Learning Discriminative Temporal Patterns in Speech: Development of Novel TRAPS-Like Classifiers

Barry Chen[1,2], Shuangyu Chang[2], Sunil Sivadas[3]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] University of California Berkeley, Berkeley, CA, USA
[3] OGI School of Science and Engineering, OHSU, Portland, Oregon, USA
{byc, shawnc}@icsi.berkeley.edu sunil@ece.ogi.edu

Abstract

Motivated by the temporal processing properties of human hearing, researchers have explored various methods to incorporate temporal and contextual information in ASR systems. One such approach, TempoRAl PatternS (TRAPS), takes temporal processing to the extreme and analyzes the energy pattern over long periods of time (500 ms to 1000 ms) within separate critical bands of speech. In this paper we extend the work on TRAPS by experimenting with two novel variants of TRAPS developed to address some shortcomings of the TRAPS classifiers. Both the Hidden Activation TRAPS (HATS) and Tonotopic Multi-Layer Perceptrons (TMLP) require 84% less parameters than TRAPS but can achieve significant phone recognition error reduction when tested on the TIMIT corpus under clean, reverberant, and several noise conditions. In addition, the TMLP performs training in a single stage and does not require critical band level training targets. Using these variants, we find that approximately 20 discriminative temporal patterns per critical band is sufficient for good recognition performance. In combination with a conventional PLP system, these TRAPS variants achieve significant additional performance improvements.

1. Introduction

While typical ASR systems use acoustic models trained on features extracted from spectral slices in time, a promising and complementary approach to acoustic modeling is TempoRAl PatternS or TRAPS [1][2][3]. Instead of extracting phonetic information from spectral slices in a short amount of time, as conventional ASR systems do, TRAPS extracts phonetic information from separate frequency channels (critical bands) spanning the full spectrum over a large amount of time (0.5 second to 1 second). Indeed, [4] shows that significant amounts of information exist at times greater than 100 ms away from the current time for phonetic classification. Using TRAPS in combination with conventional features, researchers have shown significant performance improvements in many conditions especially in the high noise conditions[1][3].

A TRAPS acoustic model as shown in Figure 1 consists of two stages of multi-layer perceptrons (MLPs). The first stage is a non-linear mapping from log critical band energy time trajectories to phonetic probabilities, and the second stage consists of another MLP that combines these critical band phonetic probabilities (one set per critical band) to obtain the overall phonetic probabilities. Let us focus our attention on the first stage of the TRAPS acoustic model. For each critical band, there is an MLP trained using the standard error back-propagation algorithm to learn phone posteriors by minimizing the cross-entropy
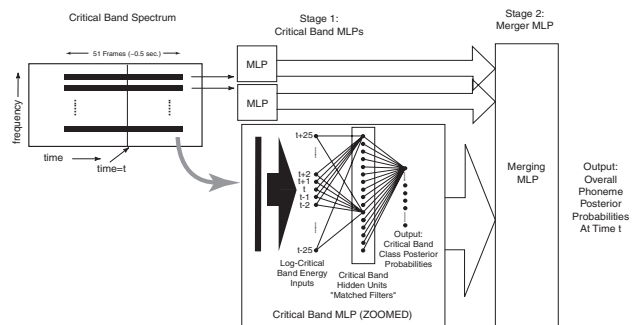

Figure 1: *TRAPS Acoustic Model*

between the network output and target distributions. Each net takes, as input, a half second (or a 1 second as in [1]) long log critical band energy temporal trajectory consisting of 51 frames (one frame per 10 ms calculated using a short-term FFT over 25 ms), and the training target is the phone label for the current frame. The input to hidden layer connections of these nets learn hyperplane separations in the input space of the 0.5 second long log critical band energy trajectories. Another way to look at it is that they learn matched filters useful for phonetic classification on the temporal evolution of the log critical band energy. In [1] these critical band MLPs learn 300 such matched filters for each critical band. The hidden to output layer of these critical band MLPs combine the outputs of the matched filters to form phone probabilities. The actual performance of these critical band MLPs on phonetic classification is actually quite low. The frame classification error rates of these nets range from 66% to 73% on TIMIT. This suggests that there is not enough information within a 0.5 second long log critical band trajectory to accurately classify a phone. This is not surprising considering that different phones may look quite similar within a narrow frequency band. Only after the second combining stage of the TRAPS model do we find low phone classification error rates since the model is able to combine the critical band phone posteriors to come up with the overall phone posteriors. Even though this conventional setup for TRAPS works well, we believe that further improvements are possible, so we consider two questions:

1.1. Can we skip the mapping from the outputs of the matched filters to critical band phone posteriors?

We have noted how the high frame classification error rates suggests that we cannot make all phone distinction given only a sin-
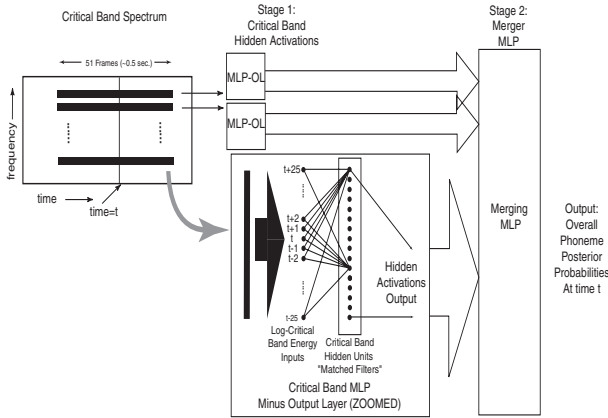
Figure 2: *Hidden Activation TRAPS (Note: MLP-OL stands for MLP minus the output layer)*

Figure 3: *Tonotopic Multi-Layer Perceptron*

gle critical band temporal trajectory. We hypothesize that whatever important phonetic information that can be gleaned from the critical band trajectory is already captured by the matched filters (critical band MLP input to hidden layer connections). The additional mapping from the matched filters to phone posteriors may be an extraneous and inaccurate mapping. Why not skip this intermediate mapping and instead use the outputs from the matched filters from every critical band as inputs for the second stage merger? In this way, we hope to find a more parsimonious model.

1.2. Is there a better way to train critical band matched filters?

Because training MLPs to learn phone posteriors from log critical band temporal trajectories is too difficult a task, what categories, instead of phones, should we train the first stage TRAPS models to learn? In [3] the critical band classifiers are trained to learn six broad categories based on manner of articulation. One can also imagine training the critical band classifiers to other linguistic feature-like classes that can be better distinguished at the critical band level; however, it would be better to learn what categories are important from data. Furthermore, any training labels that we can specify at the sub-band level based on full-band phonetic labels may be inaccurate because of potential asynchrony among the sub-bands [5]. We experiment with a new model for TRAPS which consists of a single 4-layer neural network whose architecture resembles TRAPS and whose training procedure obviates the need to specify critical band categorical targets - the log critical band matched filters are learned automatically from the data without specifying critical band level labels.

2. A Parsimonious Model: Hidden Activation TRAPS (HATS)

We have developed a version of the TRAPS acoustic model which we call Hidden Activation TRAPS (HATS). The architecture and training of HATS is very similar to conventional TRAPS. The training procedure of the first stage critical band MLPs is identical to that of TRAPS. Once the critical band MLPs are trained, we "chop off" the hidden to output layer of every critical band MLP, leaving only the output ("activations") of the hidden layer (hence, Hidden Activation TRAPS). After
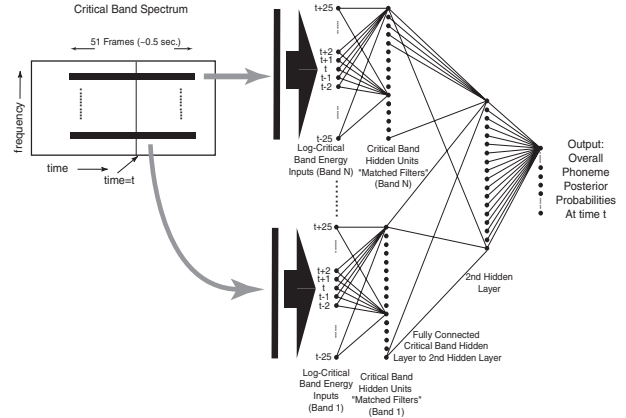
error back-propagation training, one can interpret these hidden layer activations as the outputs of discriminatively trained critical band matched filters. The second stage of HATS is just like TRAPS: a merger MLP (trained using the same training set and cross-validation set as in the first stage training) takes the hidden activations from all the critical bands and learns the mapping to phone posteriors. The HATS setup is shown in Figure 2.

It may seem that we don't gain anything from this HATS approach except reducing the number of parameters via the chopping off procedure, but we can significantly further reduce the number of parameters by reducing the number of matched filters required per critical band. In conventional TRAPS this number was set to 300 per critical band. To determine an optimal number of matched filters for HATS, we trained several HATS models that differ only in the number of matched filters per critical band. For fair comparison, we kept the total number of parameters constant (about 160,000 total neural net weights and biases). The frame classification accuracy of these HATS models on our cross-validation set has an optimal performance peak at 20 matched filters per critical band.

3. One Stage Training: Tonotopic Multi-Layered Perceptron(TMLP)

We have also created a 4 layer MLP that trains the critical band matched filters without the need for specifying critical band level targets. The first hidden layer consists of several groups of hidden units, each of which is constrained to receive the log energy inputs from only a single critical band. Note that these critical band hidden units correspond to the discriminatively trained matched filters as in the HATS system. The second hidden layer and the output layer combines the outputs of the matched filters across all frequencies to obtain phone posterior estimates. We call this model Tonotopic Multi-Layered Perceptron (TMLP) because the first hidden layer is arranged according to frequency channels. Figure 3 shows the architecture of TMLP. This is similar to HATS except that in this case the training of the first hidden layer happens as a part of the overall error back propagation training algorithm. This obviates the need to specify any kind of critical band training targets since the one stage training learns what is important implicitly.

Before continuing with the results of our experiments on HATS and TMLP, we take a brief detour to explain our experimental setup.

4. Experimental Setup

We use the TIMIT database for our experimental work. Using the recommended training set consisting of 3696 utterances, we set aside 3326 utterances for training our various MLPs and 370 utterances for cross validation. The cross validation set is used for adjusting the learning rate during MLP training and also for determining the early stopping point to prevent over-fitting.

In our experiments, we use the hybrid ANN/HMM speech recognition framework [6]. The artificial neural nets estimate phone posteriors. These posteriors are then scaled by the phone priors to produce the scaled likelihoods needed for the HMM back-end which is the Chronos decoder [7]. We also use a standard phone bigram language model during decoding.

Each of the various neural nets is trained to learn the original 61 TIMIT phones. The best phone sequence decoded by Chronos is at first a sequence of these 61 phones. In many previous studies using TIMIT, researchers map these 61 phones into a smaller set of 39 phones [8] and report their results using this smaller phone set. We perform the same mapping on the decoded output and then use SCLITE to obtain our phone error rates(%substitutions+%deletions+%insertions) on the complete TIMIT test set consisting of 1344 utterances and 51664 total phones.

In the following sections we present results in clean condition as well as in noisy and reverberant conditions. Please note, however, that all training was done using clean speech. We have experimented with two noisy conditions: Mercedes Benz noise (recorded inside the car) and exhibition hall noise (containing mainly speech babble). The noise files come from the Wall Street Journal Task for the AURORA2 evaluations. We add these noises to the clean files at different signal to noise ratios. We also convolve the clean signals with a room impulse response, corresponding to a 60 dB reverberation time of 0.8 seconds.

The features fed to our various TRAPS-like acoustic models are calculated from the clean, noisy and reverberant speech waveforms. These features are log critical band energies [9] calculated for every critical band and for each frame every 10 ms. The mean and standard deviation of the energies from each critical band are calculated and subtracted (divided in the case of standard deviation) on a per utterance basis. 51 consecutive frames of the log energies from each critical band forms the input features for our systems at the time corresponding to the 26th frame.

5. Results

We trained and tested four systems: a TRAPS baseline system, a HATS, a TMLP system, and a conventional PLP baseline system on the TIMIT training set using the original hand-labeled phones consisting of 61 distinct phones. The baseline TRAPS system is similar to the one presented in [1]. This TRAPS baseline system has 300 hidden units per critical band MLP and a merger MLP with 317 hidden units for a total of 1,032,377 parameters. The HATS system has 20 hidden units per critical band and also 317 hidden units for the merger. The total number of parameters for the HATS system is 159,935. The TMLP system also contains 20 hidden units per critical band, 317 hidden units for the merger and has the same number of parameters as the HATS system. Finally, the PLP system uses 12th order PLP [9] plus energy and first and second derivatives. These features undergo a per utterance mean and variance normalization and are then fed to an MLP with 9 frames of input context which

estimates the phone posteriors and contains roughly 160,000 parameters also. We stress that TRAPS, HATS, and TMLP differ significantly from this conventional PLP system because they focus on long narrow frequency patterns rather than the shorter spectral slices in PLP. Table 1 shows the phone recognition error rates of these four systems for the clean, reverberant, Mercedes Benz noise, and exhibition hall noise conditions

In addition to these four systems by themselves, we have some initial combination experiments using frame-wise posterior multiplication. This is the simplest combination technique in the hybrid ANN/HMM framework and has worked well in past studies. For each phone, the posterior probabilities from two different systems are simply multiplied and scaled by the square of the prior of the phone. Table 2 shows the phone recognition error rates for these combined systems

Test Condition	System Description			
	TRAPS	HATS	TMLP	PLP
Clean	32.7%	29.8%	31.0%	**29.7%**
Reverberant	56.3%	**54.2%**	58.0%	59.2%
Benz Noise				
20 dB	35.9%	**33.8%**	35.5%	36.5%
10 dB	42.7%	**42.2%**	42.8%	**42.2%**
0 dB	55.0%	56.7%	54.2%	**50.5%**
Exhib. Noise				
20 dB	41.6%	**39.9%**	41.8%	40.4%
10 dB	61.4%	63.4%	62.0%	**60.0%**
0 dB	102.2%	95.7%	**86.5%**	95.9%

Table 1: Phone error rates of the four systems on the full TIMIT test set under various conditions

Test Condition	Combination System		
	PLP+TRAPS	PLP+HATS	PLP+TMLP
Clean	27.2%	**26.5%**	26.8%
Reverberant	52.9%	**52.4%**	54.1%
Benz Noise			
20 dB	30.9%	**30.7%**	30.9%
10 dB	**35.9%**	36.2%	36.3%
0 dB	**44.9%**	45.8%	**44.9%**
Exhib. Noise			
20 dB	36.2%	**35.8%**	36.5%
10 dB	**54.4%**	55.7%	55.6%
0 dB	79.9%	**65.8%**	81.3%

Table 2: Phone error rates of the combined systems on full TIMIT test set under various conditions

6. Discussion

The three different variants of TRAPS show similar overall performance characteristics; however, HATS performs better than both TRAPS and TMLP in the case of clean or low noise conditions and also in the reverberant condition. In high noise conditions, both TRAPS and TMLP show more robustness than HATS, but in high exhibition hall noise TMLP outperforms all other systems. When compared with the conventional PLP system, these TRAPS variants perform better in reverberation. PLP is much better in very noisy car condition, while much worse in
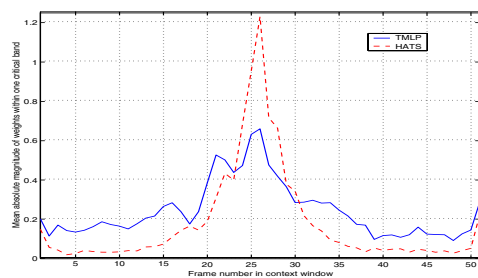
Figure 4: Comparison between HATS and TMLP for the average absolute magnitude of input to first hidden layer weights within a particular critical band (band 11, 1777 Hz). The center of the temporal window is at frame 26.

very noisy exhibition hall condition suggesting that combination of the TRAPS variants with PLP would show further gains.

The preliminary simple combination results show much promise. All combination error rates are significantly less than the corresponding error rates for the individual systems, and the clean results are close to the best published TIMIT phone recognition error rate that we are aware of. The PLP+HATS combination error rate on clean, 26.5%, is slightly greater than the best published TIMIT phone recognition error rate of 24.2% in [10].

An important result to note is that using only 20 discriminative patters per critical band in the HATS and TMLP systems, we can achieve comparable phone recognition performance to that of TRAPS which uses 300 discriminative patterns per critical band. The HATS and TRAPS systems both have 84% fewer parameters than TRAPS. From our results using the HATS system, we conclude that dropping the intermediate mapping from hidden unit activations to critical band phone posteriors, especially in clean conditions or in the presence of modest noise, is a good idea.

When training the discriminative critical band matched filters as part of the overall error back-propagation algorithm in the TMLP, we see similar performance to that of TRAPS. There does seem to be some advantage to TMLP in the highly noisy conditions (0 dB). Further analysis reveals a difference in the average temporal extent between the matched filters learned in the two-stage (HATS or TRAPS) and one-stage training (TMLP). In general, the matched filters learned in the two-stage training tend to have a much narrower temporal extent than that learned in the one-stage training. Figure 4 shows a comparison of the average absolute magnitude of input to first hidden layer weights within a particular critical band (band 11, 1777 Hz) between HATS and TMLP. It shows that TMLP matched filters are temporally broader and less sharply concentrated at the current frame than their HATS counterpart. Similar patterns are observed across all critical bands. It is not entirely clear why there is such a difference or what implication it has on the network performance. Perhaps the superior performance of TMLP in severe noise conditions (cf. Table 1) over HATS is due to the broad shape of these matched filters.

7. Conclusions

In this paper we have found that approximately 20 discriminative temporal patterns per critical band is sufficient to perform TIMIT phone recognition. We have developed two new

variants to TRAPS: HATS and TMLP. Both drastically reduce the number of parameters required while improving the phone recognition performance under clean condition. Furthermore, the HATS system outperforms TRAPS under reverberant and moderate noise conditions, and the TMLP system outperforms HATS under severe noise conditions. In addition, TMLP can be trained in one stage without the need of specifying critical band training targets. In combination with a conventional PLP system, all three variants achieve better performance than any system alone.

These results are encouraging and we are currently extending this work by experimenting with other combination techniques. Within our HATS and TMLP framework, we are also exploring using unequal number of matched filters across critical bands, thus tailoring the number of discriminative patterns to the information contained in each critical band. Finally, we would also like to extend our experiments beyond TIMIT corpus to perform phone and word recognition on conversational telephone speech.

8. Acknowledgements

9. References

[1] Hermansky, H. and Sharma, S., "Temporal Patterns (TRAPS) in ASR of Noisy Speech", Proc. ICASSP 1999.

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

[3] Jain, P.; Hermansky, H.; Kingsbury B. "Distributed Speech Recognition Using Noise-Robust MFCC and TRAPS-Estimated Manner Features", Proc, ICSLP 2002.

[4] Yang, H.; Sharma, S.; Van Vuren, S.; Hermansky, H. "Relevance of Time Frequency Features for Phonetic and Speaker Channel Classification", Speech Communication, August 2000.

[5] Mirghafori, N. and Morgan, N., "Transmissions and Transitions: A Study of two Common Assumptions and Multiband ASR", Proc. ICASSP 1998.

[6] Bourlard, H. and Morgan, N. Connectionist Speech Recognition: A Hybrid Approach, Kluwer Academic Publishers, 1994.

[7] Robinson, T. and Christie, J., "Time-First Search for Large Vocabulary Speech Recognition", Proc. ICASSP 1998.

[8] Lee, K. F. and Hon, H. W. "Speaker-Independent Phoneme Recognition Using Hidden Markov Models", IEEE Transactions on Acoustic Speech, and Signal Processing, 37(12), pp. 1641-1648, November 1989.

[9] Hermansky, H., "Perceptual Linear Predictive (PLP) Analysis for Speech", The Journal of the Acoustical Society of America, 87:1738-1752, April 1990.

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