INCORPORATING TANDEM/HATS MLP FEATURES INTO 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 (MLPs). 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. These features had previously been shown to give significant accuracy improvements for CTS recognition when used with modest amounts of training data and relatively simple recognition architectures. 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 1500 hours of speech. Second, we investigated which components of a multi-pass, 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 long-term 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, 3, 4] 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.

The paper is organized as follows. We first describe the augmented front-end features that form the basis of our work (Section 2), followed by the techniques developed to scale feature training to very large training corpora (Section 3). We then present a series of experiments aimed at optimizing the use of the features in the context of the overall recognition system using a moderate amount of training data (Section 4). Finally, we report results of full-scale systems on the Fall 2004 CTS evaluation set (Section 5).

2. MLP-BASED FRONT-END FEATURES

We have been developing features based on multilayered perceptron (MLP) derived posteriors. MLPs are trained by taking various snapshots of the time-frequency plane as input. The MLP posteriors can later be combined for higher accuracy. 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 shorter-duration chunks of 200 ms. The combined posterior goes through further transformation including log, PCA and truncation in the way described in [3], 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].

For both types of MLPs, the output targets are the 46 phones used in the SRI CTS recognition system. The MLP focusing on medium-term information takes 9 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 [6, 7] called Hidden Activation TRAPs (HATs) [8, 4]. HATs consists of two stages of MLPs. The first stage extracts phonetically discriminant information from 500 ms of critical band energies, while the second stage merges this information and produces phone posteriors. The phone posteriors from both systems are merged on a per-frame basis using a weighted average, where the weights are the inverse entropy of the phone posteriors coming from the corresponding system [9].

3. SCALING UP TO MORE TRAINING DATA

For the Fall 2004 Rich Transcription evaluation, a vast amount of new training data became available in the form of the Fisher Corpus (about 2000 hours of conversational speech). Up to this point,

we have been developing and scaling our approach using increasingly larger subsets of the approximately 400 hours of Switchboard as training data for our nets [1, 2, 3]. In these published results, we started with gender-dependent nets with 500K total parameters trained on 32 hours of speech. We progressively doubled the total number of parameters as well as the amount of training data up to 4 times the original and still found relative improvements (4%-9%) when augmenting the standard front-end feature. For the RT-04 evaluation, we planned to use a system with 16 times the original amount of data and net parameters. From our experience with scaling this approach to use larger amounts of data as well as all the later passes of the SRI recognizer, we were confident that this approach would continue to help when given more data with which to train. The challenge, however, was how to train neural nets on an order of magnitude more data. It was shown in [10] 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 have larger MLPs. Thus, the total amount of time for MLP training increases quadratically with the amount of training data. A back-of-the-envelope calculation of the amount of training time needed to train our nets on all of Fisher and Switchboard data came out to be more than one year. To shorten training time and yet maintain all the benefits of more data and more parameters, we adopted several modifications to our training recipe:

1. We modified the learning schedule for the nets.
2. We rotated the portions of the training data for each epoch of training.
3. We accelerated the training software by using architecture-specific libraries.

3.1. Learning schedule modifications

We use an early stopping training schedule for our MLP training that prevents over-fitting. The basic procedure is to start training using a relatively large learning rate for each epoch[1] 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, the total number of epochs needed decreases. For example, using 1x[2] training data and net size, 9 epochs are needed for training, while 8 epochs are needed for a 2x system that uses 2 times as much training data and 2 times as many parameters, and 7 epochs for a 4x system. To train the 16x nets, we also use training sets of incremental size among epochs, where we start from 4x training data in the first few epochs and later switched to 16x training data for the last epoch. With this knowledge, we roughly extrapolated that 6 epochs would be sufficient if we were to train a "16x" system.

[1] One complete epoch of training corresponds to having processed every frame of the training.

[2] 1x pronounced "one times" corresponds to 32 hours of training data per gender and 500K trainable weights per gender-dependent neural net

Table 1. Final CV frame accuracy with different initial learning rates for 2x MLPs

Initial Learning Rate	CV accuracy
0.016	66.83
0.008	67.78
0.004	67.97
0.002	68.08
0.001	67.69
0.0005	66.98

Table 2. Learning rate schedule and data rotation

Epoch Number	PLP/MLP Learning Rate	HATs Merger Learning Rate	Data Used
1	0.001	0.005	4x
2	0.001	0.005	4x
3	0.001	0.005	4x
4	0.0005	0.00025	8x
5	0.00025	0.000125	8x
6	0.000125	0.0000625	16x

For the 6 epochs for training 16x nets, we use the following strategy and scheduling: The first 3 epochs are trained using 4x training data (128 hours per gender) with a higher learning rate, followed by 2 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. An example of the relation between the initial learning rate and the frame accuracy on the cross-validation set (CV accuracy) is shown in Table 1 for a 2x net training with 2x data, where an initial learning rate of 0.002 gives the best CV accuracy. We further determined the optimal learning rate for 4x net with 4x data to be 0.001. Since we use 4x data for the initial 3 epochs in the 16x net training, we decided to use an initial learning rate of 0.001. With similar tuning, we determined the initial learning rate for the HATs merger net to be 0.0005.

The training schedule for PLP/MLP and HATs merger is summarized in the first three columns in Table 2.

3.2. Data rotation

A modification to our training recipe that we adopted 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 data, a major part of the available data from the Fisher and Switchboard Corpus.

Another intuition, suggested by early ICSI experiments in the late 1980s, was that early epochs, where the gradient descent error-back propagation algorithm made larger steps in parameter space,

required less data to get in the vicinity of a good minimum, but as the steps were getting finer in later epochs, more data would help the algorithm hone in on a good error minimum. This scheme was simulated and tested, on smaller-scaled 1x and 2x systems. A 2x net was first trained with 3 epochs with 1x data followed by a single epoch of training with 2x data. With the 2x data as a superset of the 1x data in training, the frame accuracy on an independent cross-validation (CV) set was 65.6%, while in the nonoverlapping case, the CV accuracy improved to 66%.

Since the HATs architecture is trained in two stages where the first stage is parallelizable and relatively quick because of 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 2.

To make the training set, only the native speakers in the Fisher Corpus are used. They are 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 the Switchboard data in the 16x training set, half of it in the 8x training set, and a quarter of it in the 4x training set, which means that the actual training sets in different epochs are not strictly nonoverlapping. Still, the total coverage was 750 hours per gender for the combined Switchboard/Fisher training of the neural networks.

3.3. Software upgrades

In addition to modifying the training schedule and employing data rotation, we took advantage of software upgrades. Chris Oei at ICSI rewrote sections of our neural net training software so that the linear algebra operations were optimized for our computer architectures. Using the Basic Linear Algebra Subroutines (BLAS) libraries, he compiled a version of our training software to utilize the Hyper Threading capabilities of our dual Intel Xeon CPUs. The current training speed with the new software is as high as 1500 to 2000 million connection updates per second (MCUPS), 3 to 4 times faster than our old software, where the improvement in speed comes roughly half from the BLAS libraries and half from Hyper Threading.

3.4. Feature computation overhead

With the increased speed for training, it took 6 weeks on four computers with dual Xeon 2.8G Hz CPUs to train four gender-dependent PLP/MLP and HATs nets. Feature generation speed is measured as 0.57x real time on a 3.0 GHz CPU. Generating the feature for the entire 2400 hours of Fisher and Switchboard took about 2 weeks at SRI, partially because of network bottlenecks.

4. SYSTEM ARCHITECTURE AND EXPERIMENTS

4.1. System architecture

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 [11]; 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 [12]; a parallel "lower" tier of decoding steps uses PLP features [13]. 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 noncrossword (nonCW) triphone models, and 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 the confusion network combination.

Table 3. Word error rate (WER) on RT-02 males using a one-pass and rescoring system.

Features	WER (bigram)	WER (4gram)
PLP only	35.2	30.5
PLP + Tandem/HATs	32.8	28.4
Relative change	-6.8%	-6.9%

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.

4.2. 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 SuperARV 4-gram [14] trained on CTS transcripts as well as Broadcast News and conversational web data [15], 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 MLPs were trained on a 120-hour male-speaker subset of the acoustic CTS training set. A corresponding female MLP was not available, 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.

As a point of reference for subsequent experiments, Table 3

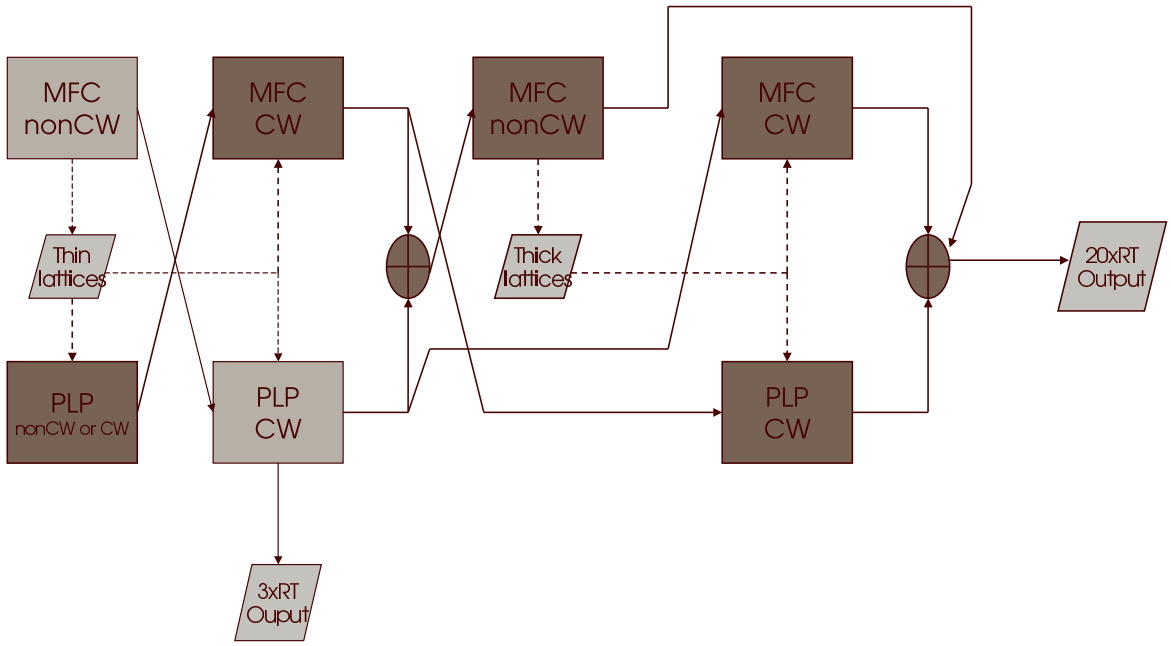

Fig. 1. SRI CTS recognition system. Rectangles represent decoding steps. Parallelograms represent decoding output (lattices or 1-best hypotheses). Solid arrows denote passing of hypotheses for adaptation or output. Dashed lines denote generation or use of word lattices for decoding. Crossed ovals denote 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.

shows results typical of our earlier work, using a simple one-pass bigram decoding and 4-gram LM rescoring system. The test set is the male portion of the RT-02 evaluation set (72 conversation sides). The baseline acoustic model uses PLP, compared with PLP augmented with MLP features. Both before and after LM rescoring the relative improvement obtained is 6.8% (or about 2% absolute).

4.3. Cross-adaptation and lattice decoding

A first question arising in any multistage system is whether a modeling improvement should be applied to all stages or just the final stage. The latter approach is attractive especially if the improvement is computationally more costly than the baseline approach. This is the case here, 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. The two stages interact in two ways: the MFCC step generates MLLR adaptation hypotheses for the PLP step, and the PLP decoding is constrained by lattices generated in the first step. We investigated the benefit of augmenting the MFCC models with the Tandem/HATs MLP features to generate MLLR reference transcriptions only, and to generate lattices.[3] To highlight the differences in acoustic models we omitted the final LM N-best rescoring that normally takes place on the PLP decoding output.

[3]Note that the same PLP-based Tandem features were used when augmenting both MFCC and PLP front ends.

Table 4. Word error rate (WER) on RT-02 males using a two-stage system with cross-adaptation and lattice decoding. The first stage uses MFCC models, and the second stage uses PLP models.

System	WER
Baseline (no MLP features)	26.9
MLP features in PLP models only	26.2
MLP features in MFCC and PLP models	
MLLR hyps only	26.0
MLLR hyps and lat. generation	25.7

The results are shown in Table 4. We observe that there is a substantial benefit to using the best features (MFCC + MLP) in all decoding passes, both to generate MLLR references and for lattice generation. This is somewhat surprising with regard to lattice generation, since the baseline lattices have a low oracle error rate of about 4%. It seems that in spite of such a low lattice error, search errors do occur in the second decoding pass, some of which can be prevented by using the improved features for lattice generation. Note that even under the best scenario, the overall improvement from MLP features is only 4.5% relative, compared to 6.9% in a one-stage system. This could be due to the fact that 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.

4.4. Results with full systems

Based on the results reported above, 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 per-

Table 5. 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

formance to the baseline system using only the standard MFCC and PLP front ends. 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 5 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 5: 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 current DARPA RT evaluations. We therefore proceeded to look for further revisions to our system architecture that would approximate the full benefit of the 6-way system combination within a 20xRT recognition framework. Since we know that a 3-way combination can be accommodated in the allowed runtime, we can ask which 3-way subset of the 6-way combination yields the most gain. A search over all 3-out-of-6 combinations showed that a combination of MFCC+MLP-nonCW, MFCC+MLP-CW, and PLP-CW subsystems yields the lower WER.

Accordingly, we ran a revised complete 20xRT system that used a combination of MFCC+MLP-nonCW, MFCC+MLP-CW and PLP-CW models in its final stage. As shown in Figure 1, this corresponds to a system that uses MLP features in all its MFCC-based decoding stages, and unmodified PLP features in all other stages. Such a system also has the desirable property that MFCC+MLP features are used in the initial and final lattice generation stages, thus ensuring the best possible lattice accuracies. The overall results with the revised 20xRT system are shown in the last row of Table 5. The absolute WER reduction over the baseline is 1.0% on RT-03, or 4.1% relative. This structure was then adopted for the final evaluation system.

5. EVALUATION SYSTEM RESULTS

5.1. New data and system update

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. To reduce overall training time for HMMs, the Fisher training set was split into two complementary halves so 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 minimal performance degradation on a single model's accuracy (0.2% absolute). 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 [16], 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.[4]

5.2. Results and Discussion

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 1500 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 6 summarizes all results, split by gender. The overall relative WER reduction on both test sets is identical, 9.9% (2.0% absolute on the evaluation set). This improvement is considerably greater than the gains reported in Section 4. The amount of MLP training data as a percentage of the total training corpus is about the same in the evaluation system as in the experiments reported before (about 60%). However, it could be the case that the MLP is better able to take advantage of the overall increase in data, and that therefore the system incorporating MLP features performs relatively better when given more data. The different scaling of MLP and HMM performance as a function of data, in turn, could be explained by the different scaling of the number of parameters. As

[4]The LM was a standard backoff 4-gram LM, rather than a SuperARV 4-gram as in preliminary experiments.

Table 6. 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
Baseline w/MLP features	16.8	14.2	15.5	19.0	17.7	18.3
Relative change	-7.2%	-12.3%	-9.9%	-5.9%	-13.2%	-9.9%

Table 7. Relative WER changes (in %) with MLP features, broken down by gender and processing stage.

	RT-04F Dev		RT-04F Eval	
System	Male	Female	Male	Female
Bigram decode	-8.5	-10.9	-10.0	-11.9
4-gram rescore	-8.8	-13.5	-8.1	-13.7
MLLR (cross-adaptation)	-5.4	-10.0	-5.3	-10.3
Final system combinations	-7.2	-12.3	-5.9	-13.2

described in Section 3, the number of MLP parameters was increased linearly with the amount of data. For the HMMs such a scaling would not have been practical given memory and runtime limitations; the number of HMM and Gaussian parameters was kept roughly constant.

A more puzzling observation concerns the relative gains achieved by the MLP features on male versus on female speakers: The relative WER reductions are about twice as big for female speakers as for males. To help us understand this phenomenon we tabulated the relative gains from MLP features at various points in the system, as shown in Table 7. The first row reflects the gain from MLP features after the initial bigram decoding (for lattice generation). When compared with the last row of Table 3 we can confirm that even for male speakers, the improvement is better than in the earlier experiments with less training data. Also, the MLP gains are fairly balanced across genders at this stage. However, the following rows of Table 7 show two diverging trends. For males, successive processing steps (and especially the application of the higher-order LM) seem to *reduce* the relative effectiveness of MLP features, whereas for females the trend is the opposite. This imbalance hints at a possible improvement of the system (e.g., by fixing a problem that might affect only the male speakers), and needs further investigation.

6. CONCLUSIONS

We have shown that Tandem/HATs features when added to standard MFCC and PLP front ends in evaluation-style STT systems, can yield considerable accuracy improvements, giving about 10% relative WER reduction. The additional features are estimated by multilayer perceptrons to maximize frame-level phone classification, and then appended to the standard feature vectors. Since the MLP training is not easily parallelized, we developed a number of engineering techniques to enable training on 1500 hours of speech in a reasonable time (about 6 weeks). Furthermore, our experiments showed that simply adding the features to all models in a multipass, 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 the purposes of system combination.

7. ACKNOWLEDGMENTS

This work was funded by DARPA under Contract MDA972-02-C-0038 and Grant MDA972-02-1-0024 (approved for public release; distribution is unlimited). We thank our colleagues at SRI, ICSI, and the University of Washington for their many contributions to the development of the STT system used in this research.

8. REFERENCES

[1] N. Morgan, B. Y. Chen, Q. Zhu, and A. Stolcke, "Scaling up: Learning large-scale recognition methods from small-scale recognition tasks", *in Proc. Special Workshop in Maui (SWIM)*, 2004.

[2] 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*, vol. 1, pp. 536–539, Montreal, May 2004.

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

[4] B. Y. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks", in S. H. Kim and D. H. Youn, editors, *Proc. ICSLP*, Jeju, Korea, Oct. 2004.

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

[6] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech", *in Proc. ICASSP*, Phoenix, AZ, Mar. 1999.

[7] H. Hermansky, S. Sharma, and P. Jain, "Data-derived nonlinear mapping for feature extraction in HMM", *in Proc. ICASSP*, Istanbul, June 2000.

[8] B. Chen, S. Chang, and S. Sivadas, "Learning discriminative temporal patterns in speech: Development of novel TRAPS-like classifiers", in P. Dalsgaard, B. Lindberg, H. Benner, and Z. Tan, editors, *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001.

[9] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR", *in Proc. ICASSP*, Hong Kong, Apr. 2003.

[10] 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.

[11] A. Stolcke et al., "Development of the SRI/ICSI/UW Fall 2004 conversational telephone speech-to-text system", these proceedings, 2004.

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

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, Apr. 1990.

[14] W. Wang, A. Stolcke, and M. P. Harper, "The use of a linguistically motivated language model in conversational speech recognition", *in Proc. ICASSP*, vol. 1, pp. 261–264, Montreal, May 2004.

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

[16] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training", *in Proc. ICASSP*, Orlando, FL, May 2002.