Learning Long-Term Temporal Features in LVCSR Using Neural Networks

Barry Chen, Qifeng Zhu, Nelson Morgan

International Computer Science Institute, Berkeley, CA, USA
{byc,qifeng,morgan}@icsi.berkeley.edu

Abstract

Incorporating long-term (500-1000 ms) temporal information using multi-layered perceptrons (MLPs) has improved performance on ASR tasks, especially when used to complement traditional short-term (25-100 ms) features. This paper further studies techniques for incorporating long-term temporal information in the acoustic model by presenting experiments showing: 1) that simply widening acoustic context by using more frames of full band speech energies as input to the MLP is suboptimal compared to a more constrained two-stage approach that first focuses on long-term temporal patterns in each critical band separately and then combines them, 2) that the best two-stage approach studied utilizes hidden activation values of MLPs trained on the log critical band energies (LCBEs) of 51 consecutive frames, and 3) that combining the best two-stage approach with conventional short-term features significantly reduces word error rates on the 2001 NIST Hub-5 conversational telephone speech (CTS) evaluation set with models trained using the Switchboard Corpus.

1. Introduction

Hynek Hermansky's group pioneered a method to capture long-term (500-1000 ms) information for phonetic classification using multi-layered perceptrons (MLP). Their approach learned temporal patterns based on consecutive frames of log critical band energies (LCBEs), and used these patterns as a basis for phonetic classification [1][2]. More specifically, they developed an MLP architecture called TRAPS, which stands for "TempoRAl PatternS". The TRAPS system consists of two stages of MLPs. In the first stage critical band MLPs learn phone probabilities posterior on the input, which is a set of consecutive frames (usually 51-100 frames) of LCBEs, or LCBE trajectory. A "merger" MLP merges the output of each of these individual critical band MLPs resulting in overall phone posteriors probabilities. This two-stage architecture imposes a constraint upon the learning of temporal information from the time-frequency plane: correlations among individual frames of LCBEs from different frequency bands are not directly modeled; instead, correlation among long-term LCBE trajectories from different frequency bands are modeled.

TRAPS by themselves perform about as well as more conventional ASR systems using short-term features, and significantly improve word error rates when used in combination with these short-term features. TRAPS complement conventional systems by performing better on speech examples that are problematic for models trained on conventional features. We worked on improving the TRAPS architecture in the context of TIMIT phoneme recognition [3]. This led us to the development of Hidden Activation TRAPS (HATS), which differ from TRAPS in that HATS use the hidden activations of the critical band MLPs instead of their outputs as inputs to the "merger" MLP.

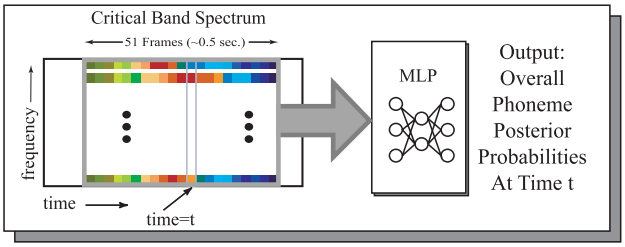

Figure 1: Architecture for Naive One Stage Approach

So instead of using critical band level phoneme probabilities, HATS uses outputs of critical band "matched filters" for inputs to the second stage merger. We found that HATS significantly outperformed TRAPS while using many fewer parameters.

In this paper, we wanted to further explore the incorporation of long-term features in the setting of large vocabulary continuous speech recognition (LVCSR). More specifically we want to explore two major questions: first, does the two-stage learning of HATS and TRAPS actually provide any advantage over a naive one-stage learning, where the latter consists of training an MLP to learn phone probabilities using 51 consecutive frames of LCBEs from all 15 critical bands in one step? And second, are the non-linear transformations of critical band trajectories, provided in different ways by HATS and TRAPS, actually necessary? For this second question, we compare linear and non-linear first stage critical band learning approaches and use these results as inputs to the second stage "merger" MLP.

We start this discussion with detailed architecture descriptions in Section 2 and experimental setup explanations in Section 3. In Section 4 frame accuracy results for each of the various long-term architectures are presented and discussed. Section 5 presents word recognition results using phone posterior features derived from the various long-term architectures, and Section 6 discusses results from using these posterior features in combination with a conventional short-term feature. Finally, Section 7 summarizes the conclusions.

2. MLP Architectures

2.1. One Stage Approach

A straightforward approach to incorporating greater temporal context is to give an MLP more frames of speech features and simply let the MLP learn what it needs to estimate phonetic posteriors. For our experiments, we chose this comparatively simple approach as the baseline architecture. In all of the experiments in this paper, we use LCBEs calculated every 10 ms on 8 kHz sampled speech which gives us a total of 15 bark scale spaced LCBEs. These are then mean and variance normalized per utterance. Figure 1 shows our baseline approach (hence-

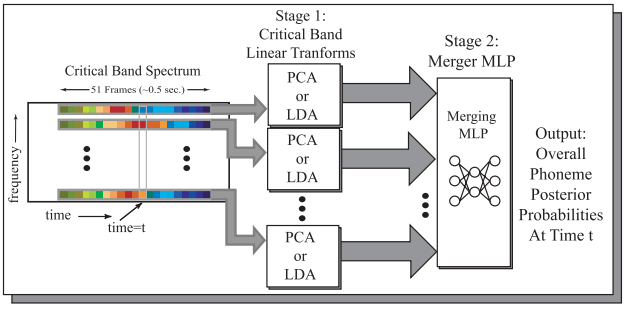

Figure 2: Architecture for Two Stage Linear Approaches

forth referred to as "15 Bands x 51 Frames") which uses 51 frames of all 15 bands of LCBEs as inputs to an MLP. These inputs are built by stacking 25 frames before and after the current frame to the current frame, and the target phoneme comes from the current frame. As is usual for this kind of use of MLPs, the network 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 in this paper, the MLPs are trained on 46 phoneme targets obtained via forced alignment from SRI's LVCSR recognizer [4], and consist of a single hidden layer with sigmoidal nonlinearity and an output layer with softmax nonlinearity.

2.2. Two Stage Linear Approaches

We hypothesize that the "15 Bands x 51 Frames" system is too unconstrained, and that it will be useful to design the learning so that the MLP is forced to represent temporal structure. We investigated several architectures that partition the learning into two constrained stages: first, learn what is important for phonetic classification given single critical band energy trajectories of 51 frames; and second, combine what was learned at each critical band to learn overall phonetic posteriors. This "divide and conquer" approach to learning splits the task into two smaller and possibly simpler sub-learning tasks.

For the first of these two-stage architectures, we calculate principal component analysis (PCA) transforms for successive 51 frames of each of the 15 individual 51 frames of LCBE resulting in a 51 x 51 transform matrix for each of the 15 bands. We then use this transform to orthogonalize the temporal trajectory in each band, retaining only the top 40 features per band. Figure 2 shows how we then use these transformed (and dimensionally reduced) features as input to an MLP. In a related approach, we replaced PCA with linear discriminant analysis (LDA) "trained" on the same phoneme targets used for MLP training. This transform projects the LCBE of a single band onto vectors that maximize the between class variance and minimize the within class variance for phoneme classes. These two two-stage linear approaches are henceforth denoted as "PCA40" and "LDA40" respectively.

2.3. Two Stage Non-Linear Approaches

Finally, we experimented with four two-stage non-linear approaches based on training critical band MLPs. Once trained, these critical band MLPs transform each of the 15 LCBE trajectories into input for a second stage merger MLP that combines and transforms this critical band information into estimates of phoneme posteriors conditioned on the entire spectrum. The two-stage MLP training approach is similar in spirit to the effective neural net training used in [5].

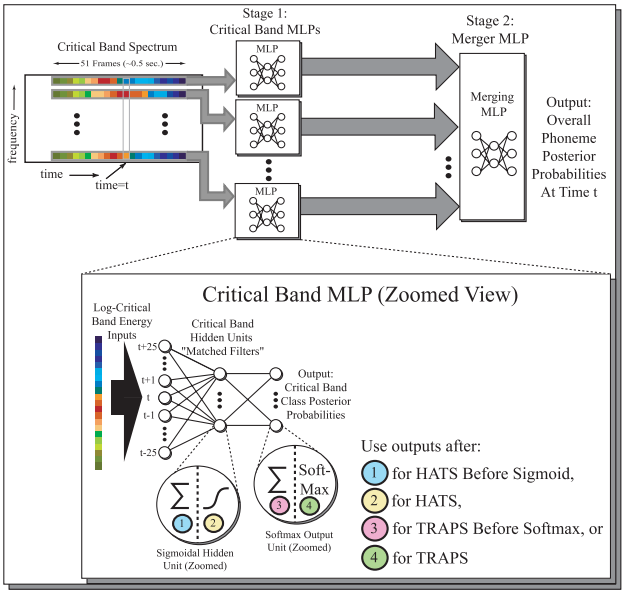

Figure 3: Architecture for Two Stage Non-Linear Approaches

Figure 3 shows each of our four non-linear two-stage architectures. In the first of these approaches, the input to the second stage is the dot product of the LCBE inputs with the input to hidden unit weights of the corresponding critical band MLP. Another way to say this is that the activation values before the sigmoid in each critical band hidden unit is used as the input to the second stage merger MLP. We refer to this architecture as "HATS Before Sigmoid". While this first approach consists of a linear matrix multiply, we categorize it in this subsection because the matrix was learned as part of a structure that included non-linear sigmoid functions, which have a significant effect on the values learned.

The second approach, "HATS", takes the outputs of each hidden unit as the input to the merger MLP. The third approach takes the activations after the hidden to output weight matrix multiplication, but just before the final softmax nonlinearity of the critical band MLPs. This approach is denoted as "TRAPS Before Softmax". Finally, the fourth approach uses the regular activations from the critical band MLPs that are phoneme posterior probabilities conditioned on the LCBE inputs. This last non-linear approach will be denoted as "TRAPS".

3. Experimental Setup

For all of the experiments reported in this paper, we show test results 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. 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. This training set corresponds to the one used in [6] without Switchboard Credit Card data. 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 long-term temporal

systems have roughly the same number of total network parameters (about 500,000 weights and biases). In preliminary experiments we found that forty hidden units per critical band for HATS was sufficient for good performance, so we made sure that all the two-stage systems had forty hidden units or in the case of PCA and LDA forty dimensions at the critical band level. The use of forty hidden units for the TRAPS system is not what has been previously reported when researchers refer to TRAPS, but for this study, we enforced this for fair comparison's sake [1].

Once the MLPs are trained, we use them to generate posterior features for an HMM back-end recognizer in a similar manner as was done in [7]. More specifically, the back-end that we used was similar to the first pass of the system described in [4], using a bigram language model and within-word triphone acoustic models. Further details on the posterior features used will be explained below.

4. Classification Accuracy Results and Discussion

In this section we examine the frame level classification of each of the various neural net architectures on the Eval2001 test set. Frame level accuracy serves as a good preliminary indicator of performance and 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. Table 1 summarizes the frame accuracy scores, relative improvement over baseline, and rank for each of the seven temporal architectures. For reference, we have included a conventional intermediate temporal context MLP that uses 9 frames of per-side normalized (mean, variance, and vocal tract length) PLP plus deltas and double deltas as inputs ("PLP 9 Frames"). This intermediate context MLP was trained on the same training data and phonetic targets as the others.

The one-stage 15 Bands x 51 Frames system serves as our naive baseline system and gets 64.73% of all frames correct. With the exception of the TRAPS system, all of the two-stage systems do better than this. From this, we can see that simply feeding an MLP classifier more frames for temporal context is suboptimal, but using the right two-stage approach is also important. HATS outperforms all other two-stage approaches at the frame level by achieving a 66.91% accuracy. HATS Before Sigmoid and TRAPS Before Softmax perform comparably at 65.80% and 65.85% respectively, while PCA and LDA approaches perform similarly at 65.50% and 65.52% respectively. At the frame level, it seems that forcing the system to focus first on learning what it can in each of the long-term narrow frequency band inputs independently is a useful constraint, particularly in the case of HATS.

5. MLP Based Feature Recognition Results and Discussion

Frame accuracy results give a good preliminary indication of performance, but can sometimes only be moderately correlated to word error rates. We performed word recognition experiments by transforming the outputs of the various MLPs and using them as features for the SRI speech recognizer. Specifically, in each case we take the log of the outputs from the MLPs and then decorrelate the features via PCA. Then we apply per-side mean and variance normalization on these transformed outputs and use the result as the front-end features in our HMM back-end. As in the previous section, we report the performance of the seven feature sets incorporating a long temporal input (500 ms), and include results for a moderate but more conventional input range (100ms for 9 frames of PLP). Table 2 summarizes the rank, word error rate (WER), and improvement over the one-stage baseline when appropriate.

System Description (Rank)	Frames Correct (%)	Baseline Improv. (% Rel.)
Baseline: 15 Bands x 51 Frames (6)	64.73	-
PCA40 (5)	65.50	1.19
LDA40 (4)	65.52	1.22
HATS Before Sigmoid (3)	65.80	1.65
HATS (1)	**66.91**	**3.35**
TRAPS Before Softmax (2)	65.85	1.73
TRAPS (7)	63.96	-1.19
PLP 9 Frames	67.57	N/A

Table 1: Frame Accuracies on Eval2001

Looking at tables 1 and 2, we can see that HATS always ranks 1 when compared to all other long temporal systems, achieving 3.35% and 7.29% relative improvement over the baseline one-stage approach in frame accuracy and WER respectively. The TRAPS (after softmax) doesn't provide an improvement over the baseline, but all of the other approaches do. This suggests that constraining the learning in the two-stage process can be helpful if the architecture is appropriate. The final softmax nonlinearity in the critical band MLPs in TRAPS is the only difference between it and TRAPS Before Softmax, so including this nonlinearity during recognition, causes significant performance degradation, though it is apparently critical to include it during training. It is likely that the softmax's output normalization is obscuring useful information that the second stage MLP needs. Since the HATS system significantly outperforms both the HATS Before Sigmoid and TRAPS Before Softmax systems, this means that the sigmoid nonlinearity is helpful whereas the extra linear mapping from hidden units to critical band phones is not. Another way to interpret this is that when using our two-stage learning approach, the best first-stage approach is to learn "probabilities" of certain critical band energy patterns. These "probabilities" correspond to the outputs of the hidden units of the critical band MLPs, and the patterns correspond to the energy trajectories represented by the input to hidden unit weights (essentially a non-linear form of matched filters).

6. Feature Augmentation Results and Discussion

Previous studies have shown time and again that systems learning temporal patterns perform reasonably well by themselves, but that in combination with the more conventional short-term full band features these temporal patterns provide significant additional performance improvements. Our current results also corroborate this previous finding. In the following experiments,

[1] It is possible that better results could be obtained by optimizing the number of critical band hidden units (or dimensions) for each of the different systems presented.

System Description (Rank)	WER (%)	Baseline Improv. (% Rel.)
Baseline: 15 Bands x 51 Frames (6)	48.0	-
PCA40 (2)	45.3	5.63
LDA40 (3)	46.5	3.13
HATS Before Sigmoid (4)	45.9	4.38
HATS (1)	**44.5**	**7.29**
TRAPS Before Softmax (4)	45.9	4.38
TRAPS (7)	48.2	-0.42
PLP 9 Frames	41.2	N/A

Table 2: WER of Systems Using Stand-Alone Posterior Features on Eval2001

System Description (Rank)	WER (%)	Baseline Improv. (% Rel.)
Baseline: Non-Augmented HLDA(PLP+3d)	37.2	-
15 Bands x 51 Frames (6)	37.1	0.27
PCA40 (2)	36.8	1.08
LDA40 (2)	36.8	1.08
HATS Before Sigmoid (2)	36.8	1.08
HATS (1)	**36.0**	**3.23**
TRAPS Before Softmax (5)	36.9	0.81
TRAPS (7)	37.2	0.00
PLP 9 Frames	36.1	2.96
Inv Entropy Combo HATS + PLP 9 Frames	34.0	8.60

Table 3: WER of Systems Using Augmented Posterior Features on Eval2001

we started with SRI's EARS Rich Transcription 2003 front-end features - 12th order PLP plus first three ordered deltas, per side mean, variance, and vocal tract length normalized, all transformed by heteroskedastic linear discriminant analysis (HLDA), keeping the top 39 features. Using these baseline features (HLDA(PLP+3d)), we performed a first pass viterbi recognition on Eval2001 and achieved a 37.2% word error rate (WER).

We then appended the top 25 dimensions after PCA on each of the temporal features described in section 5 to the baseline HLDA(PLP+3d) features. Table 3 summarizes the rank, WER, and relative improvement over the baseline HLDA(PLP+3d) features. All systems below the baseline system refer to the features that are appended to the baseline HLDA(PLP+3d) features.

When HATS features augment the conventional HLDA(PLP+3d) features, WER can be reduced by 3.23% relative, which is much better than the other long-term temporal methods tested. The one-stage approach and TRAPS lag the other two-stage approaches which perform roughly at the same level. Using the intermediate-term PLP 9 Frames system to augment HLDA(PLP+3d) features gives about the same performance as HATS. If we combine the posterior probability outputs of HATS and PLP 9 Frames systems using an inverse entropy weighting method [8], take the log followed by PCA to 25 dimension, and append to HLDA(PLP+3d) features, we get the "Inv Entropy Combo HATS+PLP 9 Frames" features. These features achieve a sizable 8.60% relative improvement over the HLDA(PLP+3d) features alone. This improvement is greater than the sum of the individual HATS augmentation and PLP 9 Frames augmentation. In combination, HATS and PLP 9 Frames features act synergistically to reduce WER.

7. Conclusions

We have compared several different approaches for incorporating long-term temporal information in MLP based front-end acoustic models and have shown that applying specific temporal constraints on the learning from time-frequency plane is important. More specifically, the one-stage approach, in which we feed 51 consecutive frames of 15 log critical band energies to the MLP, underperforms almost all two-stage approaches in which we have constrained the learning into two stages. The first of these stages extracts relevant information for phoneme classification within each long-term critical band energy trajectory, while the second stage combines what was learned in the first stage to produce the overall phoneme probabilities. We have also shown that it is important to use the hidden activations temporal patterns (HATS) as the two-stage approach. HATS significantly outperforms all other long-term temporal systems studied both as a standalone feature to an HMM back-end, but also when it is concatenated with conventional PLP features. Finally, HATS features combines synergistically with an intermediate time MLP features to achieve an 8.6% relative WER reduction on the Hub-5 2001 evaluation test set.

8. Acknowledgements

We want to thank Andreas Stolcke for all his support and help in running the SRI recognition system. This work is supported by the DARPA EARS Novel Approaches Grant: No. MDA972-02-1-0024.

9. References

[1] H. Hermansky, S. Sharma, "TRAPS - Classifiers of Temporal Patterns", in Proc. ICSLP 1998.

[2] H. Hermansky, S. Sharma, and P. Jain, "Data-Derived Non-Linear Mapping for Feature Extraction in HMM", in Proc. ICASSP 2000.

[3] B. Chen, S. Chang, and S. Sivadas, "Learning Discriminative Temporal Patterns in Speech: Development of Novel TRAPS-Like Classifiers", in Proc. Eurospeech 2003.

[4] 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 2000.

[5] C. Antoniou, "Modular Neural Networks Exploit Large Acoustic Context Through Broad-Class Posteriors for Continuous Speech Recognition", in Proc. ICASSP 2001.

[6] N. Morgan, B. Chen, Q. Zhu, and A. Stolcke, "TRAPping Conversational Speech: Extending TRAP/Tandem Approaches To Conversational Telephone Speech Recognition", in Proc. ICASSP 2004.

[7] D. P. W. Ellis, R. Singh, and S. Sivadas, "Tandem Acoustic Modeling in Large-Vocabulary Recognition", in Proc. ICASSP 2001.

[8] H. Misra, H. Bourlard, and V. Tyagi, "New Entropy Based Combination Rules In HMM/ANN Multi-Stream ASR", in Proc. ICASSP 2003.