PROGRESS IN RECOGNIZING CONVERSATIONAL TELEPHONE SPEECH

Barbara Peskin, Larry Gillick, Natalie Liberman,
Mike Newman, Paul van Mulbregt, and Steven Wegmann

Dragon Systems, Inc.
320 Nevada Street, Newton, MA 02160 - USA

ABSTRACT

This paper describes recent improvements made to Dragon's speech recognition system which have improved performance on Switchboard recognition by roughly 10 percentage points in the past year. These features include the use of rapid speaker adaptation, a move from a 20 to a 10 msec frame rate for recognition, expansion of the acoustic training set and lexicon, and the introduction of interpolated language models. Preliminary results applying this Switchboard-trained system to conversations drawn from the English CallHome corpus are also quite strong, suggesting that this technology ports well to novel tasks. Finally, the paper includes a report on several research projects currently in progress which show promise of further reducing the error rate.

1. INTRODUCTION

When the Switchboard corpus [1] was introduced five years ago, the problem of recognizing conversational telephone speech seemed forbiddingly difficult. Early attempts at automatic transcription had word error rates in the 80%'s and 90%'s [2]. But far from being the intractable problem originally believed, recognition of Switchboard data has yielded in recent years to a number of techniques which continue to improve accuracy at a rate that gives no evidence of declining. Last year, we described a series of improvements to our Switchboard recognizer which resulted in word error rates in the 40%'s and near-perfect topic classification [3]. Since then, we have been able to cut the error rate even further, dropping from the 43.9% reported in [3] down to 34.3%, as measured on the so-called "CAIP set", a test of 20 pre-chopped conversation halves.

Dragon's continuous speech recognizer has been described extensively elsewhere (e.g. see [4]). The baseline system for this report uses gender-independent speaker-normalized acoustic models and a 10k trigram language model, all trained from Switchboard data. In this paper, we focus on changes made since [3] that have resulted in the nearly 10 percentage point improvement over this baseline system. Each of the next four sections highlights a new feature of the system.

Nor are these improvements specific to the Switchboard corpus. Even though the system was trained entirely on Switchboard data, we have demonstrated strong performance on a "blind" test of English conversations from the multilingual CallHome corpus of international telephone calls. In section 6, we discuss testing of the improved system, including this cross-corpus CallHome test. Finally, in section 7 we describe some new developments currently underway.

2. 10 MSEC RECOGNITION RATE

In dealing with such corpora as Wall Street Journal, Dragon has routinely trained on 10 msec data but dropped every other frame at recognition time to speed the decoding process. It had been our experience with read-speech databases that this practice resulted in relatively little loss in recognition performance and a significant savings in recognition time. However, conversational speech has much greater variation in speaking rate and we have found a correspondingly greater benefit in moving to the 10 msec frame rate.

Concurrent with our move to a 10 msec frame rate, we also introduced a new rapid match module that uses a phonetic tree for fast lexical search. After considerable parameter tuning for the new rate and new rapid match, we were able to improve Switchboard performance by a full 4 percentage points, as illustrated in the table below (Figures give word error rate on the CAIP set with bigram and trigram language models.):

recognition mode	10k bi	10k tri
20 ms, old rapid match	45.4%	43.9%
20 ms, new rapid match	—	43.4%
10 ms, new rapid match	41.4%	39.9%

Table 1. Effect of recognition rate and rapid match redesign.

Note that the new rapid match alone reduces the error rate only slightly given the open thresholds at which these tests were run, but it does speed up the recognition enough to make the cost of doubling the

number of speech frames tolerable.

3. EXPANDED ACOUSTIC TRAINING

The baseline acoustic models were trained from 60 hours of Switchboard speech, drawn from roughly 1300 message halves. We next decided to vastly expand our acoustic training to include over 3000 conversation sides. After extracting the speech segments via an automatic acoustic chopper, this resulted in nearly 170 hours of speech data.

Our triphone models are general mixtures of gaussians, where sharing of models among contexts is decided via binary decision trees. We trained new models using the same thresholds as before for our tree-building process, which resulted in an increase from 14,000 to about 21,000 output distributions. With the added data, we also felt that the mixture models might profitably use more components, so we allowed the number of gaussians per mixture to grow from 16 to 20. The additional data resulted in a 2 percentage point improvement in performance:

60 hr train, 16 components	41.4%
170 hr train, 16 components	39.9%
170 hr train, 20 components	39.4%

Table 2. Effect of training set size (using 10k bigram language model).

These are the first models we have built from the enlarged training set, and we expect to see further improvements as we iterate on the builds and tune our decision-tree growing to the greater quantity of data.

4. RAPID ADAPTATION

Dragon's approach to rapid adaptation uses linear regression to compute a transformation of acoustic space mapping speaker-independent models to speaker-dependent ones. This work was inspired by, and represents a simplification of, the techniques described in [5]. During a "Baum-Welch style" pass over the adaptation data, speaker-specific statistics on model means are collected and then regression is used to fit the transformation that best predicts the observed speaker means from the speaker-independent ones. Phonemes can be grouped into classes and separate transformations trained for each group.

We experimented with several different transformations to determine how complex a family could be supported given only the two minutes or so of speech in a typical Switchboard message half. We found that two transformations (essentially, one for vowels and one for consonants) yielded the best performance. The results are summarized in Table 3, which gives word error rates on the CAIP set using a 10k bigram language model and an assortment of recognition modes.

	unadpt	1 tr	2 tr	4 tr	7 tr
60 hr, 20 ms	45.4	42.6	42.1	—	43.0
60 hr, 10 ms	41.4	—	38.8	—	—
170 hr, 10 ms	39.4	—	36.8	37.1	37.1

Table 3. Effect of rapid adaptation with varying numbers of transformation groups.

Using two transformations, we obtained a (relative) 6-7% reduction in word error rate by using speaker adaptation, and this was on top of the relative 6% improvement already reported from using speaker-normalized base models [6]. (For more information on Dragon's implementation of speaker adaptation, see the companion paper [7].) In addition to the benefit of applying speaker adaptation at test time, we are now seeing improvements from its use in training, as described in section 7.1 below.

5. EXPANDED VOCABULARY AND INTERPOLATED LANGUAGE MODELS

In addition to expanding the acoustic training set, we decided to expand the vocabulary. We had always trained the language model and lexicon from the full 3000+ message halves now used in acoustic training, but we next decided to include all the words that occur in the training set rather than retaining only those that occur at least four times. This led to an increase in vocabulary size from 10k to 22k (represented by about 25k pronunciations).

Using this expanded vocabulary, we built several language models and explored various interpolated forms. The three component models were (1) a basic trigram language model trained from the full training set (1.9 million words of text), (2) for each of the 66 topics represented in the training set, a bigram language model for that topic (training varied with topic: from under 5,000 words of text for some topics to nearly 60,000 for others), and (3) a general trigram language model trained from the ARPA '94 Wall Street Journal training data (227 million words, but only n-grams involving the Switchboard vocabulary were used).

The interpolated models were created using linear combinations of the component models, according to the standard formula

$$P(w) = \lambda * P_1(w) + (1 - \lambda) * P_2(w),$$

where P is the probability of a word w as predicted by the interpolated model, and P_1 and P_2 are the probabilities predicted by the component models. Interpolation coefficients were estimated based on optimization of perplexity. In the case of topic-based language models, the topic of each message was determined using the same topic classification protocol described in [3].

With this 66-way discrimination, all test messages were correctly classified.

To test the new lexicon and language models, we supplemented the usual CAIP set with an additional test set of 15 messages on 15 different topics. This second set used full (unchopped) message streams and included a number of less well represented topics, thus making it a significantly harder task. With the expanded lexicon, out-of-vocabulary rates were halved: the OOV rate dropped from 1.4% to a mere 0.7% on the CAIP set, and from 2.3% to 1.3% on the 15-topic set. The resulting improvement in recognition is tabulated below. Tests used the unadapted 60-hour acoustic models and the interpolation constants specified in the table.

language model	CAIP	15-topic
10k bigram	41.4	43.4
10k trigram	39.9	42.2
22k trigram	39.2	41.5
.60(tri) + .40(topic)	—	40.6
.85(tri) + .15(WSJ)	38.3	40.4
.60(tri) + .25(top) + .15(WSJ)	—	40.1

Table 4. Word error rates for language model variations.

6. TESTING THE COMPLETE SYSTEM

The four new components described above were incorporated into a single system which made use of a multipass recognition protocol. First speaker-independent models and a simple bigram language model were used to transcribe the data. Then the (errorful) output of this pass was used for rapid adaptation to the test speaker. While the acoustic models were being adapted, the recognition transcripts were fed into a 66-way topic classifier to determine the conversation topic. The speech was then re-recognized using the speaker-adapted acoustic models and the triply-interpolated language model.

The system performed well on a variety of test sets. In addition to the CAIP set and the 15-topic set described above, we also tested it on an independent set of 20 (unchopped) Switchboard messages covering a new set of 20 topics and on 20 messages drawn from the English subset of the CallHome corpus.

CAIP set	34.3%
15-topic set	36.1%
20-topic set	39.1%
CallHome	50.3%

Table 5. Word error rates for the combined system.

Unlike the Switchboard messages, which are domestic calls between strangers on prompted topics, the CallHome messages are spontaneous conversations between friends and family members over international phone lines. Given the differences in the corpora, we were pleased to see that performance on CallHome English lagged only about 10 points behind that on the new 20-message Switchboard test, especially since the system was trained entirely from Switchboard data. For example, the OOV rate of 3.9% for the CallHome test is triple that of the Switchboard 15- and 20-topic tests. We were particularly surprised to find that topic interpolation still worked to our advantage despite the fact that the CallHome messages were not prompted with the Switchboard topics: using only the basic trigram + WSJ interpolation without the topic component resulted in the marginally higher word error rate of 50.9% for the CallHome set.

7. ON-GOING RESEARCH

In addition to the improvements described above, we are currently engaged in several new projects which show promise of reducing the error rate still further. We briefly describe a few of these here.

7.1. Adaptation at Training

We are experimenting with applying a transformation during training which functions as a sort of "inverse" to the rapid adaptation transform described in section 4. In doing so, we hope to map each speaker's training data to a "canonical" speaker-independent model, thus allowing us to build more focused models which will transform better under rapid adaptation. Other approaches to this problem have been taken, for example, by BBN [8].

Because the regression used in training the original transformations is determined from model means, we were concerned that applying the inverse directly to the speech frames might result in erratic behavior due to the greater variability in the frames. (This was supported by some early experiments with inverse mappings.) Consequently, we devised the following "inverse" transformation: For each frame of training data, determine its associated output distribution via forced alignment and see how the corresponding model mean would map under the usual speaker-independent to speaker-specific transform. Now instead of applying this shift to the model mean, subtract the shift from the speech frame itself, i.e.

$$f' = f - (x' - x)$$

where f and x are the original frame and associated model mean, x' is the speaker-adapted mean, and f' is the backwards-transformed frame. (Actually, the situation is somewhat more complicated: The model means x and x' are themselves weighted averages over the individual components of the mixture model.)

After applying such a transformation to the 160 training speakers in our 60-hour training set, we rebuilt the acoustic models and found a modest reduction in error rate when we applied our usual rapid adaptation at test time. But the real improvement comes from the fact that the new models can profitably use a greater number of transformations at test time. Overall we measured a 1.5% absolute reduction in error rate attributable to the improved models, beyond the improvement already obtained from the use of rapid adaptation at test time alone. Results for combinations of various numbers of transformations at training and testing are presented in Table 6. (All runs used the 10k bigram language model, but recognizer thresholds were slightly different from those used in Table 3, hence the small discrepancy compared to the "60 hr, 10 ms" line there.)

# transfs for train	# transfs for test				
	0	2	4	7	16
0	41.6	39.0	—	—	—
1	—	38.6	—	37.8	—
7	—	38.4	38.4	37.5	38.8

Table 6. Effect of rapid adaptation with varying numbers of transformations at test and train.

Furthermore, we found that by iterating the rapid adaptation at test time, we could obtain an additional 0.5% improvement, producing an error rate of 37.0% using 7 transformations at training and 7 for each of two passes at test time. While small, this improvement was remarkably uniform across test speakers, significant at the $P = 0.01$ level. Taken altogether, these improvements yield a 4.6% reduction in error rate over the system using no rapid adaptation, nearly doubling the improvement reported in section 4.

7.2. Discriminative Training

We are engaged in several projects designed to use discriminative training to improve recognition. For one project, we ran phoneme recognition on the 60-hour training set in order to collect statistics on misrecognized frames. By aligning the speech data to the resulting transcripts, we can compute model means for frames incorrectly assigned to a given output distribution. Statistics from these incorrectly recognized frames are then combined with (essentially subtracted from, with an appropriate weighting) the accumulated means of the true data, obtained from a Baum-Welch style pass using the correct training transcripts. We have experimented with a number of schemes for combining the two sets of statistics but so far have obtained only a 0.7% (absolute) reduction in the error rate. However, this work is still at a very early stage.

A second project uses discriminative techniques to control mixture splitting when creating the basis components used by our mixture models. So far we have been tuning this algorithm on simpler (e.g. monophone) models, where we have obtained reductions of 2% in word error rate. Performance results on more detailed models are not yet available.

7.3. Nonparametric Density Estimation

We have begun a highly speculative project aimed at exploring the use of nonparametric acoustic modelling. This work is motivated by exploratory experiments indicating that the acoustic data corresponding to a single output distribution actually occupies only about a five or six dimensional submanifold in the much higher dimensional space in which we perform acoustic modelling. (Dragon's recognizer now uses a 24 to 44-dimensional feature vector.) It is not surprising that our current acoustic models, using mixtures of 16 - 20 gaussian components, are unable to capture much of the fine structure of this complicated 5-dimensional subspace. We are therefore exploring nonparametric techniques, estimating the probability density at a point in acoustic space directly from the training data that occurs near that point. Performance results are not yet available for this work.

REFERENCES

[1] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. ICASSP-92*, San Francisco, March 1992.

[2] L. Gillick, et al., "Applications of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification using Telephone Speech," *Proc. ICASSP-93*, Minneapolis, April 1993.

[3] B. Peskin, et al., "Improvements in Switchboard Recognition and Topic Identification," *Proc. ICASSP-96*, Atlanta, May 1996.

[4] R. Roth, et al., "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer," *Proc. Spoken Language Systems Technology Workshop*, Austin, January 1995.

[5] C.J. Leggetter and P.C. Woodland, "Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression," *Proc. ICSLP'94*, Yokohama, September 1994.

[6] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker Normalization on Conversational Telephone Speech," *Proc. ICASSP-96*, Atlanta, May 1996.

[7] V. Nagesha and L. Gillick, "Studies in Transformation Based Adaptation," *Proc. ICASSP-97*, Munich, April 1997.

[8] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," *Proc. ICSLP'96*, Philadelphia, October 1996.