Using MLP Features in SRI's Conversational Speech Recognition System

Qifeng Zhu[1] *Andreas Stolcke*[1,2] *Barry Y. Chen*[1,3] *Nelson Morgan*[1,3]

[1]International Computer Science Institute, Berkeley, CA
[2]SRI International, Menlo Park, CA [3]University of California, Berkeley, CA

Abstract

We describe the development of a speech recognition system for conversational telephone speech (CTS) that incorporates acoustic features estimated by multilayer perceptrons (MLP). The acoustic features are based on frame-level phone posterior probabilities, obtained by merging two different MLP estimators, one based on PLP-Tandem features, the other based on hidden activation TRAPs (HATs) features. This paper focuses on the challenges arising when incorporating these nonstandard features into a full-scale speech-to-text (STT) system, as used by SRI in the Fall 2004 DARPA STT evaluations. First, we developed a series of time-saving techniques for training feature MLPs on 1800 hours of speech. Second, we investigated which components of a multipass, multi-front-end recognition system are most profitably augmented with MLP features for best overall performance. The final system obtained achieved a 2% absolute (10% relative) WER reduction over a comparable baseline system that did not include Tandem/HATs MLP features.

1. Introduction

The goal of this work is to demonstrate that acoustic features estimated discriminatively as phone-level posterior probabilities can be used effectively to lower the error rate of large-vocabulary, speech recognition systems, above and beyond a host of state-of-the-art feature extraction and normalization techniques and in the context of a multipass recognition system using multiple model adaptation and system combination steps. In previous work [1, 2] we had shown that posterior features estimated by multilayer perceptrons can yield relative word error reductions ranging from 6% to 10%, but using less complex systems and smaller amounts of training data than would typically be used in a state-of-the-art system. The challenge for the present work was twofold. First, we had to scale up the (computationally expensive) feature training to very large training corpora of almost 2000 hours of speech. Second, we had to develop a system architecture that preserved (or increased) the sizeable wins seen in smaller systems in conjunction with an array of other techniques that could potentially diminish the relative gains obtained with our augmented feature stream. In fact, as we will show here, simply adding the additional features uniformly to all components of a multi-front-end, multipass recognition system does not yield the best results, and a more selective use of the augmented feature stream is advantageous.

2. MLP-based Frontend Features

We have been developing features based on multilayer perceptron (MLP) derived posteriors. Previous papers show the procedure on feature extraction using MLPs [2]. MLPs are trained by taking various snapshots of the time-frequency plane as input. We have found that posteriors from MLPs focusing on information derived from long time chunks of 500 ms can be effectively combined with posteriors from MLPs focusing on short time chunks of 200 ms. The MLP focusing on medium term information takes nine consecutive frames of PLP features, as well as their first and second deltas as inputs. We will henceforth denote this as PLP/MLP. To extract long-term information, we use a variant of the Temporal Patterns (TRAPS) MLP architecture [3] called Hidden Activation TRAPS (HATS) [4]. The combined posterior goes through further transformation including log, PCA, and truncation in the way described in [2], and is then concatenated to the traditional features such as MFCC or PLP to form the augmented feature vector, which is passed to a GMM-HMM based speech recognition system. This approach builds on the so-called TANDEM approach first proposed in [5].

3. Scaling Up to More Training Data

For the Fall 2004 Rich Transcription (RT) evaluation, a vast amount of new training data became available in the form of the Fisher corpus (about 2000 hours of conversational speech). The challenge is how to effectively train neural nets on an order of magnitude more data than we used to deal with. It was shown in [6] that an optimal ratio of the total number of trainable parameters in an MLP to the total number of training examples is about 1:20. So with more data, we should use larger MLPs. Thus, the total amount of time for MLP training increases quadratically with the amount of training data, leading to an estimated training time of over one year for the entire available data. To speed up training time and yet keep the benefits of more data and more parameters, we adopted several modifications to our training recipe, as described in the next three sections.

3.1. Learning schedule modifications

We use an early stopping training schedule for our MLP training which prevents over-fitting. The basic procedure is to start training using a relatively large learning rate for each epoch (one epoch corresponds to processing every frame of the training set once). until error reduction on an independent cross-validation set drops below a fixed threshold. At this point, the learning rate is halved before each subsequent epoch, and the training stops when the error reduction on the cross-validation set drops below that fixed threshold. When examining our previous net trainings, we found that there were inefficiencies in this approach. First, we noticed that the epoch before the change of learning rate (often the 4th epoch) was never significantly reducing the error rate on the cross-validation set. That epoch only serves to mark the start of halving the learning rate for the following epochs. Second, we noticed that with more training data, fewer epochs were needed for convergence. For example, using 32 hours of training data per gender and 500K trainable weights per MLP (our initial configuration, known as "1x"), nine epochs

Table 1: Learning rate schedule and data rotation

Epoch Number	Tandem/PLP Learning Rate	HATS Merger Learning Rate	Data Used
1	0.001	0.0005	4x
2	0.001	0.0005	4x
3	0.001	0.0005	4x
4	0.0005	0.00025	8x
5	0.00025	0.000125	8x
6	0.000125	0.0000625	16x

are needed for training, while eight epochs are needed for a "2x" system which uses twice the training data and parameters, and seven epochs for a "4x" system. To train the "16x" nets (our final configuration), we also increase the training set size between epochs (see below), from 4x training data in the first few epochs to 16x data in the last epoch. With this knowledge, we roughly extrapolated that six epochs would be sufficient if we were to train up a "16x" system.

For the six epochs of training 16x nets, we use the following strategy and scheduling: The first three epochs are trained using 4x training data (128 hours per gender) with a higher learning rate, followed by two epochs of training with 8x training data (256 hours per gender) with half of the initial learning rate, further followed by an epoch of training with 16x data (512 hours per gender) with a quarter of the initial learning rate.

Furthermore, we noticed that the initial learning rate plays an important role in the training, and as we train with more data, smaller initial learning rates gave better results. By tuning with 1x, 2x, and 4x data and extrapolation on 16x data, we determined the initial learning rate for the Tandem/PLP net as 0.001, compared with 0.008 for the 1x net. Similarly, the initial learning rate for HATS merger net is set to 0.0005.

The training schedule for Tandem/PLP and HATS merger nets is summarized in the first three columns in Table 1.

3.2. Data rotation

Another modification to our training recipe was the use of nonoverlapping subsets of increasing amounts of training data for different epochs. From our experience, having better data coverage gave better results. Usually in MLP training, the same data are used in different epochs. When 16x data (512 hours per gender) are used for training the 16x nets, only less than half of the total available data (1200 hours per gender) are used. By using nonoverlapping data in training, the total amount of used training data can cover 4x + 8x + 16x of the data, thus comprising the majority of the available training set. Again, this scheme was first verified with 1x and 2x data. Since the HATS architecture is trained up in two stages where the first stage is parallelizable and relatively quick due to smaller critical band MLPs, we trained these critical band MLPs on the union of the 4x, 8x, and 16x subsets. The second stage merger MLP is trained using the schedule summarized in Table 1.

To create the training set, only the native speakers in the Fisher corpus were used. Waveforms were randomly selected to make the nonoverlapping 4x, 8x, and 16x datasets. Because the transcription quality of the Switchboard corpus is more reliable, we decided to use all Switchboard data in the 16x training set, half of it in the 8x training set, and a quarter in the 4x training set, which means the actual training sets in different epochs are not strictly non-overlapping. Still, the total coverage was 750 hours per gender for the combined Switchboard/Fisher training set.

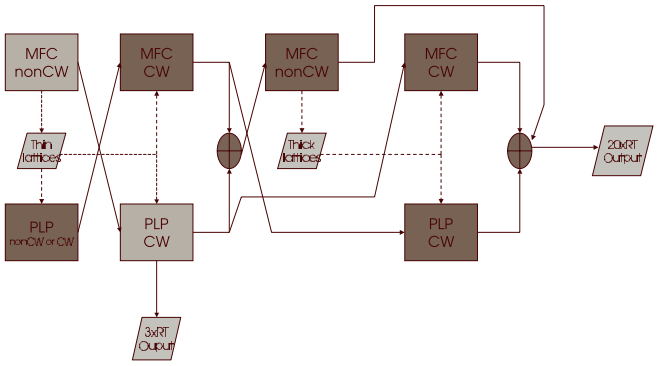

Figure 1: SRI CTS recognition system. Rectangles represent decoding steps. Parallelograms represent decoding output (lattices or 1-best hypotheses). Solid arrows denote passing of hypotheses for adaptation or output. Dashed lines denote generation or use of word lattices for decoding. Crossed ovals denote confusion network system combination. The two decoding steps in light gray can be run by themselves to obtain a "fast" system using about 3xRT runtime.

3.3. Software improvements

Finally, we also took advantage of software improvements. Chris Oei at ICSI optimized our MLP software, making use of Basic Linear Algebra Subroutine (BLAS) libraries that had been tuned for dual Intel 2.8 GHz Xeon hyperthreading CPUs. This resulted in throughput of 1500 to 2000 million connection updates per second (MCUPS), 3 to 4 times faster than before. About half the speed-up comes from BLAS libraries and half from hyperthreading. It still took 6 weeks on four computers to train the four gender-dependent PLP and HATS nets. Feature generation speed was measured at 0.5 times real time on a 3.4 GHz CPU.

4. System Architecture Development

The baseline for our work is the SRI CTS system as used in the Fall 2003 DARPA Rich Transcription evaluation and later refined for the Fall 2004 evaluation, as depicted in Figure 1. A detailed description of the system can be found in [7]; here we highlight its key aspects as relevant to the incorporation of MLP features. An "upper" (in the figure) tier of decoding steps is based on MFCC and voicing features [8]; a parallel "lower" tier of decoding steps uses PLP features [9]. The outputs from these two tiers are combined twice using word confusion networks (denoted by crossed ovals in the figure). Except for the initial decodings, the acoustic models are adapted to the output of a previous step from the respective other tier using MLLR (cross-adaptation). Lattices are generated initially to speed up subsequent decoding steps. The lattices are regenerated once later to improve their accuracy, after adapting to the outputs of the first combination step. The lattice generation steps use non-crossword (nonCW) triphone models, decoding from lattices uses crossword (CW) models. The final output is the result of a three-way system combination of MFCC-nonCW, MFCC-CW, and PLP-CW models. The entire system runs in under 20 times real time (20xRT). For many scenarios it is useful to use a "fast" subset of the full system consisting of just two decoding steps (the light-shaded boxes in the figure); this fast system runs in 3xRT and exercises all the key elements of the full system except for confusion network combination.

The baseline system structure is the result of a heuristic

optimization (which took place over several years) that aims to obtain maximal benefit from system combination and cross-adaptation, while staying within the 20xRT runtime constraint imposed by the DARPA STT evaluation. It was not feasible to redo this type of optimization from scratch using the new MLP features. We therefore decided to keep the overall processing structure and investigate systems that were obtained by replacing the features (and associated acoustic models) in the various decoding steps.

4.1. Data

For purposes of system optimization we used a version of the system and training data as was available at the time of the Fall 2003 RT evaluation. The corresponding baseline triphone acoustic models were trained on about 200 hours per gender, drawn from the LDC Switchboard and CallHome English corpora. All models were gender-dependent and trained using the minimum mutual information (MMI) criterion, on MFCC and PLP features respectively, after processing with cepstral mean and variance normalization, vocal tract length normalization (VTLN), heteroscedastic linear discriminant analysis (HLDA), and speaker-adaptive feature transformation (SAT, used in all but the first decoding step). The language model (LM) was a 4-gram trained on CTS transcripts as well as Broadcast News and conversational Web data [10], and was kept fixed for all experiments. No Fisher data was used in training this system.

Since the system design experiments were carried out in parallel with the development of large MLP training approaches (described in the previous section), we chose the largest MLPs available at the time for these experiments. These "4x" MLPs were trained on a 120-hour male-speaker subset of the acoustic CTS training set. A corresponding female MLP was not available, and thus all experiments were carried out on male-speaker test subsets. For MLLR purposes, we used a block-diagonal transform matrix that adapted the baseline and Tandem/HATs portions of the feature vector independently.

4.2. Early results

We initially tested the "4x" MLP features with PLP baseline models, using a single-stage bigram decoding and 4-gram rescoring system. Adding MLP features reduced the WER on the RT-02 test set from 30.5% to 28.4%, a 6.8% relative improvement.

In moving to a multistage, multi-front-end system, one of the first questions is whether a modeling improvement should be applied to all stages or just the final stage. The latter approach saves processing time, since the Gaussian computation is roughly proportional to the size of the feature vector, and our MLP features add 25 components to the feature vector, a 64% increase over the standard 39-dimensional baseline.

We tested the MLP features in various configurations in the fast, two-stage CTS system consisting of MFCC-nonCW decoding followed by PLP-CW decoding, with a baseline WER of 26.9%. When MLP features were used only at the PLP stage, the WER was reduced to 26.2%. When MLP features were also used to generate the adaptation hypotheses in the first stage, the result was improved to 26.0%. Finally, with MLP features added in the lattice generation run, the WER was 25.7%.

Not too surprisingly, it seems that it is important to incorporate MLP features early in the search to realize their full benefit. Note, however, that even under the best scenario, the overall improvement from MLP features is only 4.5% relative, compared to 6.9% in the one-stage system. This could be because cross-adaptation now occurs between two systems that share 40% of their feature vectors, which, while reducing each system's error rate individually, also makes their errors more correlated.

Table 2: Word error rate (WER) on RT-02 and RT-03 males using fast and full CTS systems

System	RT-02	RT-03
3xRT baseline	26.1	26.3
3xRT w/MLP features	24.8	25.5
20xRT baseline	23.7	24.6
20xRT w/MLP features	23.0	23.9
40xRT baseline w/MLP features	22.1	23.0
20xRT revised w/MLP features	22.8	23.6

4.3. Results with full systems

Based on the earlier results, we trained complete 20xRT CTS systems that use the Tandem/HATs MLP features in all acoustic models (MFCC and PLP, CW and nonCW), and compared performance to the baseline system using only the standard MFCC and PLP frontends. For completeness, the same comparison was done for the fast (3xRT) versions of the two systems. Since various parameters of the full system (such as the N-best rescoring weights) had been tuned on a subset of the RT-02 data we report results on both DARPA RT-02 and RT-03 evaluation sets (male speakers only, comprising 72 and 69 conversations sides, respectively).

The first four rows of Table 2 summarize the results from these experiments. We see that adding MLP features, when added to all models in the system, reduces WER by only about 2.8% relative, again showing diminishing returns as the system becomes more complex. As in the cross-adaptation experiment, we can attribute the loss in relative improvement to the fact that the two subsystems (MFCC and PLP-based, respectively) become more similar as both are augmented by the MLP features. Both cross-adaptation and the confusion-network combination in the full system would be negatively affected by this change.

To counteract the reduced effect of system combination we consider a new strategy: combining systems with and without MLP features, as well as those based on MFCC and PLP features. In our present setup, this can be achieved by running both the baseline system and the system with MLP features, and carrying out a final 6-way confusion network combination of all the models involved (MFCC-nonCW, MFCC-CW, PLP-CW, MFCC+MLP-nonCW, MFCC+MLP-CW, PLP+MLP-CW). The result is shown in the fifth row of Table 2: a 0.6% absolute WER reduction over the all-MLP system, resulting in a 6.5% relative gain over the baseline. Note that the relative improvement obtained is quite similar to that in our initial one-pass system. This suggests that the improvements from improved features can carry over to complex systems, provided that the system combination strategy embodied in the baseline is properly "expanded" to include the new features.

The drawback of the resulting system is of course that it no longer runs in 20xRT, thereby exceeding the stipulations for the DARPA RT-04F evaluation. Since we knew that a 3-way model combination could be accommodated in 20xRT, we looked for the best 3-way combination among the six subsystems available. This turned out to be the combination of MFCC+MLP-nonCW, MFCC+MLP-CW, and PLP-CW subsystems. As per Figure 1, this corresponds to a 20xRT system that uses MLP features in all its MFCC-based decoding stages, and unmodified PLP features in all other stages. Such a system is also desirable

Table 3: Word error rate (WER) on RT-04F development and evaluation sets

	RT-04F Dev			RT-04F Eval		
System	Male	Female	All	Male	Female	All
Baseline	18.1	16.2	17.2	20.2	20.4	20.3
w/MLP feats.	16.8	14.2	15.5	19.0	17.7	18.3
Rel. change (%)	-7.2	-12.3	-9.9	-5.9	-13.2	-9.9

because MFCC+MLP features are used in the initial and final lattice generation stages, thus ensuring the best possible lattice accuracies. Overall results with the revised 20xRT system are shown in the last row of Table 2. The absolute WER reduction over the baseline is 1.0% on RT-03, or 4.1% relative. This was the structure adopted for the final evaluation system.

5. Evaluation System

For the RT-04F evaluation all models were retrained on the full set of available CTS training data. This included all data used previously, plus about 2000 hours from the new Fisher collection. We excluded all nonnative speakers from acoustic training, since the test set was known to contain only native speakers. To reduce overall training time for HMMs, the Fisher training set was split into two complementary halves such that each half contained data from all training conversations. MFCC and PLP models were then trained on the complementary halves. Early experiments showed that this incurred only a minimal performance degradation on a single model's accuracy (0.2% absolute), and the combined system was effectively trained on the entire training set, while almost halving the required training time.

Other general improvements to the baseline (and correspondingly to the MLP-based system) were as follows. Acoustic models were trained using the minimum phone error (MPE) criterion [11], rather than with MMI. Also, triphone models were clustered using a decision-tree-based, top-down procedure, rather than SRI's traditional bottom-up "genone" algorithm. The nonCW models in the first PLP decoding step were replaced by CW models, giving a small accuracy gain and eliminating one model set to be retrained. Finally, the language model was also updated by incorporating Fisher transcripts and new Web data in training.

Two systems were trained: a baseline using standard MFCC (plus voicing) and PLP features, and a contrast system that used MFCCs augmented with Tandem/HATs MLP features. The MLP features were trained on 1800 hours of CTS data as described in Section 3. The system with MLP features was also the primary system fielded by SRI in the RT-04F evaluation (modulo minor bug fixes). Both systems were tuned on the RT-04F CTS development set (72 conversations) and then tested on the RT-04F evaluation set (also 72 conversations).

Table 3 summarizes all results, split by gender. The overall relative WER reduction on both testsets is identical, 9.9% (2.0% absolute on the evaluation set). This improvement is considerably greater than those reported in Section 4, and can be attributed to the fact that the amount of MLP training data is now commensurate with the HMM training data. (Previous experiments used MLPs trained on only a portion of the HMM data.) However, we also observe that the improvement is almost twice as big for female speakers than for males. This imbalance needs further investigation and points to a possible improvement of the system (by improving accuracy specifically on male speakers.)

6. Conclusions

We have shown that Tandem/HATs features when added to standard MFCC and PLP frontends in evaluation-style STT systems can yield considerable accuracy improvements, giving about 10% relative WER reduction. Since the MLP training is not easily parallelized, we developed a number of engineering techniques to enable training on 1800 hours of speech in a reasonable time frame (about 6 weeks). Furthermore, our experiments showed that simply adding the features to all models in a multi-pass, multi-front-end recognition system gave only meager improvements. We found that it is critical to use the improved features in early recognition passes for generating lattices and adaptation hypotheses. On the other hand, it is better to not use the MLP features in at least some of the system components to maintain diversity for purposes of system combination.

7. Acknowledgments

We thank our colleagues at SRI, ICSI, UW, IDIAP, and Columbia for contributions to the recognition system and other valuable input. This research was funded by DARPA under grant MDA972-02-1-0024 and contract MDA972-02-C-0038. Distribution is unlimited. Any opinions expressed here are those of the authors and do not necessarily reflect the views of the funding agency.

8. References

[1] N. Morgan, B. Y. Chen, Q. Zhu, and A. Stolcke, "TRAPping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition," in *Proc. ICASSP*, Montreal, May 2004, vol. 1, pp. 536–539.

[2] Q. Zhu, B. Y. Chen, and N. Morgan, "On using MLP features in LVCSR," in *Proc. ICSLP*, S. H. Kim and D. H. Youn, Eds., Jeju, Korea, Oct. 2004.

[3] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. ICASSP*, Phoenix, AZ, Mar. 1999, vol. 2, pp. 289–292.

[4] B. Chen, S. Chang, and S. Sivadas, "Learning discriminative temporal patterns in speech: Development of novel TRAPS-like classifiers," in *Proc. EUROSPEECH*, P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, Eds., Aalborg, Denmark, Sept. 2001, vol. 1, pp. 429–432.

[5] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *Proc. ICASSP*, Istanbul, June 2000, vol. III, pp. 1635–1638.

[6] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *Proc. ICASSP*, Phoenix, AZ, Mar. 1999, vol. II, pp. 1013–1016.

[7] A. Stolcke et al., "SRI/ICSI/UW Fall 2004 conversational telephone speech-to-text system," DARPA RT-04F Workshop, Nov. 2004.

[8] M. Graciarena, H. Franco, J. Zheng, D. Vergyri, and A. Stolcke, "Voicing feature integration in SRI's Decipher LVCSR system," in *Proc. ICASSP*, Montreal, May 2004, vol. 1, pp. 921–924.

[9] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[10] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *Proc. HLT-NAACL 2003*, M. Hearst and M. Ostendorf, Eds., Edmonton, Alberta, Canada, Mar. 2003, vol. 2, pp. 7–9, Association for Computational Linguistics.

[11] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, Orlando, FL, May 2002, vol. 1, pp. 105–108.