PROGRESS IN BROADCAST NEWS TRANSCRIPTION AT DRAGON SYSTEMS

Steven Wegmann, Puming Zhan, and Larry Gillick

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

ABSTRACT

In this paper we shall report on recent progress in acoustic modelling and preprocessing in our Broadcast News transcription system. We have gone back to basics in acoustic modelling, and re-examined some of our standard practices, in particular the use of IMELDA and frequency warping, in the context of the Broadcast News corpus. We shall also report on some preliminary experiments with a generalization of IMELDA, "semi-tied covariances". In combination, these improvements lead to a 3.5% absolute improvement over our eval97 models. We shall also describe our attempts to fix our rather primitive, silence-based preprocessing system, including initial results using a new speaker-change detection algorithm based on Hotelling's T^2-test.

1. INTRODUCTION

In our 1997 Broadcast News transcription system [10], we used IMELDA [7] to decorrelate our input features. Section 2 of this paper compares systems built with and without IMELDA and also with a generalization of IMELDA, called "semi-tied covariances". We shall show that moving from IMELDA parameters to non-IMELDA to semi-tied covariance yields improvements at every stage. Our 1997 evaluation system also used frequency-warped data, and section 2 includes a comparison of frequency warping with gender-dependent modelling. The results of these experiments are ambiguous but intriguing.

Broadcast News data comes to us in long unsegmented speech streams, as more or less unadulterated TV and radio broadcasts. We process these broadcasts into smaller, homogeneous segments that are then clustered into speaker-like parcels of data that may be used, for example, for speaker adaptation. Our 1997 evaluation system used only information about silence to make decisions about when to chop, which led to a small but damaging number of inhomogeneous segments. As we shall see, these inhomogeneous segments led to spurious mixed-speaker clusters, which degraded adaptation performance. We have been working towards eliminating multiple speakers in our automatically generated segments by incorporating a speaker-change detection algorithm into our system. We shall report on some promising preliminary experiments in section 3.

This work was supported by the Defense Advanced Research Projects Agency. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the U. S. Government and no official endorsement should be inferred.

2. ACOUSTIC MODELLING

2.1 Preliminaries

The front end used in all of the experiments in this paper produced 36 features every 10 milliseconds: 12 cepstral features, 12 cepstral differences, and 12 cepstral second differences. These cepstra are PLP-based [6], computed in the style of Cambridge/HTK [12]. Our usual practice is to apply an IMELDA transformation to this feature set, and then project the resulting 36-dimensional feature set down to a 24-dimensional one.

All of the acoustic models that we shall be describing were trained from the 1997 Broadcast News acoustic training corpus (about 70 hours of training data). These models are general mixture models, and use decision-tree state clustering [9]. To facilitate meaningful comparisons, all of these models are about the same size as our 1997 evaluation models: 9000 states with up to 32 gaussians per state resulting in 190,000 gaussians.

The language model that we used in all of these experiments is a small bigram language model trained from the 1997 Broadcast News acoustic training transcripts. The lexicon has 57k words obtained from these texts together with the 1997 Broadcast News language model training corpus.

All of these models are tested on the 1996 PE devtest. For the results in this section, we use the known speaker identities to segment the data and to cluster it for adaptation. This test set consists of a little over two hours of speech from 106 speakers. We will be reporting word error rates (WER) on this test set broken out by the Broadcast News focus conditions [5]:

Focus	*Description*
F0	Studio speech (clean, planned)
F1	Spontaneous speech (clean)
F2	Reduced bandwidth speech (clean, planned)
F3	Speech in the presence of background music
F4	Speech under degraded conditions
F5	Non-native speech (clean, planned)
FX	All other speech

2.2 Studies on IMELDA

Recall that the IMELDA transformation is a linear transformation that is a product of two orthogonal transformations. The first diagonalizes and then rescales the average within-class covariance matrix to the identity, and the second diagonalizes the resulting between-class covariance

matrix.

How should we use IMELDA, if at all, in our acoustic model training? If we use IMELDA should we keep all of the features, or should we project down to an uncorrelated subspace? Our current practice of using 24-dimensional IMELDA features came from our experiences with the WSJ corpus. We'll re-examine these questions in the Broadcast News corpus, by testing if IMELDA helps in the training of our state clustering and/or in the training of our output distributions.

We built four sets of models, all gender-independent using unwarped data. The first set of models, "24 IMELDA", were trained from 24-feature IMELDA-transformed data, including the state clustering. The second set of models, "36 IMELDA", were trained from 36-feature IMELDA-transformed data, including the state clustering. The third set of models, "36 Non-IMELDA", share the same state clustering as the "36 IMELDA" models, but otherwise were trained from the raw, untransformed 36-dimensional feature set. The fourth set of models, "36 Full Non-IMELDA", were entirely trained from the raw, untransformed training data, including the state clustering.

	24 Imelda	36 Imelda	36 Non-Imelda	36 Full Non-Imelda
F0	23.9	23.2	23.0	23.5
F1	38.8	37.4	37.3	37.6
F2	49.4	48.6	47.6	49.1
F3	42.3	43.3	40.4	39.7
F4	33.8	33.2	32.5	31.8
F5	42.2	42.2	41.9	41.9
FX	64.5	63.6	64.7	65.5
Total	41.1	40.5	40.1	40.4

Table 1: Variations on use of IMELDA. (Figures give word error rates.)

Table 1 summarizes the performance of these models. On the Broadcast News corpus at least, it appears that IMELDA is best used only when training the state clustering.

2.3 Semi-tied Covariances

In the last section we saw that IMELDA does not appear to work very well with our current modelling techniques on the Broadcast News corpus. In this section we shall explore using a linear transformation that is a generalization of IMELDA due to Gales [3] and Kumar [8], and first applied to the Broadcast News corpus by Gopinath [4]. The main point is that since we use diagonal covariances in the multivariate gaussians we use in our output distributions, we should try to find a linear transformation that makes this diagonal covariance assumption as reasonable as possible. We can also try to construct multiple, class-specific transformations. In fact, if each gaussian gets its own transformation, then this is equivalent to using full covariances in our models, which is why Gales named this technique "semi-tied covariances". The details of the required numerical optimizations are provided in [3], [8], and [4]. We shall refer to this technique as semi-tied covariances, as in [3], but we shall call the resulting transformations "diagonalizing" transformations.

For the experiment described below, we trained a single transformation, using the state-level alignments produced before we fit the mixtures of gaussians to each state. We trained this transformation using numerical techniques as described in [8] and [4]. In brief, we write down the likelihood of the training data where we fit a single multivariate gaussian per state using the sample mean and diagonal of the sample covariance matrix of the transformed data (i.e. the maximum likelihood estimates (mle)). We want to maximize this likelihood with respect to the transformation, so we differentiate the likelihood equation with respect to the transformation and use conjugate gradient descent to find a local maximum. We have also tried using a separate transformation for each phoneme (as in [3]), but so far this has been less successful than using a single transformation.

If in addition, we insist that all of the states share the same covariance matrix, then the mle for this shared covariance matrix is the average within-class covariance matrix. It is a simple exercise to show that the resulting likelihood equation has a single global maximum value. The orthogonal transformation which diagonalizes the average within-class covariance matrix, i.e. the IMELDA transformation, is one transformation that realizes the maximum.

Table 2 compares the performance of these models, "36 SemiTied", with three models from the previous section.

	24 Imelda	36 Imelda	36 Non-Imelda	36 SemiTied
F0	23.9	23.2	23.0	23.2
F1	38.8	37.4	37.3	36.2
F2	49.4	48.6	47.6	45.0
F3	42.3	43.3	40.4	39.4
F4	33.8	33.2	32.5	33.1
F5	42.2	42.2	41.9	40.5
FX	64.5	63.6	64.7	62.1
Total	41.1	40.5	40.1	39.0

Table 2: Semi-tied covariance vs. IMELDA.

2.4 Gender-Dependent vs. Warped Models

Another useful technique that Dragon routinely uses is frequency warping ([9], [11]). In this section we shall compare gender-dependent (GD) models with frequency-warped models and examine how these techniques interact with the semi-tied covariance technique.

In the second and third columns of table 3 we compare the performance of two sets of models. The "36 SemiTied" models were described in the previous section, while the "36 SemiTied GD" models are GD versions of these models created by adapting the "36 SemiTied" models to the gender-specific training data (in the style of [13]). This gives an impressive 2.2% (absolute) improvement, far more than we typically see for Broadcast News data when we move to gender-dependent models. (We typically see only about a 1-point improvement, due to the extreme gender imbalance in this corpus.)

In the fifth column we used warped test data, which results in an additional small improvement, but one that vanishes when we use unsupervised rapid adaptation (with one transformation).

	36 SemiTied	36 SemiTied GD		36 SemiTied GD warped test data	
			Adapted		Adapted
F0	23.2	21.6	20.4	21.6	20.5
F1	36.2	33.4	32.6	33.0	32.2
F2	45.0	43.2	38.0	40.6	36.9
F3	39.4	38.0	35.3	38.7	36.4
F4	33.1	30.1	28.0	30.3	28.7
F5	40.5	37.7	35.0	37.3	35.1
FX	62.1	60.0	56.7	59.5	57.4
Total	39.0	36.8	34.4	36.2	34.5

Table 3: Effect of gender-dependent models and frequency-warping of test data.

We wanted to verify that warping and the semi-tied covariance technique work together. To do this we built baseline acoustic models from warped training data, "36Warp". We then trained a single diagonalizing transformation using warped training data and the state assignments defined by the "36Warp" state clustering. The "SemiWarp" models used this diagonalizing transformation in training, and the "36Warp" state clustering. Table 4 shows that we get an absolute improvement of 1.2% from using a diagonalizing transformation, which is comparable to the improvement that we saw in the unwarped case (1.5%).

Note that the "SemiWarp" models have the same overall performance as the "36 SemiTied GD" models. We built GD versions of the "SemiWarp" models, the "SemiWarp GD" models, by adapting them to the (warped) male and female portions of the training data. Table 4 shows that we do get an improvement, but after we adapt it disappears.

	36 Warp	SemiWarp	SemiWarp GD	
				Adapted
F0	22.5	21.9	21.6	20.5
F1	35.6	36.0	34.5	34.7
F2	41.9	40.6	39.6	37.4
F3	38.2	38.6	36.7	34.5
F4	31.7	29.9	30.1	27.5
F5	41.1	38.1	36.1	34.2
FX	61.9	57.9	58.1	54.6
Total	38.0	36.8	36.0	34.4

Table 4: Warping training data, with standard processing, semi-tied covariance, and semi-tied gender-dependent models.

Even though the adapted "SemiWarp GD" models and the adapted "36 SemiTied GD" models with and without warped test data all perform at the same WER, they make different errors. We took advantage of this fact by combining the hypotheses from these three recognizers using NIST's ROVER software [2]. This combined system had a WER of 33.7%. Thus, rather than choosing between warping and gender-dependent modelling, we may profit from combining them in various ways.

To get a sense of how much progress we have made, we can compare these systems with last year's acoustic models. When using the same small bigram language model, our 1997 HUB4 evaluation acoustic models have an adapted WER of 35.9%, but they used speaker-adaptive training (SAT) and included supplementary training data from WSJ and WSJCAM0. The models that we used to seed the SAT process were trained only from the Broadcast News acoustic training corpus, so they provide a cleaner measure of our improvements. They had an adapted WER of 37.2%.

3. PREPROCESSING

The Achilles heel of our 1997 Broadcast News evaluation system was our segmentation algorithm. To create segments, we looked for long stretches of silence in the output of a phoneme recognizer. Because we only looked for silence, we could easily create segments that had multiple speakers in them, which could potentially create problems for our automatic clustering algorithm. Also, because we used a phoneme recognizer, we often broke in the middle of words, which caused problems for the actual recognition pass. We estimate that we lost about 2% (absolute) in WER due to these preprocessing errors, with about 1 percentage point due to chopping errors (excluding speech for various reasons) and the other 1 percentage point due to clustering errors [10].

Various reasons for the clustering errors were conjectured, but after careful study of the clusters that we created and the segments that were fed to the clustering algorithm, we now believe that the culprit was the inhomogeneity of the segments we were creating; a small but damaging number of segments contained mixed speakers.

Motivated by [1], we have been exploring the use of various speaker-change detection algorithms in the segmentation process, in an attempt to create more homogeneous segments. We have examined the Bayesian information criterion (BIC) described in [1], and are currently evaluating a related, but somewhat simpler, method based on Hotelling's T^2-test.

Like BIC, this new method looks at a moving window sweeping through the speech stream. Within this window, we might hypothesize a speaker break-point splitting the frames into speaker A and speaker B. For each hypothesized break-point, we compute the T^2-statistic

$$T^2 = (\mu_A - \mu_B)^T [\ \Sigma\ (1/n_A + 1/n_B)\ ]^{-1} (\mu_A - \mu_B)$$

where μ_A and μ_B are the vectors of parameter means for the two pieces, n_A and n_B are the frame counts, and Σ is a "universal" within-speaker covariance matrix. We decide to break at peak values of T^2. This is a multivariate analogue of the more familiar *t*-test, where n_A and n_B correct for the relative sizes of the two segments.

An advantage of this method over BIC is that the (full) covariance matrix Σ can be trained once and for all from a large representative selection of known single-speaker training segments, rather than requiring that it be re-estimated in each window, generally on far too few frames to make the estimation robust. The new system is computationally cheaper and allows better decisions on small segments. As a first step, we have implemented it still re-estimating the covariance on an utterance-by-utterance basis, but plan to move to a universal covariance next.

In table 5, we compare the properties of BIC and the T^2-test with our original segmentation algorithm as well as the reference turn-mark chopping on the 1996 PE devtest. In all cases we used an automatic algorithm to cluster the resulting segments: '#Clusters' is the number of resulting clusters. 'Spk/seg' is the average number of speakers per segment, 'Seglen' is the average length of the segments measured in seconds, and 'Change Misses' gives the number of speaker changes where we failed to break (out of the 487 changes in this test set).

	#Clusters	Spk/seg	Seglen (sec)	Change Misses
reference	100	1.0	14.0	0
original	195	1.1	4.1	141
BIC	133	1.2	9.3	126
T^2-test	135	1.1	8.3	97

Table 5: Comparison of speech segmentation algorithms.

Examining table 5, we see that the T^2 method misses fewer speaker changes, and creates longer segments that cluster more efficiently than our older method. Although the T^2 method is still in an early development stage, we are encouraged by these statistics.

Table 6 demonstrates the resulting improvement to recognition and adaptation. It reports the WER of acoustic models that are almost identical to the 36 Non-Imelda models of Table 1, before and after unsupervised adaptation within the automatically generated clusters. (This test setup is otherwise the same as that in section 2, e.g. we are using the same bigram LM.) We should warn the reader that the test set we are using is somewhat peculiar since there are a few untranscribed regions of speech that our automatic segmenter includes but the reference segments do not, which contribute about 0.5% absolute to the WER for all automatic segmentation results.

Incidentally, the adapted WER of 37.7%, reported in the 'reference' row, is identical to the adapted WER when we use the true speaker identities to produce the clusters. In other words, our automatic clustering is working well, given pure input segments.

	Unadapted	Adapted
reference	39.9	37.7
original	41.4	39.7
BIC	40.8	38.7
T^2-test	40.8	38.4

Table 6: Recognition performance using various segmentation methods. (Figures give word error rates.)

4. FUTURE RESEARCH

The frequency-warping versus GD-modelling experiments are very intriguing. It is quite surprising that we end up with (overall) equivalently performing models, even if they do make different errors. However, our method for warping is probably not optimal (see [10]), so this deserves further study.

We shall be adding bandwidth, music and gender detection to our segment generator, which should lead to further improvements in our preprocessing system.

5. REFERENCES

[1] Chen, S., and Gopalakrishnan, P. "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion". *Proceedings of the Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, pp. 127-132. Feb 1998.

[2] Fiscus, J. "A Post-Processing System to Yield Reduced Word Error Rates". *Proc. IEEE ASRU Workshop*, Santa Barbara, pp. 347-352. Dec 1997.

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

[4] Gopinath, R. "Constrained Maximum Likelihood Modeling with Gaussian Distributions". *Proceedings of the Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, pp. 110-115. Feb 1998.

[5] Garofolo, J., Fiscus, J., and Fisher, W. "Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora". *Proceedings of the Speech Recognition Workshop*, Chantilly, VA, pp. 15-21. Feb 1997.

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

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

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

[9] Roth, R., et al. "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer". *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, pp. 116-120. Jan 1995.

[10] Wegmann, S., et al. "Dragon Systems' 1997 Broadcast News Transcription System". *Proceedings of the Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, pp. 60-65. Feb 1998.

[11] Wegmann, S. et al. "Speaker Normalization on Conversational Telephone Speech". *Proc. ICASSP-96*, Atlanta. May 1996.

[12] Woodland, P. et al. "The Development of the 1996 HTK Broadcast News Transcription System". *Proceedings of the Speech Recognition Workshop*, Chantilly, VA, pp. 73-78. Feb 1997.

[13] Woodland, P., et al. "The 1997 HTK Broadcast News Transcription System". *Proceedings of the Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, pp. 41-48. Feb 1998.