TONOTOPIC MULTI-LAYERED PERCEPTRON: A NEURAL NETWORK FOR LEARNING LONG-TERM TEMPORAL FEATURES FOR SPEECH RECOGNITION

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

[1]International Computer Science Institute, Berkeley, CA, USA
[2] University of California Berkeley, Berkeley, CA, USA
{byc, qifeng, morgan}@icsi.berkeley.edu

ABSTRACT

We have been reducing word error rates (WERs) on conversational telephone speech (CTS) tasks by capturing long-term (˜500ms) temporal information using multi-layered perceptrons (MLPs). In this paper we experiment with an MLP architecture called Tonotopic MLP (TMLP), incorporating two hidden layers. The first of these is tonotopically organized: for each critical band, there is a disjoint set of hidden units that use the long-term energy trajectory as the input. Thus, each of these subsets of hidden units learns to discriminate single band energy trajectory patterns. The rest of the layers are fully connected to their inputs. When used in combination with an intermediate-term (˜100ms) MLP system to augment standard PLP features, the TMLP reduces the WER on the 2001 Nist Hub-5 CTS evaluation set (Eval2001) by 8.87% relative. We show some practical advantages over our previous methods. We also report results from a series of experiments to determine the best ranges of hidden layer sizes and total parameters with respect to the number of training patterns for this task and architecture.

1. INTRODUCTION

Traditional feature extraction methods for automatic speech recognition (ASR), e.g. PLP and MFCC, compute features over a very small amount of time spanning the entire spectrum. Researchers have shown, using information theoretic analysis, that there is significant discriminant information about the identity of the current phone at times up to several hundred milliseconds away [1, 2]. This is the motivation for a family of long-term information extracting neural architectures. Neural TRAPS (mnemonic for "TempoRAl PatternS") learns discriminant long-term (˜500ms-1000ms), narrow frequency patterns for ASR [3, 4] using two stages of multi-layered perceptrons (MLPs). We improved upon the TRAPS architecture and developed "Hidden Activation TRAPS" (HATS) for conversational telephone speech (CTS) tasks. HATS (described in section 3) are effective as a complementary source of information to traditional short-term features. When we append a transformed version of the phonetic posterior outputs coming from HATS to PLP features, we can reduce the word error rate (WER) on the 2001 Nist Hub-5 CTS evaluation set (Eval2001) by 4.3%. In combination with an intermediate-term (˜100ms) MLP system (with MLP-based features derived from 9 frames of PLP), HATS can further reduce the WER by 7.53%.

In this paper we develop an MLP architecture with the same weight connections as HATS, but trained using a single run of error-back propagation. This eliminates the need to specify sub-frequency band level training targets, and consequently removes the need to store such targets for training (hence reducing disk requirements). We call this MLP architecture the Tonotopic Multi-Layered Perceptron (TMLP). We will show how the TMLP performs better than either HATS or a simpler unconstrained MLP-based method, achieving an 8.87% relative reduction in WER on Eval2001. We then explore the various configurations of the TMLP to find: 1) the optimal number of first layer hidden units; 2) the relationship between the optimal number of first layer hidden units, the total size of the TMLP, and the amount of training data; and 3) what ratio of training frames to TMLP parameters produces the best accuracies given a fixed amount of training time.

2. TONOTOPIC MLP DESCRIPTION

Inspired by the tonotopic organization of the human peripheral auditory system, where different positions in the cochlea are sensitive to different frequencies, we developed the TMLP. As noted earlier, the first hidden layer of the TMLP is tonotopically organized into several sets of hidden units. Each of these sets is constrained to see inputs coming only from a single frequency band, and together, all of the sets span the frequency range of speech. The second hidden layer, as well as the output layer are fully connected with their previous layers. Figure 1 shows the structure of a TMLP.

We have been using log critical band energies as inputs to the TMLP. After computing the log critical band energies of speech every 10 milliseconds and normalizing the mean and variance over each utterance, we take 51 consecutive frames (˜500 ms) of these normalized energies as the input layer of the TMLP. The output of the ith first layer hidden unit for frame f is given by equation (1):

$$O_{layer1,i} \stackrel{\text{def}}{=} sig\left(\sum_{t=f-25}^{f+25} in_{freq(i),t} W_{layer1,t,i} + B_{layer1,i}\right) \quad (1)$$

where $sig(x)$ is the logistic sigmoid function. $in_{freq(i),t}$ is the tth frame of energy in the one and only one frequency band that the ith first layer hidden unit is constrained to see. $W_{layer1,t,i}$ and $B_{layer1,i}$ are the trainable weights and bias respectively for the ith unit.

The second layer of hidden units takes the outputs of all first layer hidden units as inputs. The output of the jth second layer hidden units is given by equation (2):

$$O_{layer2,j} \stackrel{\text{def}}{=} sig\left(\sum_{I} O_{layer1,i} W_{layer2,i,j} + B_{layer2,j}\right) \quad (2)$$

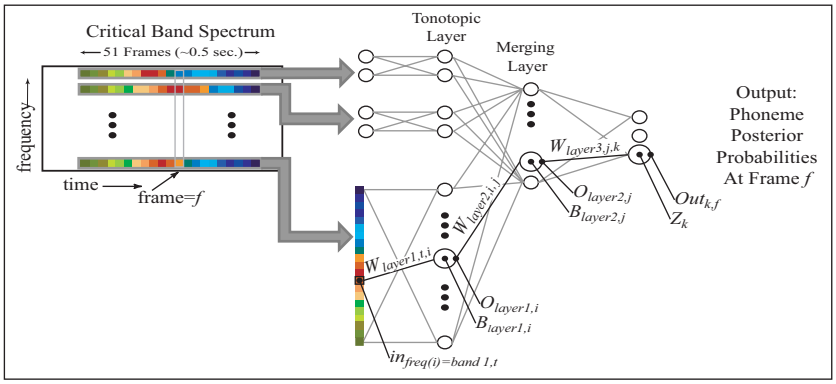

Fig. 1. *Tonotopic Multi-Layered Perceptron*

$W_{layer2,i,j}$ and $B_{layer2,j}$ are the trainable weights and bias respectively for the jth second layer hidden unit. Finally, the outputs of the TMLP are given by equation (3):

$$Out_{k,f} \stackrel{\text{def}}{=} \frac{\exp(Z_k)}{\sum_K \exp(Z_k)} \tag{3}$$

where Z_k is given by equation (4):

$$Z_k \stackrel{\text{def}}{=} sig\left(\sum_J O_{layer2,j} W_{layer3,j,k} + B_{layer3,k}\right) \tag{4}$$

$W_{layer3,j,k}$ and $B_{layer3,k}$ are the trainable weights and bias for the kth output unit.

As is typical for MLPs trained to estimate posteriors, the TMLP is trained with output targets that are "1.0" for the class associated with the current frame, and "0" for all others. For all of the systems described here, the MLPs are trained on 46 phoneme targets obtained via forced alignment from SRI's large vocabulary recognizer [5]. Also, all MLPs are trained to minimize cross entropy error by using the error back propagation algorithm. It is important to note that TMLP imposes a constraint upon the learning of temporal information from the time-frequency plane: correlations among individual frames of energies from different frequency bands are not directly modeled. Instead, the TMLP models correlation between long-term energy trajectories from different frequency bands.

3. COMPARISONS TO OTHER LONG-TERM SYSTEMS

We compare the performance of TMLP to two other MLP systems that learn long-term temporal information. In our previous work [6], we developed the Hidden Activation TRAPS (HATS) architecture. It is identical to TMLP with respect to the structure; i.e., the way each of the hidden units are connected is the same in HATS and TMLP. The difference arises in training. The HATS system learns the weight connections and biases in a two stage approach. First, we train single hidden layer MLPs for each critical band on the phone targets. Second, using the outputs of the hidden units of the first stage MLPs as inputs, we train another single hidden layer MLP on the phone targets. In contrast, the TMLP is trained in a single stage without specifying critical band level training targets. Since the structure of both TMLP and HATS are the same, they both impose the same learning constraint that forces the system to first learn discriminant narrow-frequency temporal patterns. Both the HATS and TMLP systems in this section use 40 hidden units per critical band and 750 hidden units at the merging layer giving about 500k total parameters. We compare HATS and TMLP to an unconstrained MLP that also has 500k parameters. This unconstrained MLP has a single hidden layer, and its inputs comes from 51 frames of log energies over all 15 critical bands. We refer to this long-term neural architecture as "15 Bands x 51 Frames". Also, for comparison purposes, we include results from our more traditional MLP which takes as inputs 9 frames of 12th order PLP plus energy, deltas, and double deltas. This MLP learns information about the speech over an intermediate-term scale (˜100ms), and we refer to this as "PLP 9 Frames".

3.1. Experimental Setup

We have seen that the best way to utilize the MLP posterior estimates as features has been to apply a series of transformations and concatenate them to the traditional PLP front-end features [7]. The transformations, designed to better fit the MLP outputs with Gaussian mixtures, require taking the log followed by principal component analysis (PCA) to orthogonalize and reduce the dimensionality from 46 to 25. The back-end that we used was similar to the first pass of the system described in [5], using a bigram language model and within-word triphone acoustic models. For these experiments we didn't use the multi-pass system which greatly improves WER, but in other work with HATS our improvements on the simpler system have largely carried over once later decoding passes and adaptation stages were employed.

The training set that we use for both MLP and HMM training consists of about 68 hours of conversational telephone speech data from four sources: English CallHome, Switchboard I with transcriptions from Mississippi State, and Switchboard Cellular. Training for both MLPs and HMMs was done separately for each gender, and the test results below reflect the overall performance on both genders. We hold out 10% of the training data as a cross validation set in MLP training. For fairness in comparison, all of the neural nets systems have roughly the same number of total network parameters (about 500k trainable parameters). The test results are on the 2001 Hub-5 evaluation data (Eval2001), a large vocabulary conversational telephone speech test set consisting of a total of 2,255,609 frames and 62,890 words.

System Description	WER (%)	Baseline Improv. (% Rel.)
Baseline: Non-Augmented HLDA(PLP+3d)	37.2	-
15 Bands x 51 Frames	36.6	1.61
HATS	35.6	4.30
TMLP	35.5	4.57
PLP 9 Frames	35.6	4.30
Inv Entropy Combo 15 Bands x 51 Frames + PLP 9 Frames	34.8	6.45
Inv Entropy Combo HATS + PLP 9 Frames	34.4	7.53
Inv Entropy Combo TMLP + PLP 9 Frames	33.9	8.87

Table 1. WER of Augmented Posterior Feature Systems on Eval2001

3.2. Results

Table 1 summarizes the WER of systems that use front-end features created by the concatenation of various transformed MLP outputs and baseline PLP features. The baseline PLP features are the heteroskedastic linear discriminant analysis (HLDA) transformed 12th order PLP plus energy and the first 3 deltas. All of the resulting features after concatenation are mean and variance normalized on a per conversation side basis. "HLDA(PLP+3d)" is the baseline feature system which gives a 37.2% word error rate. When we append the transformed outputs from "15 Bands x 51 Frames", WER drops by 1.61% relative. Appending the transformed HATS, TMLP, and "PLP 9 Frames" features causes relative drops in WER by 4.30%, 4.57%, and 4.30% respectively. Combining our MLP-based long-term (500 ms or more) features with intermediate term (100 ms) networks has yielded great improvements in the past, so using an inverse entropy weighted posterior combination technique to scale stream weights based on confidence [8], we combine "PLP 9 Frames" separately with "15 Bands x 51 Frames", HATS, and TMLP, apply log and PCA, and then append these to the baseline "HLDA(PLP+3d)" features. Using the combination of "PLP 9 Frames" with "15 Bands x 51 Frames" to augment the baseline features, we get a WER of 34.8% (see "Inv Entropy Combo 15 Bands x 51 Frames + PLP 9 Frames" in table 1). The combination of "PLP 9 Frames" with HATS and TMLP, give word error rates of 34.4% and 33.9% respectively.

3.3. Discussion

From table 1, we see that both the HATS and TMLP features outperform the "15 Bands x 51 Frames" features (1.61% improvement vs. 4.30% and 4.57%). One plausible explanation for this result is that the structural constraints imposed by HATS and TMLP on the learning of long-term critical band energies allow them to more efficiently utilize the number of trainable parameters. Comparing the TMLP features with the HATS features, TMLP features perform slightly better on word error rate, but the difference is not statistically significant. In terms of frame accuracy [1], however, the TMLP network performs at 68.2% accuracy on Eval2001, while the HATS network performs at 66.9%. By training all parameters in one overall back-propagation, TMLP is better able to maximize frame accuracy. One practical benefit to TMLP compared with HATS is that there is no need for storing large intermediate files that are created after the 1st stage training used as inputs for the 2nd stage training in HATS. These take ˜40 Gigabytes for this training set, and much more for the huge data sets that are now being used for CTS evaluations.

The "PLP 9 Frames" features also give a significant reduction in WER when used to augment the baseline "HLDA(PLP+3d)" features (4.30% relative). However, the information that it provides is complementary to the information provided by the neural nets that utilize longer-term inputs. We see this from the results that all combinations with the long-term systems yield greater relative improvement than just using "PLP 9 Frames" alone. Of all the combinations, combining with the long-term TMLP system works the best, adding another 4.57% relative improvement over using "PLP 9 Frames" alone.

4. AN EMPIRICAL STUDY OF TMLP

In this section we explore the relationships between the amount of training data, the total number of parameters, and the number of first hidden layer units per critical band in the TMLP. We created four different training sets for the TMLP. The first set consists of about 124.9 hours of conversational telephone speech data from: English CallHome, Switchboard I with transcriptions from Mississippi State, and Switchboard Cellular. Subsampling the 124.9 hour set by 2, 4, and 8 results in a 62.4 hour, 31.2 hour, and 15.6 hour training set. We trained TMLPs with 20, 30, 40, 50, and 60 hidden units per critical band (layer 1), and for each of these cases we chose the second hidden layer size such that the total number of parameters was either 250k , 500k, 1 million, or 2 million. Training for each TMLP setting was done separately for each gender, and the frame accuracies reflect the overall performance on both genders. For testing, we report frame accuracies on the Eval2001 test set as described in section 3.1. Figure 2 shows four graphs of frame accuracy on Eval2001 versus the number of hidden layer 1 units per critical band of TMLPs for the four different amounts of training data.

All curves in Figure 2 exhibit a max accuracy between 30 and 50 units except for the 1M parameters/15.6 hour case which has a max at 60. Only the 500k parameters/15.6 hour and 1M parameters/15.6 hour cases show trends that may indicate higher accuracies for greater than 60 hidden layer 1 units. A reasonable question to ask is whether the optimum number of hidden layer 1 units grows with more training data. From figure 2, we see that the opposite seems closer to the truth. Looking at all the lines for TMLPs with 500k, 1M, and 2M parameters, the best number of hidden layer 1 units seems to even decrease as the training data increases. However, it does appear to be the case that as the number of total parameters increases, the best number of hidden layer 1 units increases slightly. This can be seen clearly in the 62.4 and 124.9 hour panels. See how the best number of hidden layer 1 units go from 30 for 250k parameters to between 30-40 for 500k parameters, and to 40 for 1M and 2M parameters.

In a previous empirical study on training MLPs for use in a hybrid HMM/MMLP system on Broadcast News [9], we found

[1] Frame accuracy is the ratio of the number of correctly classified frames to the total number of frames, where classification is deemed correct when the highest output of the MLP corresponds to the correct phoneme label.

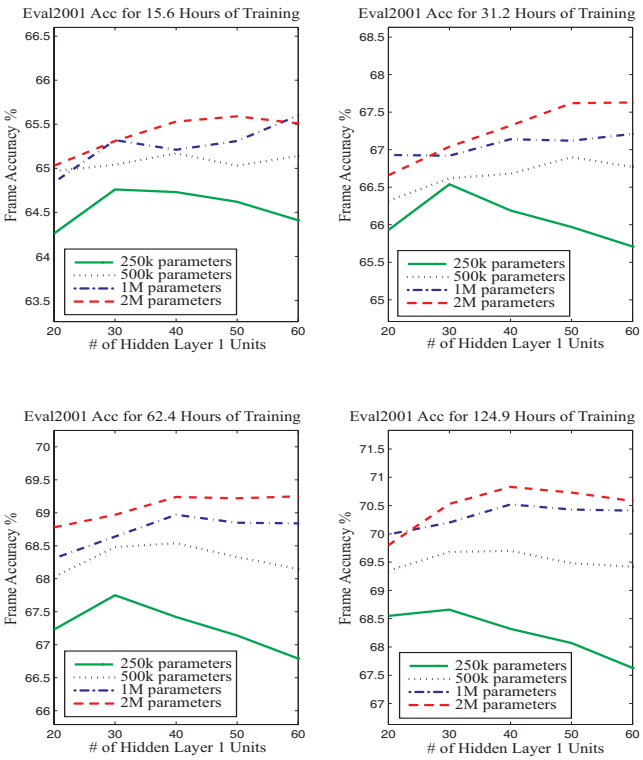

Fig. 2. *Frame Accuracy Results on Eval 2001 for Various TMLPs*

that the optimal ratio of number of training frames to number of parameters was in the range of 10 to 40 for a constant product of training frames and parameters, or equivalently the number of connection updates (CUPs) per complete epoch of training. The product of training frames and parameters gives a measure of how long it takes to train an MLP. We have plotted the average frame accuracies for TMLPs of constant CUPs versus the ratio of frames to parameters in figure 3. From this figure we can see a slowing of accuracy improvements as the ratio of frames per parameter increases. There is a decrease in accuracy for the 5 TeraCUP line when frames per parameter is greater than 20. 10 frames per parameter is definitely not the best ratio; however, it is unclear where the exact optimal ratio lies. It is somewhere between 20 and 80. It is interesting to note that the systems with 40 frames per parameter (i.e. 124.9 hours/500k parameters, 62.4 hours/500k parameters, and 62.4 hours/250k parameters) have 30-40 as the best number of hidden layer 1 units.

5. CONCLUSIONS

The tonotopic multi-layered perceptron (TMLP) is a competitive alternative to other long-term information capturing systems like TRAPS [3, 4] or hidden activation TRAPS (HATS). When used to complement traditional short and medium-term front-end features for the recognition of conversational telephone speech, TMLP achieves 8.87% relative WER reduction on Eval2001. This is slightly better than HATS and is a more practical technique in terms of the storage required for training. We also have studied relationships between the dimensionality required for the architecture, particularly in relation to major task parameters. In particular, we found that the empirically optimum number of critical

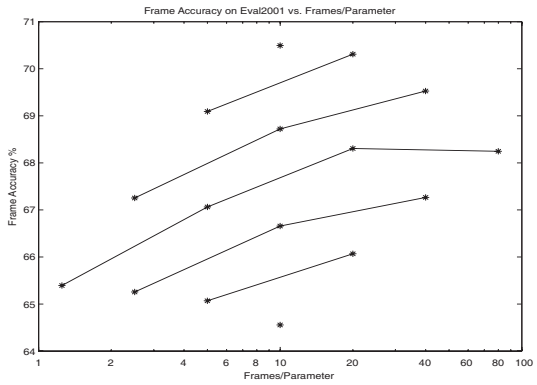

Fig. 3. *Frame Accuracy on Eval2001 For TMLPs of Equal Training Complexity*

band hidden units does not grow with increasing training data, but it slightly increases with an increase of parameters. We have also found that the optimal ratio of training frames to parameters is between 20 and 80 and that TMLPs trained in this range have best accuracies when the number of critical band hidden units is between 30-40.

6. ACKNOWLEDGEMENTS

We want to thank Andreas Stolcke for all his support and help in running the SRI recognition system, and Shawn Chang for his implementation of TMLP using ICSI's Quicknet software. Hynek Hermansky and his students, who pioneered this kind of long-term feature extraction, have always been extremely helpful. This work is supported by the DARPA EARS Novel Approaches Grant: No. MDA972-02-1-0024.

7. REFERENCES

[1] H. H. Yang, S. V. Vuuren, S. Sharma, and H. Hermansky, "Relavance of time-frequency features for phonetic and speaker-channel classification," *Speech Communications*, vol. 31, pp. 35–50, 2000.

[2] J. Bilmes, "Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling," in *Proc. ICASSP-1998*.

[3] H. Hermansky and S. Sharma, "TRAPs: Classifiers of TempoRAl Patterns," in *Proc. ICSLP-1998*.

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

[5] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[6] B. Y. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks," in *Proc. ICSLP-2004*.

[7] 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-2004*.

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

[9] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *Proc. ICASSP-1999*.