The Role of Neural Network Size in TRAP/HATS Feature Extraction

František Grézl*

Brno University of Technology, Speech@FIT, Czech Republic
grezl@fit.vutbr.cz

Abstract. We study the role of sizes of neural networks (NNs) in TRAP (TempoRAl Patterns) and HATS (Hidden Activation TRAPS architecture) probabilistic features extraction. The question of sufficient size of band NNs is linked with the question whether the Merger is able to compensate for lower accuracy of band NNs. For both architectures, the performance increases with increasing size of Merger NN. For TRAP architecture, it was observed, that increasing band NN size over some value has not further positive effect on final performance. The situation is different when HATS architecture is employed – increasing size of band NNs has mostly negative effect on final performance. This is caused by merger not being able to efficiently exploit the information hidden in its input with increased size. The solution is proposed in form of bottle-neck NN which allows for arbitrary size output.

1 Introduction

The neural network (NN) based features are gaining more and more importance in today's ASR systems. Their era started more than a decade ago by introducing the TANDEM approach [1], where outputs of one classifier were treated as features for the second classifier. The first classifier is a Neural Network (NN) (or a structure of several NNs) trained to classify phonetically motivated classes and its outputs are estimates of class probabilities. The second classifier is a standard GMM-HMM system. As probabilities do not have the desired Gaussian distribution, they were usually processed by logarithm and decorrelated by Principal Component Analysis. The resulting features are called probabilistic features.

The TRAP (TempoRAl Pattern) feature extraction was one of the first employed in TANDEM scheme [2,3]. TRAP features are derived from long temporal context (up to 1s) of primary features, mostly outputs of Mel-filter bank critical band energies (CRBE). The temporal evolution of energy in one critical band forms *TRAP vector*. This *TRAP vector* is converged into phoneme probability estimates by its own NN (band NN). This is done for all coefficients/bands. Outputs from all band NNs are concatenated into one vector and, after logarithm nonlinearity, form input to Merger NN.

* This work was partly supported by Technology Agency of the Czech Republic grant No. TA01011328, Czech Ministry of Education project No. MSM0021630528, Grant Agency of Czech Republic projects Nos. GP102/09/P635 and 102/08/0707, and by BUT FIT grant No. FIT-11-S-2.

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 315–322, 2011.

Merger NN combines all band-conditioned estimates into one final set of probability estimates.

The performance of probabilistic features in the ASR system is closely tied (although the direct relation does not exists) to the classification accuracy reached by NN during the training. Thus the improvements of probabilistic features were focused on reaching higher classification accuracy of the NN. In the context of TRAP, several proposals addressing different stages of the processing were made. Different ways of *TRAP vector* extraction were for example addressed in [4,5]. Processing of several *TRAP vectors* by one band-NN was evaluated in [6,7]. Different structures of band NNs and Merger were studied in [8]. Finally, one can always play with increasing the NNs size [9].

As there is always the question *"What happens, if you make the NN bigger?"* we would like to address the last point in our analysis. There are two kinds of NNs in TRAP processing - band NNs and Merger and so the analysis can be split into two parts:

- Changing the size of band NN and keeping Merger size constant can tell what classification accuracy can be reached by band NN and how it influences the final accuracy of the Merger. The minimum sufficient size of band NNs can be found in this way.
- Altering Mergers size while keeping the band NNs constant can show how the classification accuracy change having the same input, e.g. what is the maximum accuracy one can reach with given band-NNs accuracy.

Finally, it would be possible to tell to what extent is the merger able to compensate the lower classification accuracy of band NNs and to find optimal sizes of NNs in the architecture.

The effect of altering NNs sizes is not observed only on classification accuracies but also on Word Error Rate (WER) of Large Vocabulary Continuous Speech Recognition (LVCSR) systems on meeting speech recognition as defined by NIST RT'07 speech-to-text evaluations.

2 Probabilistic Features

Ideally, we would like such features, that have maximum mutual information between the feature vector $\mathbf{x}$ and the class Q_i they belong to. It has been shown, that maximizing the *aposteriori* probability of class maximizes also the mutual information $I(\mathbf{x}, Q_i)$, under the condition that all classes $Q_i, i = 1 \ldots K$ are equally likely [10].

An ideal feature extraction should be able to reduce the error to its theoretical limit given by Bayes' error [11]. For K class problem, the Bayes classifier compares aposteriori probabilities of vector $\mathbf{x} : p(\mathbf{x}|Q_i)$ for all classes and classifies $\mathbf{x}$ to the class with maximum aposteriori probability. Since aposteriori probabilities are not linearly independent, as

$$\sum_{i=1}^{K} p(\mathbf{x}|Q_i) = 1, \tag{1}$$

only $K - 1$ probabilistic features would be the ideal set of features which would give the Bayes' error.

To estimate class aposteriori probabilities, the discriminative connectionist model – artificial neural network (NN) – is used. This model learns the transform of the input vector $\mathbf{x}$ to aposteriori probability directly from the data.

The discriminative training of the model focuses on the boundary between the classes where the differences are magnified, whereas the details in the "middle" of the class are rather minimized. This transformation makes the resulting probabilistic features more separable. This issue was discussed in [1].

3 System Description

The recognition task is meeting speech recognition as defined by the NIST RT'07 STT evaluations. The independent head set microphone (IHM) condition with reference segmentation was used in our experiments.

The **Critical Band Energies** (CRBE) are computed from 25ms of speech every 10ms. The speech signal is sampled at 16 KHz and there are 23 filters in the filter-bank analysis. CRBE are subject to mean and variance normalization on speaker basis.

Post-processing of Mergers output consists of logarithm and Heteroscedastic Linear Discriminant Analysis (HLDA) decorrelation and dimensionality reduction to 30 dimensions. The HLDA treats every state of corresponding HMM model as class.

The **Recognition system** is based on AMI-LVCSR used in NIST RT'07 evaluation [12] which is quite complex system running in many passes. For these experiments, the process stopped after the first decoding pass and estimation of VTLN warping factor. The system was simplified by omitting the constrained MLLR adaptation and lattice generation followed by four-gram Language Model (LM) expansion. Full decoding using bi-gram LM was done instead. The LM scale factor and the word insertion penalty estimated on RT'05 were used here.

The **training set** consists of the complete NIST, ISL, AMI and ICSI meeting data – about 180 hours. The NN were trained on subset of 173 hours. The transcription for NN training were obtained by forced alignment of training data using enhanced PLP features [12].

The **features** used in recognition system are the post-processed outputs from Merger only. Although delta parameters or concatenation with cepstral features improve the performance, for the purpose of our analysis, it is better to use only outputs from the system under evaluation.

3.1 TRAP/HATS Neural Network Architectures

The concept of **TRAP architecture** was given in Sec. 1, a more detailed description follows. The length of TRAP vector is 51 frames which covers 0.5 second of speech signal. There are 23 band-NNs which are trained towards 45 phoneme classes including silence. All used NNs have three layers. The scheme is shown in Fig. 1.

The **Hidden Activation TRAPS architecture** (HATS) further improved the performance of resulting probabilistic features [13]. As the name suggests, the outputs of band NN hidden neurons (after sigmoid nonlinearity) are taken to create inputs for Merger. The logarithm between the first stage outputs and second stage inputs is omitted.

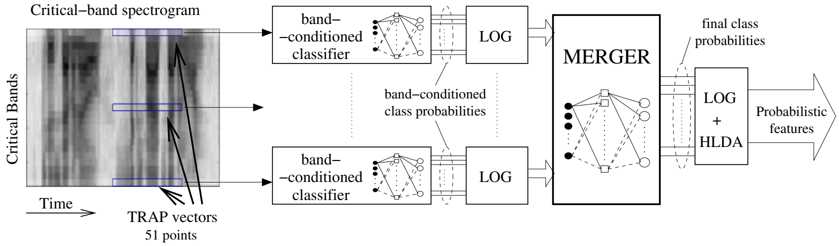

Fig. 1. Scheme of basic TRAP architecture

Table 1. Frame cross-validation accuracy of 6^{th} band NN [%]

Total band NNs weights	100 K	200 K	500 K	1 M	2 M
neurons in hidden layer	45	90	226	452	904
cross-validation acc	27.7	29.0	30.1	30.7	32.3

The numbers of weights assigned to band NNs (sum of weights in all band NNs) were 100K 200K 500K 1M and 2M. The numbers of weights in merger were 1M 1.5M 2M and 3M.

4 Experimental Results and Discussion

First, the frame accuracy on cross-validation data[1] of 6^{th} band NN is shown together with number of NN hidden units in Tab. 1. The classification accuracy increases with increasing number of weights in NN.

Next, Mergers with different numbers of weights are trained on each set of band NNs. The respective cross-validation accuracies are given in left part of Tab. 2. Then, the LVCSR system is trained on probabilistic features from each Merger and WERs are obtained. See right part of Tab. 2.

The following observations are made from these results:

- the system performance increases with increasing size of the Merger
- increasing band NNs size over 200 K weights does not further increase the performance
- the best system is not the biggest one

These observations suggest that higher classification accuracies of band NNs either cannot be utilized by the Merger, or are not necessary because the Merger is able to obtain the information by combining all band NNs outputs. In both cases, it would be interesting to find out where the band NNs improvements come from. We focused on this issue in the following section.

The HATS architecture was also evaluated. Note that the number of inputs to HATS Merger is changing with changing size of band NNs and thus the number of Mergers

[1] 10% of training data on which the NN is not trained which serves for measuring of improvements and early stopping of NN training.

Table 2. The Merger Cross-Validation frame accuracies and LVCSR WERs [%]

Cross-Validation frame accuracies [%]

band	Merger weights / hidden units			
NNs	1 M	1.5 M	2 M	3 M
weights	925	1388	1851	2777
100 K	63.3	64.1	64.5	65.5
200 K	64.2	65.1	65.6	66.3
500 K	64.0	65.6	65.5	**66.5**
1 M	64.6	65.5	66.0	66.3
2 M	64.5	65.2	65.7	66.2

LVCSR WER [%]

band	Merger weights / hidden units			
NNs	1 M	1.5 M	2 M	3 M
weights	925	1388	1851	2777
100 K	39.7	39.0	38.8	38.2
200 K	39.0	38.5	38.3	**37.7**
500 K	39.2	38.6	38.9	37.9
1 M	39.6	38.5	38.4	38.2
2 M	39.7	38.9	38.4	38.2

Table 3. The Merger Cross-Validation frame accuracies and LVCSR WERs [%]

Cross-Validation frame accuracies [%]

band NNs	Merger weights			
weights	1 M	1.5 M	2 M	3 M
100 K	65.1	66.1	66.7	**67.5**
200 K	63.9	65.2	66.0	67.0

LVCSR WER [%]

band NNs	Merger weights			
weights	1 M	1.5 M	2 M	3 M
100 K	37.6	37.4	36.8	36.7
200 K	38.7	37.5	37.1	**36.6**

hidden units is different for every experiment. The HATS Merger's hidden layer sizes are the same as for TRAP Merger's for 100 K weights in band NNs and roughly half for 200 K weights in band NNs. The HATS Merger cross-validation accuracies and WERs are given in Tab. 3.

The performance of HATS systems is also increasing with increasing size of Merger NN. On the other hand, increasing the size of band classifiers has negative effect on the performance of the whole system - only the architecture accommodating the largest Merger was able to provide comparable performance to a system with smaller band NNs.[2] This shows that the HATS Merger was overloaded by increased number of inputs. Although these inputs carry more information which was able to improve TRAP systems, it cannot be utilized by HATS Merger and, contrary, more inputs seems to bring larger confusion and impair the overall performance.

4.1 Detailed Analysis

This section is focused on the band NNs accuracies. Tab. 1 gives the classification accuracies of the 6^{th} band NN. The accuracies of band NNs in all bands are shown in Fig. 2. It can be seen, that the classification increases in all bands with increasing NN size, so stagnation in Mergers accuracy cannot be assigned to the degradation of NNs in other bands.

The following analysis was focused on classification accuracy of individual classes by one band NN – 6^{th} band was used. The cross-validation data were used for this analysis. The percentage of correctly classified frames per individual class are shown

[2] Further experiments with larger band NNs were not performed as stagnation was observed for TRAP architecture and degradation for HATS.

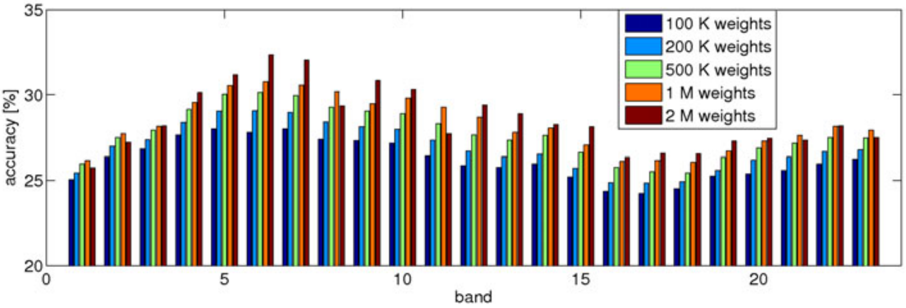

Fig. 2. Cross-validation accuracies [%] of all band NNs

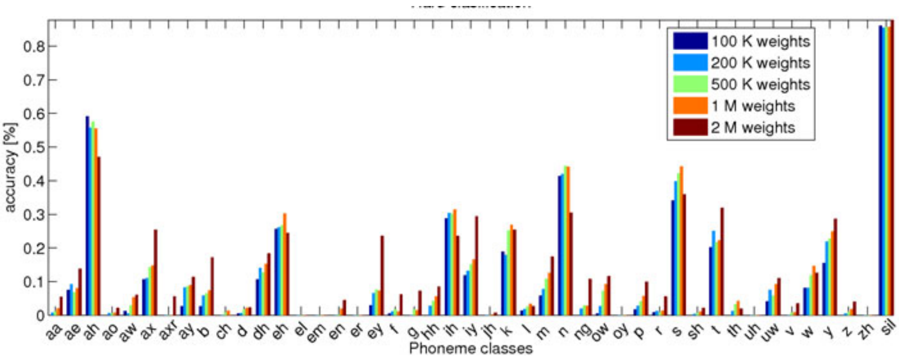

Fig. 3. Classification accuracies per individual class [%] – 6^{th} band

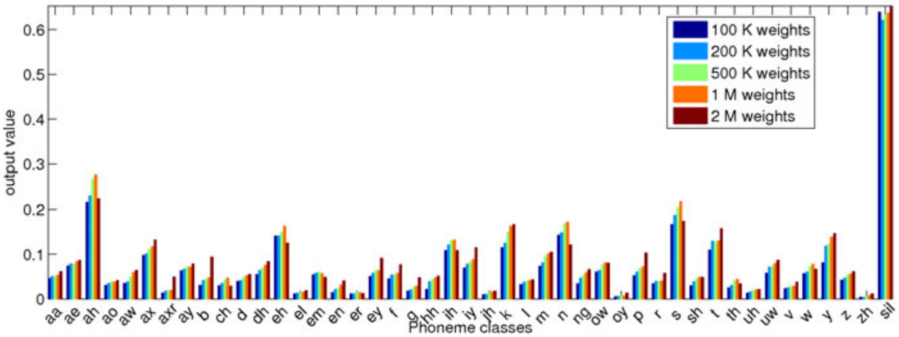

Fig. 4. Average value of corresponding NN output for given class [%] – 6^{th} band

in Fig. 3. In the next step, the average value of output corresponding to given class was computed over all input vectors labeled as given class, see Fig. 4.

For both analysis can be seen that increased size of band classifier increased classification accuracies/average output value for almost all classes. The exception are band

NNs of size 2M weights which seems to perform significantly better for some classes and have worse performance for others. Overall can be said that the patterns are about the same.

5 Conclusions

In this study we have investigated the role of NN size in TRAP/HATS probabilistic feature extraction scheme. This investigation covers both parts of the processing – band NNs and Merger. The band NNs creates some kind of filter which let only particular information to Merger. If the information is lost here, Merger will not be able to achieve high classification accuracy. Thus it is important to use band NNs of such size, which will preserve all necessary information. The Mergers task is to combine particular probability estimates into final ones. It thus has to have enough power to perform this task.

The results obtained on TRAP architecture shows that the system performance increases with both, band NN size and Merger size. But the enlargement of band NNs over some size does not have further positive effect, the performance saturates. The increased size of the merger can compensate for poor band NNs performance to some extent. But the cost in terms of used weights is much higher compared to what is added to band NNs. Over all, it can be said that having more parameters in band NNs does not hurt and leads to good system performance.

Unfortunately, this cannot be said for HATS architecture. Although much better performance was obtained by this architecture when band NNs with 100 K weights were used, increasing size of band NN did not improve the performance. Contrary, degradation was mostly seen and only the architecture with largest merger gained comparable results. Why the improvement seen for TRAP systems is not observed when HATS architecture is used instead? We know, that useful information is contained in activation outputs of larger band NNs, but giving it directly to Merger is not the right way to present it. The HATS system seems to benefit from compact information on Mergers input. From this point of view, the tuning of HATS system in [14] might be questionable, but the authors did not provide the NN sizes to give us a clue where their operation point is. It can be recommended to prefer smaller band NNs when designing the HATS architecture and to validate the designed architecture experimentally.

It would be beneficial to present compact information to the Merger regardless the size of band NNs. Such solution have been already proposed in the form of bottle-neck NN structure [15]. It effectively separates the output size from other parameters of the NN such as number of classes (which is fixed in TRAP architecture) and size of the NN. The possible problem with this approach lies in usage of five-layer NN. It might be difficult to train more complex NN on evolution of just one parameter (a single critical band energy) and also proper setting of sizes of all layers would be more complicated. Nevertheless, this approach seems to be another step in TRAP/HATS feature extraction techniques.

References

1. Hermansky, H., Ellis, D.P.W., Sharma, S.: Tandem connectionist feature extraction for conventional HMM systems. In: Proc. ICASSP 2000, Turkey (2000)
2. Sharma, S.R.: Multi-stream approach to robust speech recognition, Ph.D. thesis, Oregon Graduate Institute of Science and Technology (October 1999)

3. Hermansky, H., Sharma, S., Jain, P.: Data-derived nonlinear mapping for feature extraction in HMM. In: Proc. Workshop on Automatic Speech Recognition and Understanding, Keystone (December 1999)
4. Athineos, M., Hermansky, H., Ellis, D.P.W.: LP-TRAP: Linear predictive temporal patterns. In: Proc. ICSLP 2004, Jeju Island, KR, pp. 949–952 (October 2004)
5. Tyagi, V., Wellekens, C.: Fepstrum representation of speech signal. In: Proc. of IEEE ASRU, San Juan, Puerto Rico, pp. 44–49 (December 2005)
6. Jain, P., Hermansky, H.: Beyond a single critical-band in TRAP based ASR. In: Proc. Eurospeech 2003, Geneva, Switzerland, pp. 437–440 (2003)
7. Grézl, F., Hermansky, H.: Local averaging and differentiating of spectral plane for TRAP-based ASR. In: Proc. Eurospeech 2003, Geneva, Switzerland (2003)
8. Zhu, Q., Chen, B., Grézl, F., Morgan, N.: Improved MLP structures for data-driven feature extraction for ASR. In: Proc. INTERSPEECH 2005, Lisbon, Portugal (September 2005)
9. Ellis, D., Morgan, N.: Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition. In: Proc. ICASSP 1999, Phoenix, Arizona, USA, pp. 1013–1016 (March 1999)
10. Bourlard, H., Morgan, N.: Connectionist Speech Recognition: A Hybrid Approach. Kluwer International Series in Engineering and Computer Science, vol. 247. Kluwer Academic Publishers, Dordrecht (1994)
11. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press Professional, Inc., San Diego (1990)
12. Hain, T., et al.: The AMI system for the transcription of speech meetings. In: Proc. ICASSP 2007, Honolulu, Hawaii, USA, pp. 357–360 (April 2007)
13. Chen, B., Zhu, Q., Morgan, N.: Learning long-term temporal features in LVCSR using neural networks. In: Proc. ICSLP 2004, Jeju Island, KR (October 2004)
14. Zhu, Q., Stolcke, A., Chen, B., Morgan, N.: Using MLP features in SRI's conversational speech recognition system. In: Proc. INTERSPEECH 2005, Lisbon, Portugal (September 2005)
15. Grézl, F., Karafiát, M., Kontár, S., Černocký, J.: Probabilistic and bottle-neck features for LVCSR of meetings. In: Proc. ICASSP 2007, Honolulu, Hawaii, USA, pp. 757–760 (April 2007)