Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems

James Baker, Janet Baker, Paul Bamberg, Kathleen Bishop, Larry Gillick, Vera Helman, Zezhen Huang, Yoshiko Ito, Stephen Lowe, Barbara Peskin, Robert Roth, Francesco Scattone

Dragon Systems, Inc.
Newton, Massachusetts 02160

ABSTRACT

In this paper we present some of the algorithm improvements that have been made to Dragon's continuous speech recognition and training programs, improvements that have more than halved our error rate on the Resource Management task since the last SLS meeting in February 1991. We also report the "dry run" results that we have obtained on the 5000-word speaker-dependent Wall Street Journal recognition task, and outline our overall research strategy and plans for the future.

In our system, a set of output distributions, known as the set of PELs (phonetic elements), is associated with each phoneme. The HMM for a PIC (phoneme-in-context) is represented as a linear sequence of states, each having an output distribution chosen from the set of PELs for the given phoneme, and a (double exponential) duration distribution.

In this paper we report on two methods of acoustic modeling and training. The first method involves generating a set of (unimodal) PELs for a given speaker by clustering the hypothetical frames found in the spectral models for that speaker, and then constructing speaker-dependent PEL sequences to represent each PIC. The "spectral model" for a PIC is simply the expected value of the sequence of frames that would be generated by the PIC. The second method represents the probability distribution for each parameter in a PEL as a mixture of a fixed set of unimodal components, the mixing weights being estimated using the EM algorithm. In both models we assume that the parameters are statistically independent.

We report results obtained using each of these two methods (RePELing/Respelling and univariate "tied mixtures") on the 5000-word closed-vocabulary verbalized punctuation version of the Wall Street Journal task.

1. INTRODUCTION

This paper presents "dry run" results of work done at Dragon Systems on the Wall Street Journal (WSJ) benchmark task. After we give a brief description of our continuous speech recognition system, we describe the two different kinds of acoustic models that were used and explain how they were trained. Then we present and discuss the results obtained so far and review our plans for further research.

In our system a set of output distributions, known as the set of PELs (phonetic elements), is associated with each phoneme. The HMM for a PIC (phoneme-in-context) is represented as a linear sequence of states, each having an output distribution chosen from the set of PELs for the given phoneme, and a (double exponential) duration distribution. The model for a particular hypothesis is constructed by concatenating the necessary sequence of PICs, based on the specified pronunciation (sequence of phonemes) for each of the component words. Thus our system models both word-internal and cross-word co-articulation. When a model for a PIC that is needed does not exist, a "backoff" strategy is used, whereby the model for a different, but related, PIC is used instead.

The two methods to be compared in this paper constitute different strategies for representing and training the output distributions to be used for the nodes found in the PIC models. The first method involves generating a set of (unimodal) PELs for a given speaker by clustering the hypothetical frames found in the spectral models for that speaker, a step we call "rePELing", and then constructing speaker-dependent PEL sequences to represent each PIC as an HMM, which we call "respelling". The spectral model for a PIC can be thought of as the expected value of the sequence of frames that would be generated by the PIC, normalized to an average length. The second method, a univariate version of tied mixtures, represents the probability distribution for each parameter in a PEL as a mixture of a fixed set of unimodal components, the mixing weights being estimated using the EM algorithm [9]. In both the RePELing/Respelling and the tied mixture models, we assume that the parameters are statistically independent. A more detailed explanation of these two methods can be found in sections 3 and 4.

2. OVERVIEW OF DRAGON TRAINING AND RECOGNITION

The continuous speech recognition system developed by Dragon Systems was presented at the June 1990 DARPA

*This work was sponsored by the Defense Advanced Research Projects Agency and was monitored by the Space and Naval Warfare Systems Command under contract N00039-86-C-0307.

SLS meeting ([5], [6], [11]) and at the February 1991 DARPA SLS meeting ([4]). The version presented in this paper is speaker-dependent, and was demonstrated to be capable of near real-time performance on a 1000-word task when running on a 486-based PC. When running live, a TMS320C25-based board performs the signal processing and the speech is sampled at 12kHz. In the experiments reported in this paper, the speech was sampled at 16kHz, the speech waveforms having been supplied in a standard format by NIST.

An important contribution to our improved performance in the last year was our switch to 32 signal processing parameters (consisting of our eight original spectral parameters together with 12 cepstral parameters and their estimated time derivatives). The cepstral parameters were computed via an inverse Fourier transform of the log magnitude spectrum. At recognition time, the parameters are computed every 20 ms, while for purposes of training, 10 ms data was used.

The recognition algorithm relies on frame-synchronous dynamic programming (an implementation of the forward pass of the Baum-Welch algorithm) to extend sentence hypotheses subject to the elimination of poor paths by beam pruning. In addition, the Continuous Speech Recognizer uses the DARPA-mandated digram language model ([15]), which is a modification of the backoff algorithm from [13]. The rapid matcher, as described in [11], is another important component of the system. For any frame, it limits the number of word candidates that can be hypothesized as starting at that frame. For purposes of this paper, which is primarily concerned with the quality of our modeling, most of the rapid match errors have been eliminated by passing through long lists of words for the detailed match to consider, at the cost of considerable additional computation. Similarly, most of the pruning errors have been eliminated by running with a high threshold. A companion paper [10], that appears in this volume, describes a new strategy for training the rapid match models directly from the Hidden Markov Models specified by the PICs. This new strategy shows promise for reducing the average length of the rapid match list that must be returned at any given time, and thus, speeding up the recognizer.

In the experiments described below, models were trained for each of the 12 speaker-dependent Wall Street Journal speakers, using the approximately 600 training sentences (300 with verbalized punctuation and 300 without). Testing was done using the approximately 40 recorded sentences (per speaker) available as the 5000-word closed-vocabulary verbalized punctuation development test set.

In order to incorporate context information at the phoneme level, triphone structures were constructed that include information about the immediate phonetic environment that affects a phoneme's acoustic character. These augmented triphones, called "PIC"s, are the fundamental unit of the system, and are closely related to other approaches that have appeared in the literature ([16] and [14]). The information that the PICs currently contain is the identity of the preceding and succeeding phonemes, and, optionally, an estimate of the degree of the phoneme's prepausal lengthening. Each PIC is represented acoustically by a sequence of nodes. Each node is taken to have an output distribution specified by a PEL, and a duration distribution. PIC models representing the same phoneme may share PELs, but PELs can never be shared across phonemes. The parametric family used for modeling the probability distributions of the durations as well as of the individual acoustic parameters is assumed to have the double exponential form

$$P(x) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}},$$

where μ is the mean and σ is the mean absolute deviation.

A detailed description of the original models for PICs and how they were formerly trained can be found in [6]. The following sections explain how a variety of modifications have been made to the original PIC training algorithm.

The English phoneme alphabet used by the system includes 26 consonants (including the syllabic consonants, /L/, /M/, and /N/) and three levels of stress for each of 17 vowels, constituting a total of 77 phonemes. Approximately 10% of the lexical entries for the 5000-word WSJ task have multiple pronunciations, because of stress differences in the vowels and expected pronunciation variations.

Of course, the number of possible PICs that can appear in hypotheses at recognition time (including cross-word PICs) is vast compared to the number of PICs that typically appear in 600 sentences of Wall Street Journal training data. This paper reports results when around 35,000 PICs are built for the rePEL/respell models and when around 14,000 PICs are built for the tied mixture models. When the recognizer asks for a model for a PIC that has not been built, a backoff strategy is invoked which supplies a model for a related PIC instead.

3. REPELING/RESPELLING

In earlier reports [6], [7], we described a straightforward procedure that generated speaker-dependent models via several passes of adaptation of the reference speaker's models. The adaptation process modified the PEL probability distributions and the PIC-dependent duration distributions. However, no new PELs were created, nor was the PEL sequence for a given PIC allowed to change. The sharing of PELs by different PICs was determined by the acoustics of the reference speaker's speech, and was assumed to generalize to other speakers.

At the last SLS meeting in Feb 1991 [4], we reported on a method for choosing the sequence of PELs for a PIC in a speaker-dependent fashion, essentially in the same manner as had been done for the reference speaker. This step could be performed once the original PELs had been adapted using the reference speaker's PIC spellings. To the extent that differences in PEL sequences for a given PIC can reflect different choices of allophones, this extra step can capture allophonic variation among different speakers, and lifts the restriction that the sharing of PELs be the same for all speakers. This change produced a significant improvement in performance.

In order to take full advantage of our new more informative signal processing parameters, however, a further change was required. We needed to construct a new set of PELs to serve as the class of output distributions for the HMMs to be constructed. It was not adequate to simply extend, by adaptation, the 8 parameter PELs we had been working with, to 32 parameter PELs, as this would prevent us from making distinctions that could not even be seen with the old signal processing.

In the previous reports [6] and [4], we described how a set of PELs for the reference speaker was initially hand-constructed while running an interactive program for "labeling" spectrograms of the reference speaker's speech. We needed to be able to construct a new set of PELs automatically; thus, we implemented a k-means clustering algorithm whose purpose was to create a new set of (32 parameter) PELs for each speaker whose models were to be trained. This step involved clustering the frames in the "spectral models" for all of the PICs to be constructed for that phoneme. A spectral model for a PIC is obtained by performing linear stretching and shrinking operations on PIC tokens (examples of the given PIC and of related PICs, available from a prior segmentation of the training data, based on the best models then available) and then averaging the resulting transformed tokens (which have a common length), to obtain a kind of "expected" PIC token.

The primary motivation behind the rePELing step was to make it likely that each spectral frame would have at least one PEL that matched it fairly well. As each of the 77 phonemes was limited to having only 63 PELs available for building PICs, about 4500 PELs were created per speaker.

Once the new set of PELs had been created, a dynamic programming algorithm was used for converting the spectral model to an HMM containing up to six nodes, with each node assigned a PEL and a duration distribution. This respelling step drew on about 4000 of the 4500 PELs in constructing the HMMs.

A summary of the overall training procedure is outlined below, with rePELing and respelling appearing as steps 4 and 5:

1. Six passes of adaptation were run on each speaker's training data, starting with the reference speaker's models, using the old 8 parameter signal processing.
2. Segmentation of each speaker's data was performed, using the best available models (originally, those produced in step 1).
3. Spectral models were built for each PIC, using all 32 parameters, based on the segmentation in step 2.
4. RePELing was done for each speaker in order to generate a speaker-dependent set of output distributions.
5. For each speaker, respelling was performed to determine the PEL sequences that would be used in the resulting HMMs.
6. For each speaker, one additional pass of adaptation was performed in order to better estimate the mean absolute deviations for each parameter for each PEL.
7. Steps 2 - 6 could then be repeated, if desired.

Results for this method appear in section 5.

4. TIED MIXTURES

Were the model described in section 3 correct, the 32 parameters in each acoustic frame corresponding to a given PEL would be distributed as if they were generated by 32 independent (unimodal) double exponential distributions. However, graphical displays reveal that the frame distributions for many PELs have multiple modes. Furthermore, it is well known that the parameters within a frame are correlated. In order to deal with the multimodality of the data and to capture the dependence among parameters, Dragon has implemented

a modeling strategy in which the output distributions are represented in a more flexible way. This representation, similar to other tied mixture models developed elsewhere ([8], [12]), also provides the basis for achieving speaker independence.

If we divide the parameters into groups or "streams", with the property that parameters in different streams can be assumed to be independent, then our new modeling strategy represents the probability of a frame in a given state as the product of probability densities for each stream, and the probability density for a stream is assumed to be a mixture distribution over a fixed set of basis distributions specific to the stream.

More formally, we let $f(x)$ represent the probability density of a PEL, where x is a frame, and we assume that $f(x)$ is the product of s probability densities, $f_i(x_i)$, one for each stream:

$$f(x) = \prod_{i=1}^{s} f_i(x_i).$$

Furthermore, we assume that each f_i can be represented in terms of a set of basis distributions g_{ij}:

$$f_i = \sum_{j=1}^{C_i} \lambda_{ij} g_{ij},$$

where C_i is the number of components for stream i.

At the present time, we are using 32 streams; i.e., each parameter is assumed to be statistically independent of every other parameter in a given state. We have assumed the 32 parameters to be independent both as a way of relating our new results to our old results (which were also based on the same strong independence assumption), and as a debugging tool. We chose our basis distributions to be equally spaced double exponential distributions with a fixed mean absolute deviation, arranged so as to cover the full range of each parameter. Thus, when a mixture distribution was estimated, it was easy to see what values in the space were relatively likely or unlikely. In the system reported here, the set of basis components is the same for each stream, which would not be the case in a more general setting.

The tied mixture PIC models were assumed to be either 1-node or 2-node models, with the number of nodes being determined based on the proportion of very short PIC tokens. At the present time, no PEL is used as an output distribution for more than one node. Each tied mixture PIC model was built via the EM algorithm from instances of the given PIC found in the training data for the given speaker (based on segmentations obtained using the best available models). Unfortunately, most of the PICs that occur in the training data occur very few times, and, not surprisingly, most of the PICs that could in principle occur never in fact do.

Thus, two key problems that must be solved in training the recognizer are (1) the smoothing problem and (2) the backoff problem. The maximum likelihood estimator (MLE), together with many related asymptotically efficient estimators, has the defect of being a rather poor estimator when it is given only a small amount of data to work with: think of estimating the probability of "heads" from only one coin flip. Thus, it is important to smooth the MLE when there is clearly an insufficient supply of data. We have chosen to implement a smoothing algorithm with a strong Bayesian flavor. In this paper we will not address the backoff problem in any detail; at the present time, when we do not have a model for a PIC available to the recognizer, we substitute a "generic" PIC model, which has less specific context information.

The Bayesian solution to the coin flip problem amounts to representing the prior information we may have about the probability of "heads" as a prior number of flips, of which a certain number are taken to be heads, and then combining those "prior" flips with the real flips. We have taken a similar approach to the problem of estimating the mixing probabilities in our tied mixture models. We build the more common PICs before we build the less common PICs (see below). At the time that we are ready to build a given PIC, we make our best judgement as to what the mixing probabilities are for each stream of each state in the PIC. This guess is based on the models that have already been built for related PICs. Not only do we guess the mixing probabilities, but we also make a judgement about the "relevance" of our estimate, which is to say, the number of frames of real data that we judge our guess to be worth. We then use these prior estimates to initialize the EM algorithm, and in addition, we combine the accumulated fractional counts for each mixture component with the prior counts based on our prior guess, in forming the estimate to be used during the next iteration. Thus we have as our re-estimation formula:

$$\hat{\lambda}_{ij} = \frac{k\lambda^*_{ij} + n_{ij}}{\sum(k\lambda^*_{il} + n_{il})},$$

where λ^*_{ij} is the *a priori* estimate based on the PICs that have already been built, k is the relevance of this estimate, and n_{ij} is the accumulated fractional count for the jth component when estimating the distribution for

the ith parameter in a given node.

PICs are currently built in a prescribed order in our system: we build those for which there is the most data first. Thus, we begin by building the doubly-sided generic PICs, i.e. models for phonemes averaged over all left and right contexts. Then we move on to build singly-sided generic PICs, i.e. models for phonemes where the context is specified only on the right or on the left; we use the doubly generic PIC models to smooth the models for the singly generic ones. Finally we build our fully contextual PICs, but again we build the most common ones first, using the doubly and singly generic PICs to smooth the fully contextual ones. When building a relatively uncommon fully contextual PIC, it is useful to smooth the model using models of related fully contextual PICs which share some of the context or have closely related contexts.

5. RESULTS ON WSJ DATA

This section contains results on the 5000-word closed-vocabulary speaker-dependent verbalized punctuation version of the Wall Street Journal task, using the development test data. Table 1 lists results for all the WSJ speakers, displaying the word error rates using three different models. The first column contains the results of the first recognition run we did using models obtained by merely adapting our reference speaker's original models, using our old 8 parameter signal processing, yielding an overall word error rate of 35.6%. The second column contains our best 32 parameter unimodal models using the rePELing/respelling training strategy, after several iterations of training, with an overall error rate of 16.4%. Finally the last column contains the results of our first experiment recognizing Wall Street Journal sentences with the 32 stream tied mixture models described above, but based on only one segmentation step (segmentation into phonemes). This produced a word error rate of 14.8%. It is encouraging that the tied mixture models yielded better performance than did the unimodal models on 11 out of the 12 speakers, given that there has not yet been any opportunity for parameter optimization.

Speaker	Adapt only	RePEL Respell	Tied Mixtures
001	24.8	8.0	6.7
002	21.9	9.6	6.7
00A	64.5	24.6	26.8
00B	36.8	22.3	21.8
00C	47.1	28.6	28.1
00D	56.5	23.0	20.5
00F	43.3	20.6	16.9
203	27.1	14.4	13.9
400	27.6	12.5	12.4
430	30.5	13.5	9.6
431	29.4	14.5	9.8
432	18.3	5.5	4.7
AVG	35.6	16.4	14.8

Table 1: Summary of Wall Street Journal Results. 5000-word speaker-dependent closed-vocabulary development test set word error rate (%) using verbalized punctuation.

6. CONCLUSIONS

The training paradigm outlined above in the description of our tied mixture modeling has only recently been fully implemented at Dragon. Many aspects of the training strategy await full exploration, but the early results we have described are very encouraging. Already we have improved our performance relative to our old modeling and training paradigms.

In the coming months we plan to focus on a number of different aspects of training. First, we will be constructing basis distributions for streams with more than one parameter and studying the effect of this modeling on performance. We anticipate that we should obtain improved performance as we will then be modeling the dependence among parameters in an individual frame. We will also be studying a variety of backoff strategies, which involve substituting fully contextual PICs instead of generic PICs, when a PIC model has not been built. Another issue of importance will be the nature of our Bayesian smoothing, which we hope to implement in a more "data driven" way. Furthermore, we expect that the use of tied mixture modeling will allow us to develop a high-performance speaker-independent recognizer, an important goal for the coming year.

REFERENCES

1. X. L. Aubert, "Fast Look-Ahead Pruning Strategies in Continuous Speech Recognition", *Proc. ICASSP*, May 1989, Glasgow.
2. L. Bahl et al., "Large Vocabulary Natural Language Continuous Speech Recognition", *Proc. ICASSP*, May 1989, Glasgow.
3. L. Bahl, P.S. Gopalakrishnan and D. Kanevsky, "Matrix Fast Match: A Fast Method for Identifying a Short List of Candidate Words for Decoding", *Proc. ICASSP*, May 1989, Glasgow.
4. J.K. Baker, J.M. Baker, P. Bamberg, L. Gillick, L. Lamel, F. Scattone, R. Roth, D. Sturtevant, O. Ba and R. Benedict, "Dragon Systems Resource Management Benchmark Results - February 1991", *Proceedings of the DARPA Speech and Natural Language Workshop*, February 1991, Pacific Grove, California.

5. P. Bamberg, Y.L. Chow, L. Gillick, R. Roth and D. Sturtevant, "The Dragon Continuous Speech Recognition System: A Real-Time Implementation", *Proceedings of the DARPA Speech and Natural Language Workshop*, June 1990, Hidden Valley, Pennsylvania.

6. P. Bamberg and L. Gillick, "Phoneme-in-Context Modeling for Dragon's Continuous Speech Recognizer", *Proceedings of the DARPA Speech and Natural Language Workshop*, June 1990, Hidden Valley, Pennsylvania.

7. P. Bamberg and M. Mandel, "Adaptable phoneme-based models for large-vocabulary speech recognition",*Speech Comm.*, 10 (1991) 437-451.

8. J.R. Bellegarda and D.H. Nahamoo, "Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated-Speech Recognition", *Proc. ICASSP*, May 1989, Glasgow.

9. A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, 39(B), pp.1-38, 1977.

10. L. Gillick, B. Peskin and R. Roth, "Rapid Match Training for Large Vocabularies", *Proceedings of the DARPA Speech and Natural Language Workshop*, February 1992, Harriman, New York.

11. L. Gillick and R. Roth, "A Rapid Match Algorithm for Continuous Speech Recognition", *Proceedings of the DARPA Speech and Natural Language Workshop*, June 1990, Hidden Valley, Pennsylvania.

12. X. D. Huang and M. A. Jack, "Semi-continuous Hidden Markov Models for Speech Recognition",*Computer Speech and Language*, Vol. 3, 1989.

13. S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", *ASSP-35*, pp.400-401, March 1987.

14. K.F. Lee et al., "The SPHINX Speech Recognition System", *Proc. ICASSP*,May 1989, Glasgow.

15. D.B. Paul, "New Results with the Lincoln Tied-Mixture HMM CSR System", *Proceedings of the DARPA Speech and Natural Language Workshop*, February 1991, Pacific Grove, California.

16. R. Schwartz et al., "A Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", *Proc. ICASSP*, April 1985.