|k_i)$.

If we multiply the original pdf row vector $\mathbf{p}(s)$ by the smoothing matrix, we get a smoothed pdf row vector.

$$\mathbf{p}_{smooth}(s) = \mathbf{p}_{orig}(s) \times \mathbf{T}. \tag{3}$$

In our experiments we use a separate smoothing matrix for each phoneme. This matrix is combined with the phoneme-independent matrix to ensure robustness.

The amount of training available for different models varies considerably, from one or two tokens for the majority of the triphone-dependent models to hundreds of tokens for the more common models. Clearly, we don't want to smooth a model as much if it was estimated from a large number of training tokens. Therefore we *recombine* the smoothed pdf above with the original pdf using a weight $w(s)$ that depends on the number of training tokens of the model. Thus the final pdf used is given by

$$\mathbf{p}_{final}(s) = w(s)\mathbf{p}_{orig}(s) + [1-w(s)]\mathbf{p}_{smooth}(s). \tag{4}$$

The weight w is made proportional to the log of the number of training tokens, N_T.

$$w(s) = min[0.99,\ 0.5\ \log_{10}N_T(s)] \tag{5}$$

This equation is illustrated in Figure 1.

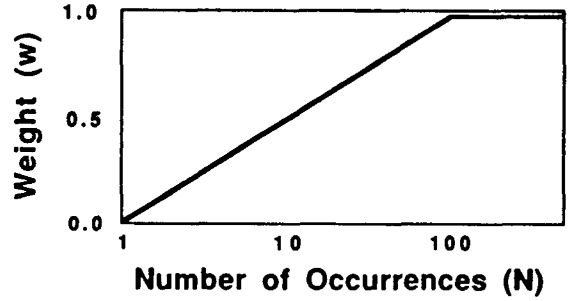

Figure 1: Weight w for original model as a function of the number of training tokens, N_T

Estimating the Matrix

We have tried three techniques for estimating the smoothing matrix: *Parzen smoothing*, *self adaptation cooccurrence smoothing*, and *triphone cooccurrence smoothing*. These methods were presented in a talk at Arden House in May 1988 and are described in detail in [10]. Since the third method worked best in our initial experiments, we will discuss only that method.

After performing forward-backward training, we have a large number of context-dependent phonetic models. Most of these (about 2,500) are triphone-dependent models. Each model has three different pdfs. These models contain a record of all of the VQ-index spectra that occurred for one part (one state) of a particular triphone. Thus, according to the Markov model, these spectra freely cooccur. For each pdf of each triphone model we count all permutations of two VQ spectra in that pdf, weighted by their probabilities and by the number of training tokens of the model. Figure 2 illustrates this process for one pdf of one model.

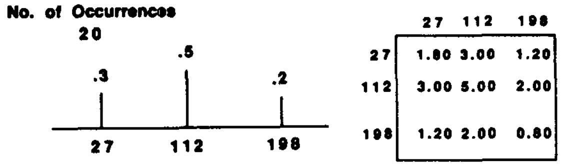

Figure 2: Triphone Cooccurrence Matrix Estimation. pdf shown results in matrix increments shown.

For example the pdf shown has VQ indices 27, 112, and 198 with probabilities 0.3, 0.5, 0.2 respectively. The model occurred 20 times in the training set. Therefore, we add 0.3 * 0.5 * 20 = 3.0 to entries (27,112) and (112,27) in the matrix. As with the second method, we keep a separate matrix for each phoneme and one phoneme-independent matrix. Each row is normalized to create probabilistic matrices. A method similar to this was developed independently by Lee [11]. However, in his method there was only one smoothing matrix, instead of one for each phoneme, and he estimated the matrix from context-independent models instead of triphone-dependent models. We believe that these differences result in too much smoothing.

Recognition experiments using the word-pair grammar were performed with and without triphone cooccurrence smoothing on all three test sets. These results are shown below in Table 5.

Total word error rate (%)

	Word-Pair		No Grammar	
Test Set	**Baseline**	**Smooth**	**Baseline**	**Smooth**
Oct. '87 (8 spkrs)	3.6	3.0	18.0	19.2
May '88 (12 spkrs)	3.4 *	2.7	16.2 *	15.2
Feb. '89 (12 spkrs)	2.9 *	3.1 *	15.3 *	13.8 *

*** Official test**

Table 5: Recognition Results With and Without Smoothing

5. Conclusions

We have described the BYBLOS Continuous Speech Recognition System. As expected, we found that adding the derivative and energy parameters in separate codebooks reduced the error rate by a factor of two, relative to using the steady state spectral parameters alone. The resulting word error rate was 3.4% and 2.9% on two successive formal tests. We presented an algorithm for smoothing discrete probability densities when the training is insufficient. However the algorithm provided only a small gain in recognition accuracy when 30 minutes were available for training. The HMM systems based on nonparametric discrete densities resulted in higher accuracy than the system that used continuous densities, leaving open the question of whether it is harmful to quantize the spectral parameters. The error rate of the speaker-dependent system when trained with 30 minutes of speech was less than half that of similar speaker-independent systems trained on over 100 speakers with five to seven times the amount of speech.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Office of Naval Research under contract number N00014-85-C-0279.

References

1. R.M. Schwartz,.Y. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Diego, CA, March 1984, pp. 35.6.1-35.6.4.

2. R.M. Schwartz, Y.L. Chow, O.A. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, March 1985, pp. 1205-1208, Paper No. 31.3.

3. Y.L. Chow, R.M. Schwartz, S. Roucos, O.A. Kimball, P.J. Price, G.F. Kubala, M.O. Dunham, M.A. Krasner, and J. Makhoul, "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, April 1986, pp. 1593-1596, Paper No. 30.9.1.

4. S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 4, August 1980, pp. 357-366.

5. J. Makhoul, S. Roucos, and H. Gish, "Vector Quantization in Speech Coding", *Proc. IEEE*, Vol. 73, No. 11, November 1985, pp. 1551-1588, Special Issue on Man-Machine Speech Communication.

6. S. Furui, "Speaker-Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, April 1986, pp. 1991-1994, Paper no. 37.10.

7. V.N. Gupta, M. Lennig, and P. Mermelstein, "Integration of Acoustic Information in a Large Vocabulary Word Recognizer", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Dallas, TX, 1987, pp. 697-700.

8. P. Price, W.M. Fisher, J. Bernstein and D.S. Pallett, "The DARPA 1000-Word Resource Management Database for Continous Speech Recognition", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, New York, NY, April 1988, pp. 651-654.

9. K. Sugawara, M. Nishimura, K. Toshioka, M. Okochi, and T. Kaneko, "Isolated Word Recognition Using Hidden Markov Models", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa FL, March 1985, pp. 1-4.

10. R. Schwartz, O. Kimball, F. Kubala, M. Feng, Y.L. Chow, C. Barry, J. Makhoul, "Robust Smoothing Methods for Discrete Hidden Markov Models", *IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, Scotland, May 1989.

11. K.F. Lee, *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The Sphinx System*, PhD dissertation, Carnagie-Mellon University, April 1988, CMU-CS-88-148