Presented at the 1999 DARPA Broadcast News Workshop, Herndon VA, 1999feb28

Reducing errors by increasing the error rate: MLP Acoustic Modeling for Broadcast News Transcription

Nelson Morgan, Dan Ellis, Eric Fosler-Lussier, Adam Janin, and Brian Kingsbury
Email: {morgan,dpwe,fosler,janin,bedk}@icsi.berkeley.edu

International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
University of California at Berkeley, EECS Department, Berkeley, CA 94720
Tel: (510) 643-9153, FAX: (510) 643-7684 *

ABSTRACT

We describe some aspects of a Broadcast News recognition system based on hybrid HMM/MLP acoustic modeling. These include the use of novel 'modulation spectrogram' features which are combined with conventional models at the posterior probability level, some experiments with nonlinear segment normalization, and an investigation of the interaction of model size and training set size for an multi-layer perceptron (MLP) acoustic classifier. We also report preliminary results of incorporating gender-dependence into this system.

1. Background

In recent years, we and our colleagues have promoted the exploration of novel, poorly understood, but promising approaches to speech recognition [2]. While such deviations from incremental improvements might initially hurt performance, the subset of the new methods that would ultimately prove useful would not be found without such explorations. This past year, we attempted to follow this advice, while still developing a system with reasonable performance on the automatic transcription of Broadcast News speech. An additional goal was finding approaches that would work well in combination with components developed by our SPRACH partners at Cambridge and Sheffield. Finally, previous published results seemed to indicate that, while the hybrid HMM/connectionist approach was successful for moderate sized training corpora, it did not appear to take advantage of significant increases in the size of the corpus. Recently improved computational capabilities at ICSI permitted tests to determine if this was true.

Given these considerations, we developed experimental Broadcast News systems that incorporated:

- a new feature extraction module incorporating auditory characteristics in a somewhat different way than was done for Perceptual Linear Prediction (PLP) [4] or RASTA-PLP [5];
- multiple-stream analysis with probability combination at the frame level
- nonlinear segment normalization, implemented by adding statistics for the whole segment as inputs to the acoustic model training; and
- extended experiments to determine the advantage provided by increased training data and/or larger number of model parameters.

All of these systems used a single large multi-layer perceptron (MLP), trained by back-propagation to produce estimates of context-independent phone class probabilities. Additionally, to form a contrast with the Cambridge system, all of our networks were trained on telephone bandwidth (using all the Broadcast News data, but low-pass filtered at 4 kHz).

The recognizer that resulted from combining our system with the one developed by our partners at Cambridge performed moderately well, despite our choice of approaches that had little history for large vocabulary systems. Subsequent to the evaluation, we have developed additional models that exceed this level of performance using only our MLP approach, confirming that the evaluation results were not due to some chance interaction between the two systems.

There are a host of commonly-used approaches that we did NOT incorporate, as their refinement for our system would have required significant time. They include context-dependent modeling (which was, however, used by our Cambridge colleagues for their recurrent networks that formed one of the streams for our joint evaluation system); and adaptation (though we did try some new normalization schemes, as described below). We would expect to ultimately include such components in future Broadcast News transcription systems, but since they are already known to help, we considered them to be secondary in importance to testing less-explored territory. However, after the evaluation, we experimented with another standard enhancement, gender-dependent modeling. This had the expected effect of moderate reductions in error, which we will describe below.

*This work was supported by the SPRACH grant from the European Union, as well as by National Science Foundation grant 9712579.

2. Experimental Systems

2.1. Adding new features

A new set of features called the Modulation-filtered Spectrogram (MSG) were developed at ICSI over the last few years [7]. While brevity does not permit a full explanation of these features, essentially they provide explicit automatic gain control and modulation filtering for each critical band energy contour. The gain control characteristics were provided implicitly in RASTA-PLP [5], but were explicitly designed in MSG to optimize recognition performance under a range of acoustic conditions. (As with the PLP features, each feature dimension was also normalized to zero mean and unit variance over each segment). The temporal characteristics of these features tend to be more 'sluggish' than those for PLP. This sometimes hurt performance under acoustic conditions that matched the training set, but in tests with smaller tasks these features seem to often improve performance in combination with PLP or RASTA-PLP [12].

The MSG features are represented in spectral rather than cepstral form, and, as we used them, consisted of two groups of 14 features each. Each group employed a different modulation filter; one was a lowpass filter constraining the modulation spectrum to 16 Hz, and the other was a bandpass filter from 2 to 16 Hz. In each case, the filter processed a sequence energies from each critical band. Multiple AGCs were used to limit the gross amplitude variability and to emphasize transitions.

We and our partners did a number of pilot experiments with combination schemes, but for the purposes of the evaluation used a simple framewise product of likelihoods determined separately for each feature set; PLP for the Cambridge RNN system, and MSG for the ICSI MLP system. Additionally, three different subsystems based on these features were combined at the hypothesis level (described in [3]).

2.2. Nonlinear segment normalization

A major source of improvements in Broadcast News systems over the past few years has been through local adaptation of acoustic models to the characteristics of each segment [11] (where a segment is defined as an utterance by a single speaker in constant acoustic conditions). The most popular technique used for this purpose with Gaussian-mixture models is Maximum Likelihood Linear Regression (MLLR) [8], but since neural network models do not have an explicit representation of class means, that algorithm cannot be directly employed.[1] At the level of input feature transformation, mean and variance normalization have become common, and are used in our systems as well. We wondered if there were more useful ways to combine framewise and segment-length features to reduce across-speaker variability. Given the ability of multi-layer perceptrons to model arbitrary nonlinear dependencies, we tried an alternative approach: We augmented our framewise feature vectors with inputs that we believed to be correlated with 'speaker character', then applied our standard learning algorithm, which minimizes the relative entropy between the framewise phone posteriors and the posteriors given by the forced alignments.

We conducted several experiments in which the input to our MLP classifier, consisting of 9 time-frames of the per-segment-normalized MSG feature vector, was enhanced with a small number of additional inputs intended to provide information on the overall character of the segment. We tried several different segment-level features, including pitch estimate histograms (to differentiate males and females), spectral maxima histograms (to indicate the formant positions), and 10th percentiles of the spectral level distribution in each band (to estimate the noise floor). In preliminary investigations, the best performing feature was the standard deviation of each feature element, prior to segment-level normalization; the means also provided some improvement. We used these two features for full tests.

2.3. Training set vs number of parameters

We and others have previously expressed concerns that our context-independent hybrid HMM/ANNs could not easily exploit large amounts of training data for tasks such as Broadcast News. To test this, we used PLP features to train MLPs for every combination of 4 network sizes (hidden layers of 500, 1000, 2000 and 4000 units) and 4 training-set sizes (corresponding to 1/8, 1/4, 1/2 and all of the 74 hours of 1997 training data). This data is plotted as a surface in Figure 1, which showed worthwhile gains to be had from increasing model size in step with training data over the entire range tested. We repeated this investigation with MSG features, obtaining very similar results. This strongly suggested that using the full 142 hours of 1997 and 1998 training data, and doubling the network size once again to 8000 hidden units (HUs), should be a worthwhile effort. This training run took 21 days on custom hardware we had developed, and thus we had time for only one attempt before the evaluation deadline. Again the larger networks had lower error on our development test set than the smaller ones; while our best 4000 HU net achieved an error rate of 29.3% in combination with the Cambridge acoustic model, the error rate for the final 8000 HU net was 26.8%, reflecting both the increased model complexity as well as the improved training targets made available through iterative realignments.

3. Results and Discussion

In the limited time before the Broadcast News evaluation, we were able to develop some of these approaches to the point

[1] Gradient-based adaptation techniques have sometime been used to similar effect for hybrid HMM/connectionist systems [9].

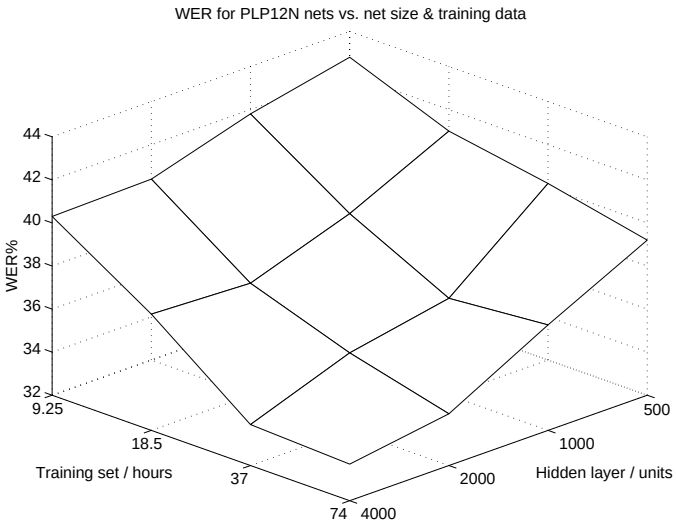

Figure 1: Surface plot of system word error rate as a function of the amount of training data and the hidden layer size.

of providing improvements. We briefly report here some of these results, and in addition we mention some of the tentative failures.[2]

3.1. MSG and PLP

As the tables show, MSG features (as used with an MLP and with reduced bandwidth data) were not as good as PLP features from the full bandwidth data (as used with the RNN). However, even in this case, the word error rate was significantly reduced by combining these two subsystems. The first columns of each table show this improvement. However, the second column shows that none of this improvement was obtained on the prepared, studio component of the shows. The third column underscores this point; essentially, all of the improvement was achieved for signals that were degraded from the clean read speech condition (F0).

system	ALL	F0	non-F0
RNN using PLP	25.1%	12.7%	27.5%
MLP using MSG	27.6%	16.1%	29.8%
Frame likelihoods product	23.3%	12.9%	25.3%

Table 1: Word error rates for RNN subsystem using PLP features, MLP subsystem using MSG features, and combined system (multiplying probabilities). F0 is the studio quality prepared speech condition. The scores given here differ slightly from the official scores due to some minor differences in segementation for the system reported here. Scores are for the first component of the 1998 evaluation set.

[2] Modesty and space limitations prevent our providing a complete list of the great ideas that didn't work.

system	ALL	F0	non-F0
RNN using PLP	23.9%	16.8%	29.5%
MLP using MSG	26.4%	19.3%	32.0%
Frame likelihoods product	22.7%	16.7%	27.5%

Table 2: Same as the previous table, except for the second component of the 1998 evaluation set.

3.2. Nonlinear segment normalization

Initial tests with using segment means and variances as MLP inputs were disappointing; performance actually got a bit worse. However, an analysis of the word errors as a function of segment length revealed that the augmented net was doing better than the baseline on longer segments, but a greater number of errors on the shorter segments was erasing this advantage. This observation seemed to make sense, since the segment-level statistics are estimated poorly for very short utterances. It also suggested the strategy of using the conventional system to recognize very short segments, and the augmented system only for those segments long enough for it to handle well. Pilot tests suggested a breakpoint of at least 22 seconds.

We trained a 4000 hidden unit network on 40 hours of 1998 BN acoustic data from segments longer than 25 seconds. Performance on our development test set was only slightly better, at 35.4%, than a comparable network that lacked the segment features (36.0%). Even this small gain was not observed consistently over other test sets, and so we did not use this approach for the evaluation. We have tentatively concluded that the standard normalization with means and variance may already be doing much of what we wanted, and that other features and possibly other methods for incorporating them may be necessary to compensate for speaker identity.

We have chosen this negative result to report (out of the many that we experienced) because we feel that the work is complete and that conclusions may be reasonably drawn. Another example of work in progress that has not yet yielded a positive result includes a significant effort on multi-band recognition; however, there are many aspects of this that we still do not understand.

3.3. Gender-based training and recognition

Subsequent to the November 1998 evaluations, we experimentally determined a procedure for incorporating gender-dependent models. In pilot studies we found that the following scheme worked best: (1) Train networks separately for speech from male and female speakers, starting each training with weights from a partially-trained gender-independent network. (2) During recognition, choose either the female or male net based on the average per-frame entropy of the clas-

sifier outputs; then choose between that gender-dependent model or a gender-independent one based on the overall utterance likelihood calculated in the decoding. We then trained large nets (8000 hidden units) on MSG features computed from 142 hours of training data. Rather than using the gender tags, we automatically labelled each segment as male or female using the same average model entropy measure used to choose between the gender-dependent nets in recognition, since we were actually more interested in clustering the data than the 'true' underlying gender. (This entropy measure agreed with the tags on 94.8% of the segments in the test set.) Further experiments revealed that a simple frame-level combination between gender-dependent and gender-independent nets (i.e., averaging the log posterior probabilities) was a equally effective way of combining the two models, and avoided the considerable expense of additional decodes required for combination schemes relying on utterance likelihood. Schemes of this kind improved the best gender-independent MSG net from 29.7% WER to 27.9%.

4. Conclusions

Briefly, we found that:

- Although MSG yielded higher error rates than MSG when used on its own, it was effective when used as a second stream with PLP, and combined at the level of frame likelihoods. All of this improvement came from the performance under conditions other than the read studio speech case (F0).
- Nonlinear segment normalization (in the form we have tried) does not seem to provide performance improvements over a simple normalization of the input features using their means and standard deviations over a segment.
- From a subset of 17 hours up to the full range of 140 hours of training data that we tried, doubling the number of training patterns used and/or doubling the size of the context-independent neural network continues to significantly improve the word error rate.
- As with earlier hybrid HMM/connectionist systems [1], and also Gaussian mixture HMM-based systems [6], incorporation of gender information can be used to provide moderate reductions in word error rate.

Finally, we note that this work required an immense effort, both on the part of our group and on the part of the time spent to do the computation for training. Our largest network took 3 weeks to train on a custom system using 4 Torrent vector microprocessors (developed at ICSI [10]), requiring roughly 3 x 10^{15} arithmetic operations.

5. Acknowledgments

By combining with the Cambridge system, we avoided the very poor results that might have resulted by starting off on our own; in particular we are very grateful to Gary Cook, who got us started in very short order.

References

1. Abrash,V., Cohen, M., Franco, H., Morgan, N., and Konig, Y., "Connectionist Gender Adaptation in a Hybrid Neural Network/Hidden Markov Model Speech Recognition System," *Proc. International Conference on Spoken Language Processing*, pp. 911-914, 1992.
2. Bourlard, H., Hermansky, H., and Morgan, N., "Towards Increasing Speech Recognition Error Rates," *Speech Communication*, pp. 205-231, May 1996.
3. Cook, G., Christie, J., Ellis, D., Fosler-Lussier, E., Gotoh, Y., Kingsbury, B., Morgan, N., Renals, S., Robinson, A.J., and Williams, G., "The SPRACH System for the Transcription of Broadcast News," *Proc. 1999 DARPA Broadcast News Workshop*, in press.
4. Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," *Journal Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, 1990.
5. Hermansky, H. and Morgan, N., "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4 pp. 578-589, 1994.
6. Huang, X., Alleva, F., Hayamizu, S., Hon, H.-W., Hwang, M.-Y., and Lee, K.-F., "Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 327-331, 1990.
7. Kingsbury, B., Perceptually-inspired Signal Processing Strategies for Robust Speech Recognition in Reverberant Environments. PhD thesis, University of California, Berkeley, CA, 1998.
8. Leggetter, C.J., and Woodland, P.C., "Maximum Likelihood Linear Regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, April, 1995.
9. Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., and Robinson, A.J., "Speaker Adaptation for Hybrid HMM-ANN Continuous Speech Recognition System," *Proc. Eurospeech '95*, pp. 2171-2174, 1995.
10. Wawrzynek, J., Asanović, K., Kingsbury, B., Beck, J., Johnson, D., Morgan, N., "SPERT-II: A Vector Microprocessor System," *IEEE Computer*, vol. 29, no. 3, pp 79-86, March 1996.
11. Woodland, P.C., Hain, T., Johnson, S.E., Niesler, T.R., Tuerk, A., Whittaker, E.W.D., and Young, S.J., "The 1997 HTK Broadcast News Transcription System," *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pp. 41-48.
12. Wu, S., Kingsbury, B., Morgan, N., and Greenberg, S., "Performance Improvements Through Combining Phone- and Syllable-Scale Information in Automatic Speech Recognition," *Proc. International Conference on Spoken Language Processing*, pp. 459-462, 1998.