IMPROVEMENTS IN RECOGNITION OF CONVERSATIONAL TELEPHONE SPEECH

Barbara Peskin, Michael Newman, Don McAllaster,
Venkatesh Nagesha, Hywel Richards, Steven Wegmann, Melvyn Hunt, Larry Gillick

Dragon Systems, Inc.
320 Nevada Street, Newton MA 02460 - USA

ABSTRACT

This paper describes recent changes in Dragon's speech recognition system which have markedly improved performance on conversational telephone speech. Key changes include: the conversion to modified PLP-based cepstra from mel-cepstra; the replacement of our usual IMELDA transformation by a new transform using "semi-tied covariance"; a new multi-pass adaptation protocol; probabilities on alternate pronunciations in the lexicon; the addition of word-boundary tags in our acoustic models and the redistribution of model parameters to build fewer output distributions but with more mixture components per model.

1. INTRODUCTION

Recognition of conversational telephone speech has progressed dramatically since the introduction of the Switchboard corpus [1] in the early 90's. While originally on the receiving end of improvements largely developed for other domains, such as DARPA's Wall Street Journal task, the Switchboard task has progressed so far and recognition output is now so intelligible, that work on conversational telephone speech has become a standard domain for gauging general speech recognition improvements, as well as a prime candidate for commercial applications. There is now a rich exchange of techniques and ideas between the Switchboard development effort and WSJ's descendant, the recognition of Broadcast News.

Two years ago, Dragon presented improvements to its Switchboard recognition system in [2]. Since that time, error rates have continued to decline and the development set has been augmented by the addition of English language calls from the more challenging CallHome corpus, a multilingual corpus of international telephone conversations between friends and family. We now present results on our next stage of improvements, which result in substantial reductions in error rate over the system reported in [2].

In the sections which follow, we focus on some of the key improvements: a revision of the front-end to use modified-PLP cepstra and a new "diagonalizing" transformation using semi-tied covariance, inclusion of word-boundary tags in the acoustic models and a shift to fewer output distributions with more mixture components per distribution, a new multi-pass adaptation protocol, and the introduction of probabilities on pronunciation variants in the lexicon.

The reader should be warned that we have found testing of improvements highly volatile on this corpus: one can get widely differing readings as one varies the test set. (Consequently, results cited in the literature only on small tests, especially in the absence of reported significance levels, should be viewed with caution.) We therefore report most results below on a variety of tests: the 20 conversation-side "CAIP" set of Switchboard calls (about 9k words), a test of 15 two-sided Switchboard conversations we refer to as "SWB'95" (roughly 13k words), and a collection of 20 two-sided CallHome calls "CH'96" (20k words). Although the results on individual test sets vary a great deal, we find the aggregated results over several tests a generally reliable indicator of performance.

2. IMPROVING THE FRONT-END

Dragon's standard signal processing at the time of [2] generated a 44-parameter feature vector every 10 msec: 8 spectral parameters, 12 mel-cepstral parameters, 12 mel-cepstral differences, and 12 mel-cepstral second differences. These parameters were mapped down to a family of 24 under an IMELDA transformation [3], which had the benefits of reducing the degree of correlation within the feature set and reducing the number of features needed as well. Since then, we have improved our cepstral calculation and have taken a new look at the role of "diagonalizing" transformations in our models.

2.1 PLP Parameters

In 1997 we converted our standard front-end to use modified PLP-based cepstra rather than our usual mel-cepstra. We found that PLP ("perceptual linear prediction") coefficients [4] provided a more robust front-end, especially for mismatched train/test conditions and/or when the amount of training data was limited. Our implementation follows the modification of PLP processing reported by Cambridge/HTK [5].

Table 1 shows the effect of the switch to PLP processing on a set of models trained from only 13 hours of Switchboard data as well as more "evaluation-quality" models trained from 170 hours. We show results both on the "CAIP" set and on a test of CallHome/English conversations before the addition of CallHome data to our acoustic training set.

	13-hr models		170-hr models	
	orig	PLP	orig	PLP
CAIP	45.0	43.6	39.4	39.7
CH'96	60.4	58.5	54.6	53.4

Table 1: Improvement from moving to modified PLP-based cepstra. (Figures give word error rates.)

The win from PLP processing continues to hold up even as our acoustic modelling improves, as shown in Table 2 for a recent set of 60-hour models. (These models also use a pure Switchboard-trained trigram language model instead of the bigram models used in Table 1 -- a better model for Switchboard, but not as good a fit to CallHome data.)

	CAIP	SWB'95	CH'96	overall
original	38.6	41.1	56.6	47.7
PLP	37.5	40.4	54.0	46.1

Table 2: Further improvements from modified PLP-based cepstra, now for better 60-hour acoustic models.

2.2 Diagonalizing Transformations

Recently, we have been re-examining our use of the IMELDA transformation and, in parallel with our work on the Broadcast News corpus (reported in [6]), have explored more general "diagonalizing" transformations.

Inspired by the work of Kumar [7] and Gales [8] on generalizations of LDA, we have been testing transformations based on what Gales calls "semi-tied covariance". The main idea is that, because we assume a diagonal covariance in the multivariate gaussians used in our acoustic models, we should seek a representation of acoustic space that most closely realizes this assumption. For more detail on Dragon's implementation of semi-tied covariance, see [6].

Table 3 shows the improvement from switching to a transformation based on semi-tied covariance. The "44/24" models start from our standard 44-parameter processing, mapped to 24 under either our usual IMELDA transform (IM) or using a semi-tied covariance mapping (ST) trained on the 24-parameter IMELDA-ized space. The "36" models use only the 36 cepstral and difference parameters, leaving out the spectral features. All models are trained from the 170-hour training set and use a trigram language model. (They also incorporate the acoustic modelling improvements described in the next section.)

	CAIP	SWB'95	CH'96	overall
44/24 IM	34.3	36.8	50.8	42.7
44/24 ST	33.6	35.6	49.5	41.6
36 IM	34.6	36.4	50.1	42.2
36 ST	32.8	35.7	49.8	41.6

Table 3: IMELDA vs. semi-tied covariance for spectral+cepstral and pure cepstral parameters.

This implementation of semi-tied covariance is quite new and wrinkles are still being ironed out, but already results look quite promising.

3. CHANGES TO ACOUSTIC MODELS

The acoustic models we are using are triphone models with 2 to 4 nodes arranged linearly (but with single skips allowed), each node having an output distribution, which we call a 'PEL' for "phonetic element", and a duration distribution. Which PEL model to employ in a given position of a triphone is determined based on decision trees whose nodes ask linguistic questions about neighboring phonemes. The PEL models themselves are mixtures of multivariate gaussian distributions. All models described here are gender-independent, speaker-normalized models.

The decision trees used to determine the sharing of models among phonetic contexts now include the capability of asking questions about the position of word-boundary as part of the phonetic context. The improvement from using such word-boundary tags appears to be small but significant, as demonstrated in the first two lines of Table 4 for a set of models trained from 60 hours of Switchboard data and using a trigram language model.

word bndry	# PELs	Max Comps	recognition WER			
			CAIP	SWB'95	CH'96	overall
no	14.4k	20	37.4	39.5	54.7	46.1
yes	14.4k	20	36.3	39.3	54.4	45.6
yes	7.5k	64	34.9	38.8	52.7	44.4

Table 4: Effect of word-boundary modelling and fewer PELs with more components.

We have also been exploring different schemes for allocating parameters in our acoustic models. In earlier years, we built models with tens of thousands of PEL models and up to 20 gaussian components per mixture. We are now seeing substantial improvements from using fewer PELs but with many more components per PEL. The last line of Table 4, above, shows a sample result for the 60-hour models.

Table 5 shows the performance of "evaluation-size" models trained from 170 hours of data but using our older (non-PLP) signal processing. All models use word-boundary markers. The first line shows models like our earlier evaluation-style models with many PELs and up to 20 components per PEL. The other lines show the effect of markedly cutting back the number of PELs and then increasing the number of components. Note that the models with 7.5k PELs and up to 64 components/PEL involve roughly the same number of model parameters as the 24k PEL / 20 component models.

# PELs	Max Comps	recognition WER			
		CAIP	SWB'95	CH'96	overall
24k	20	36.8	38.0	53.7	45.0
7.5k	20	37.2	39.0	54.8	45.9
7.5k	64	35.0	37.7	52.6	44.0
7.5k	96	34.6	37.0	52.1	43.5

Table 5: Effect of reducing #PELs and increasing #components per PEL.

4. ANOTHER LOOK AT ADAPTATION

While reviewing our procedure for rapid adaptation and speaker-adaptive training (SAT), as earlier described in [2], we made a rather remarkable discovery. We took two models: model A was built directly from the training data, and model B used our SAT

algorithm to transform the data. We then ran the following series of experiments:

(1) recognize with A, adapt and re-recognize with A
(2) recognize with A, adapt and re-recognize with B
(3) recognize with B, adapt and re-recognize with A

Experiment	First round error rate	Second round error rate
(1) A → A	39.9	37.1
(2) A → B	39.9	36.4
(3) B → A	39.8	36.2

Table 6: Effect of permuting models between adaptation stages. Use first model to recognize, then adapt and re-recognize with second.

Two years ago, we showed that experiment (2) had an error rate about 1.5 points lower than experiment (1). At the time, we (and various other sites) claimed that this was because model B was "more focused" and therefore better for rapid adaptation. However, we now find that experiment (3) gets the same error rate as experiment (2), and both are better than experiment (1). (Note that these numbers are on the SWB'95 test set, and the gain from SAT is somewhat smaller than seen earlier. But it is still real: a matched-pairs test shows that both (2) and (3) are better than (1) at a significance level of $P < 0.01$, while (2) and (3) are not significantly different from each other.)

It seems clear that SAT is not doing what we originally thought. The best explanation that we can come up with is "jiggle". Models A and B are equivalent, but different. In particular, they make different errors, and when we use the recognition results from one to adapt the other, we are effectively "interpolating" between the two sets of models.

Additional gains come from iterating the adaptation/recognition step, and we have found two techniques for enhancing this effect. The first is jackknifing on the recognition transcript, to avoid "locking in" the recognition errors. For example, we adapt on all but the first utterance, then recognize the first, and so on. A less computationally expensive alternative is to bundle the utterances into a reasonable number of clusters of approximately equal size. In our experience we get best results if we jackknife between the first and second rounds of recognition (but not between second and third), and typically we see results between 0.5 – 1.0 points better than without jackknifing.

To get the full benefit of iterating, we find we need to change the clustering (that we use for the rapid adaptation classes) between the two rounds of adaptation. Other sites (see, for example, [5]) have reported getting better results from iterating when they gradually increase the complexity of their transformations. It may well be that the issue is not necessarily complexity, but rather that changing the transformation classes between rounds is another way of mixing things up and avoiding locking in errors. In Table 7 we show the results of combining jackknifing and switching classes with the results of leaving out one of these two steps. The recognition both here and in Table 6 above use 60-hour models and our standard Switchboard trigram language model. (Note that the results of the "best" run are significantly better than "no jackknifing" with $P < 0.02$). Overall on this test set we see a full 5 points from adaptation, though results vary with the models, and in particular we appear to get smaller gains using models built from more training data.

	1st round	2nd round	3rd round
Best	39.9	35.9	34.9
No Jackknifing	39.9	36.4	35.5
Same classes	39.9	35.9	35.8

Table 7: Improvements from jackknifing and from changing transformation-class assignments between passes.

An obvious question is how far we can get by iterating the adaptation beyond the third round. If we continue one more time, we get a statistically insignificant improvement of 0.2 points. We can get an approximate lower bound by adapting on the correct transcript. It is well-known that the naïve cheating experiment simply "locks in" the correct transcript, resulting in an artificially low error rate. We can avoid this pitfall by jackknifing as described above, and the result is 34.3%. This number is astonishingly close to the best error rate obtained above. It is telling us that given the limitations of the rapid adaptation algorithm, the recognition errors have almost no effect on the adapted models.

5. PRONUNCIATION MODELLING

One of the greatest challenges in recognizing conversational telephone speech lies in correctly modelling the informal, generally highly reduced, pronunciations one encounters there. There have been many reports on pronunciation modelling efforts (see, for example, [9]), but performance has generally been mixed: although reduced pronunciations may provide a better match to actual speech data than the "standard" dictionary pronunciations, they add a great deal of acoustic confusability to the lexicon which may result in hurting as much as it helps.

We have run a number of experiments introducing reduced forms of common words (our pronunciation dictionary already included alternate forms for a number of words, but nothing close to the diversity found in natural speech). Initial results demonstrated some gain, on the order of 1.0-1.5% absolute. However, we have since discovered that most -- though not all -- of the benefit came simply from introducing probabilities on the variant pronunciations. Formerly, all alternate pronunciations of a word had been treated as equi-probable, but we realized that we would need probabilities on the pronunciations in order to control how frequently and in what contexts these highly reduced forms were hypothesized.

Table 8 shows the effect of adding probabilities for alternate pronunciations into the language model at the unigram and the bigram levels. The lexicon used is the same as that in the preceding two sections: a 28k vocabulary including 32k pronunciations. No additional reduced forms were added. The language model here is somewhat richer than that used above -- hence the lower baseline error rate. It is an interpolation of trigram models trained from Switchboard, CallHome, and Broadcast News data.

	CAIP	SWB'95	CH'96	overall
no probs on prons	31.1	33.3	45.0	38.2
unigram probs	30.6	33.2	44.0	37.6
bigram probs	30.4	32.8	43.6	37.2

Table 8: Effect of imposing probabilities on alternate pronunciations at the unigram and bigram levels.

6. TESTING THE COMPLETE SYSTEM

The improvements described above were incorporated into a single system fielded in the 1998 Hub 5 evaluation. The system uses the 44/24ST processing described in section 2. The acoustic models include 8500 full-triphone PELs with up to 96 components each, trained from 170 hours of data, including about 12 hours from CallHome/English. Matched sets of regular and SAT models were constructed.

The system runs in several recognition stages: A first recognition pass is run using speaker-independent models and somewhat tighter thresholds than in later passes. The resulting recognition transcripts are extracted and used for (unsupervised) adaptation during the second pass. This second pass adapts SAT models to the recognition output. It makes use of the jackknifing protocol described above and uses a family of 8 transformations for each conversation side, with transformation classes determined automatically based on clustering PELs via a distance metric. The resulting recognition from these second-stage models is then used to adapt the SAT models once again, now without jackknifing but using 7 expert-determined transformations per side. A third recognition pass is run to produce the final recognition hypothesis. All passes use the same interpolated trigram language model referred to in section 5.

In Table 9, we compare the performance of the resulting system to that of our 1997 evaluation system, which included the PLP processing but none of the other improvements described here. The test set is the 1997 Hub-5E evaluation set, composed of 40 5-minute conversations, 20 from Switchboard-II and 20 from CallHome (about 38k words total). (Switchboard-II is a relatively new corpus currently being collected by the Linguistic Data Consortium, similar in style to Switchboard-I but somewhat more challenging.)

	CH	SWB-II	overall
eval'97 system	57.4	39.9	48.9
eval'98 system			
pass 1	55.8	38.3	47.3
pass 2	52.0	34.5	43.5
pass 3	51.0	33.6	42.6

Table 9: Comparison of word error rates for 1997 and 1998 evaluation systems on eval'97 data.

The table shows the benefit of the series of adaptation stages and demonstrates the remarkable consistency of the improvements across the two corpora.

7. 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] B. Peskin et al., "Progress in Recognizing Conversational Telephone Speech," *Proc. ICASSP-97*, Munich, April 1997.

[3] M.J. Hunt et al., "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination," *Proc. ICASSP-91*, Toronto, May 1991.

[4] H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis for Speech," *J. Acoust. Soc. Amer.*, vol. 87, 1990, pp. 1738-1752.

[5] P.C. Woodland et al., "The Development of the 1996 HTK Broadcast News Transcription System," *Proceedings of the Speech Recognition Workshop*, Chantilly, Feb 1997.

[6] S. Wegmann et al., "Progress in Broadcast News Transcription at Dragon Systems," submitted to these *Proceedings*.

[7] N. Kumar, "Investigation of Silicon-Auditory Models and Generalizations of LDA for Improved Speech Recognition," *Ph.D. Thesis*, Johns Hopkins University, 1997.

[8] M. Gales, "Semi-tied Covariance Matrices," *Proc. ICASSP-98*, Seattle, May 1998.

[9] B. Byrne et al., "Pronunciation Modelling for Conversational Speech Recognition: A Status Report from WS97," *Proc. IEEE ASRU Workshop*, Santa Barbara, Dec 1997.