PROBABILISTIC AND BOTTLE-NECK FEATURES FOR LVCSR OF MEETINGS

František Grézl, Martin Karafiát, Stanislav Kontár and Jan Černocký

Speech@FIT group, Brno University of Technology, Czech Republic

{grezl,karafiat,cernocky}@fit.vutbr.cz

ABSTRACT

In recent years, probabilistic features became an integral part of state-of-the-are LVCSR systems. In this work, we are exploring the possibility of obtaining the features directly from neural net without the necessity of converting output probabilities to features suitable for subsequent GMM-HMM system. We experimented with 5-layer MLP with bottle-neck in the middle layer. After training such a neural net, we used outputs of the bottle-neck as features for GMM-HMM recognition system. The benefits are twofold: first, improvement was gained when these features are used instead of the probabilistic features, second, the size of the system was reduced, as only part of the neural net is used. The experiments were performed on meetings recognition task defined in NIST RT'05 evaluation.

Index Terms— Probabilistic features, bottle-neck features, TRAP-based features, LVCSR, meeting recognition.

1. INTRODUCTION

The probabilistic features – class probabilities converted to the form suitable for following GMM-HMM system – were introduced to speech recognition research in TANDEM feature extraction [1]. The class probabilities are usually estimated by a neural network and the classes are context-independent phonemes. The conversion of probabilities to probabilistic features is usually done by log followed by PCA de-correlation, possibly with dimensionality reduction.

TRAP processing [2] is a way to obtain probabilistic features. The novelty of this approach is in processing of temporal patterns (hence TRAP) of log-energy from each critical band independently. The temporal pattern is classified by a band-conditioned nonlinear classifier (multi-layered perceptron – MLP) into a phoneme class. The outputs from all band-conditioned classifiers are then merged. The final classification is done by a merging classifier.

The first stage — band-conditioned classifier — can be seen as a temporal feature extractor and the merging net as a classifier of these temporal features. The use of band-conditioned class probabilities as temporal features was studied in [3]. Authors examine the usefulness of intermediate products of band-conditioned neural net for overall classification and recognition performance. Good results were obtained with the hidden (middle) layer activation outputs. This technique is called Hidden Activation TRAPS – HATS.

In further research [4], the HATS neural net structure was created as one net – Tonotopic MLP. In this four-layered MLP, the first (input) and the second (first hidden) layers are not fully connected. The bands are processed independently and the following fully-connected layers combine information derived from each critical band to output class probabilities. In this case, the temporal features are hidden and the neural net derives them in the training process as a part of the overall training for maximization of classification accuracy.

The study in [5] pushes the extraction of temporal features out of the classifier. Instead of complex non-linear transform in form of neural net, a simple discrete cosine transform (DCT) is used on top of each critical band temporal trajectory. This transform, which was earlier proposed as pre-processing of temporal patterns for band-conditioned classifiers, provides sufficient information about the underlying temporal pattern and the classifier – a fully connected four-layer MLP – focuses only on the classification task.

The use of phoneme model states as classes is proposed in [6]. This modification is supported by the fact that temporal trajectory is quite different at the beginning and at the end of the phoneme. By classifying input features into more compact classes (in the feature space), such as phoneme model states, higher classification accuracy is achieved.

Our goal is to investigate, if better representation than posteriors can be found also in the net feeding directly the GMM/HMM models.

The outcomes of previous research are used in designing the system. The temporal patterns are transformed by DCT before entering the neural net classifier. Phoneme-states are used as classes for better classification accuracy. Finally, Heteroscedastic Linear Discriminant Analysis (HLDA) [7] technique was used instead of PCA for feature de-correlation similarly as in [8].

2. BOTTLE-NECK FEATURES

The classes-probability estimates or features derived from them do not have to be the best for subsequent classification.

We have seen, that in HATS [3], the authors resigned from using the band-conditioned posteriors at the output of the neural net and tried to find more suitable representation for the following merging classifier. Obviously, if the estimates from the first stage were perfect, the subsequent classification would not be necessary in the case when both stages have the same output classes.

The point is that band-conditioned estimators, as they see the information only from one critical band and the shapes of the temporal patterns can be very similar for several classes [9], provide rather poor classification accuracy. The usual accuracy of the classifiers is around 25%. Also, since the input is small and higher accuracy cannot be achieved, the classifier can be small. In this case, we can use the outputs of hidden layer directly as features for subsequent classification.

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication) and Caretaker project (FP6-027231), by Grant Agency of Czech Republic under project No. 102/05/0278 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under project No. 119/2004, No. 162/2005 and No. 201/2006

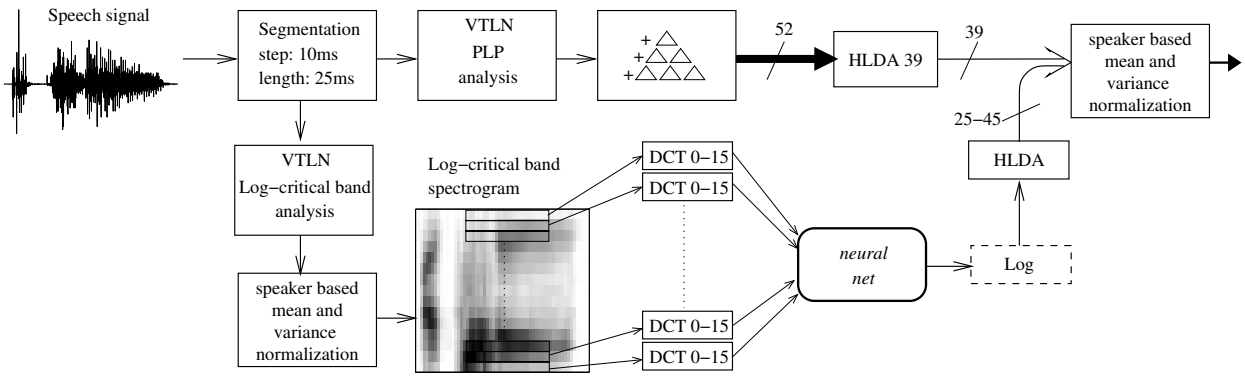

Fig. 1. The scheme of the feature extraction

The situation is different for the merging classifier. The subsequent GMM models do not have the same classes – they are trained to distinguish between context-dependent phonemes. For this, more informative features than the context-independent class posteriors are needed. The merging classifier is also quite big. The input vector is concatenation of all band-conditioned outputs and sufficient size of the hidden layer is also needed as it affects the final classification accuracy [10]. For practical reasons, it is not possible to pass several hundreds features to the GMM-HMM recognizer.

The dimensionality reduction techniques, which have to be used to obtain reasonable size of feature vector for the recognizer, have also drawbacks: first, they are only linear: the optimal dimensions for classes separations may not be found as the information in highly dimensional vectors may not by linearly separable. Furthermore, the PCA technique relies only on the variance of individual features, and the high dimensionality of input vectors causes problems when more sophisticated techniques such as HLDA are used.

If we do not want to use the dimensionality reduction techniques, and want to obtain the features suitable for the classification as outcome of neural net training process, a **bottle-neck** has to be created in the neural net structure. The neural net has the ability of nonlinear compression of the input features and of classification of such compressed features. If the trained neural net with bottle-neck has a good classification accuracy, we know that the bottle-neck outputs represents the underlying speech well.

For the reasons mentioned above, the use of widely employed three-layer MLP is not possible. A four-layer net can have bottle-neck in its first or second hidden layer. Net with bottle-neck in first hidden layer will have only limited power – one matrix multiplication and the nonlinearity – to extract suitable features from the input and the following layers may fail to correctly classify the input data. The bottle-neck in second hidden layer brings the disadvantage of poor classification of low dimensionality features into higher number of classes in only one layer. In both cases, the training procedure may not find the best features.

A four-layer MLP with bottle-neck in second hidden layer was used in [11]. Authors report improvement in small-vocabulary isolated-word recognition over PLP features. But the size of their bottle-neck (27 to 64 units) was similar to the output size (47 classes).

We decided to use **five-layer MLP** with the bottle-neck in the middle hidden layer. Such structure has enough power for extracting the internal features and also for their efficient classification.

To obtain the features, the neural net outputs are taken after the matrix multiplication and bias, before the sigmoid nonlinearity.

3. FEATURE EXTRACTION

The feature extraction follows the scheme in Fig 1. The upper branch of the diagram shows features consisting of 12th order PLP coefficients [12] plus energy computed over a 25 ms frame window every 10 ms. The Vocal Tract Length Normalization (VTLN) is used to reduce speaker variability. 1st, 2nd and 3rd order derivatives are calculated and appended to yield 52 dimensional feature vector. The dimensionality of these features is reduced to 39 by HLDA [7] and the resulting features are denoted HLDA-PLP .

The lower branch describes the computation of TRAP-based features. Here, we will not distinguish between the probabilistic and bottle-neck features as the processing steps are the same.

First, the power spectrum is computed over a 25 ms frame window every 10 ms. The power spectrum values are then integrated by 23 triangular filters and logarithm of their outputs is taken. Also here, the VTLN technique is used. The log-critical band energies are normalized to have zero mean and unity variance for each speaker. 31 consecutive frames from each critical band are transformed with 16 DCT bases including the 0^{th} base (DC offset). The transformed vectors creates input to the classifier with input size $16 \times 23 = 368$. The total number of weights in classifier was one million in all cases. The classifier outputs are states of 45 phoneme models, so there are 135 training targets. They are associated with the central frame of input critical band trajectories. Neural net outputs are de-correlated by HLDA which can also reduce the dimensionality of output features. The classes for HLDA are defined by tied states of context-dependent phoneme models. In case of probabilistic features, the logarithm precedes the HLDA.

Then, both streams are concatenated and features are mean and variance normalized per speaker.

4. EXPERIMENTAL SETUP

The system is based on AMI-LVCSR system used in NIST RT'05 evaluation [13]. Here we only summarize the main features of the system:

Data – The task is the recognition of meetings defined in NIST RT'05 evaluation. Independent headset microphone (IHM) data with reference segmentation were used. The test contains about 2 hours of speech. The training set consists of complete NIST, ISL, AMI and ICSI meeting data – about 114 hours.

bottle-neck size	25	30	35	40	45
accuracy [%]	47.5	47.7	48.0	47.9	48.2

Table 1. Classification accuracy of neural net with bottle-neck.

Recognition system works in two passes: The first pass, which is fixed in our experiments, generates wide latices with VTLN HLDA-PLP features (upper branch in Fig. 1), gender independent discriminatively trained cross-word acoustic models and bi-gram language model. These latices are then expanded by 4-gram language model.

The acoustic models contain 7700 tied states with 16 Gaussians components per state. The language model consist of 50K unigrams, 13M bigrams, 20M trigrams and 22M fourgrams. Its perplexity is 85.4 on NIST RT'04 development data.

New features are generated as described in section 3 for the second pass. Then, new models are trained starting with single-pass-retraining from the HLDA-PLP models used in the first pass followed by eight iterations of Baum-Welsh re-estimation. With these new models, the latices are acoustically rescored. The NIST RT'05 scoring is used to obtain the final word error rate (WER).

To compare the performance of the bottle-neck features with the probabilistic features, we used the four-layer MLP to estimate class probabilities. HLDA is then used for de-correlation and dimensionality reduction. The feature size of both bottle-neck and probabilistic features varied from 25 to 45 to find the optimal size.

We also compared the performance of each feature kind – i.e. HLDA-PLP, probabilistic and bottle-neck – on its own. The second pass of the system is performed for each feature kind.

The recognition constants – scale factor and word insertion penalty – need to be tuned for each dimensionality of feature vector. It was also found that probabilistic and bottle-neck features are very distinct and using the same constants for both kinds of features leads to suboptimal solution. Therefore, the constants are tuned also for each feature kind.

4.1. Neural net training

For the neural net training, one third of data from each site was used – about 38 hours together.

The training of the neural nets is split into two parts. First, only 10 hours of data randomly selected from training set were used. The standard back-propagation algorithm and the "newbob" learning rate scheduling[1] was used. The training stopped when the increment of classification accuracy on held-aside cross-validation (CV) set between two epochs was smaller than 0.5%.

In the second part, all neural net training data were used and another four epochs of training were performed. The learning rate started on the last value from the first part and was multiplied by 0.45 in subsequent epochs.

The neural nets are trained in exactly same way, only their structures differ. For probabilistic features, a four-layer MLP is used, two hidden layers have the same size. For bottle-neck features, a five-layer MLP with bottle-neck in the second hidden layer is trained. The first and third hidden layers have the same size. After the training, the MLP is used only till the bottle-neck nonlinearity (the nonlinearity is not applied). The portion of the net used to derive the bottle-neck features is 70%, the last 30% classify the features into the classes.

We used the SNet [14] training software which allows for parallel training. The time needed for complete training of one neural net was about 13 hours on four computers.

[1]The learning rate is kept fixed till the increment in cross-validation accuracy is bigger than a threshold. For the subsequent epochs, learning rate is halved till the increment falls below stopping threshold.

feature size (NN output)	64 (25)	69 (30)	74 (35)	79 (40)	84 (45)
probabilistic	26.1	25.9	25.6	25.7	25.7
bottle-neck	25.2	25.2	**24.9**	25.2	25.0

Table 2. WER for probabilistic and bottle-neck features in full feature extraction framework

feature size	25	30	35	39	40	45
HLDA-PLP	—	—	—	28.7	—	—
probabilistic	28.9	28.1	27.9	—	27.5	27.3
bottle-neck	27.3	26.9	26.6	—	26.6	**26.2**

Table 3. WER for each feature stream itself

5. COMPARISON OF PROBABILISTIC AND BOTTLE-NECK FEATURES

The first test of bottle-neck features quality is the comparison of classification accuracy of the neural net with bottle-neck and 4-layer MLP. Tab. 1 shows the cross-validation accuracies at the end of the training. The CV accuracy of the 4-layer MLP is 51.0%. So the bottle-neck neural nets are loosing about 3% in classification accuracy. If we compare these numbers with mostly used 3-layer MLP which has accuracy of 49.3%, the results are not bad mainly if we consider that for the classification of bottle-neck features, only 30% of MLP parameters are used.

Second test of the bottle-neck features quality is their perfomance in the LVCSR. They are compared to probabilistic features of the same size. In probabilistic features, the HLDA reduces the 135 dimensions of the neural net output to the desired size. The WERs of both features are shown in Tab. 2.

Finally, the performance of individual feature streams is tested. The results are shown in Tab. 3.

As mentioned above, the recognition constants are tuned for each dimensionality and each feature kind to give the lowest WER. Since there is no development set in our experimental setup, the tuning is done directly on the test set.

6. CONCLUSIONS AND DISCUSSIONS

The validation of classification performance of neural net with bottle-neck was done first. The cross-validation accuracy of five-layer MLP with bottle-neck is only about 3% worse than the classification accuracy of four-layer MLP. The drop in classification accuracy decreases with growing bottle-neck size. It means that more relevant information is available for classification in more dimensional bottle-neck. Considering that only 30% of all neural net parameters are used for classification of bottle-neck outputs, the classification accuracy is quite high.

The expected good performance of bottle-neck features compared to probabilistic features was confirmed in second test. In

LVCSR of meetings, they outperform the probabilistic features for all feature sizes by at least 0.5% which is good improvement for given task. The performance of both features kinds is increasing with the increasing size of features obtained from neural nets. The best performance is found for 35 neural net outputs: 74 dimensions in whole feature vector when probabilistic/bottle-neck features are used in combination with PLP features. Further increase of vector size does not bring improvement in WER.

Next, the recognition using only one feature kind was performed. The proposed bottle-neck features perform the best from all three feature kinds. Also, the probabilistic features outperformed the HLDA PLP features. This is actually the first time we have seen that probabilistic features outperform the PLP ones in LVCSR task. The WER is decreasing with increasing feature size when probabilistic or bottle-neck features are used itself. This shows that more of relevant information is passed from MLP to GMM-HMM system. But from the results with combined features we see that this information is partly present in the PLP features, as the increase of feature vector size above certain value does not bring further improvement in WER.

Finally, we performed experiments, where dimensionality of bottle-neck features was decreased by HLDA. As primary features, the bottle neck features of size 45 we chosen because they contain most of the information needed for classification as can bee seen from recognition results with individual kinds of features. While dimensionality reduction by HLDA brings slight improvement when only bottle neck features are used, no improvement or slight degradation is observed when used in combination with PLP features.

This suggests, that HLDA can pick up better information for subsequent GMM-HMM system but such information is already presented in PLP stream and the complementarity of the streams is partly lost. This behavior is caused by the choice of classes used in both techniques – MLP uses the states of context independent phonemes as targets, whereas HLDA accumulates statistics for the context-dependent tied states which are much closer to the targets modeled by GMM-HMM system.

The reduction of system size is also gained when bottle-neck features are used instead of probabilistic features. Only 70% of five-layer MLP is used to generate the bottle-neck outputs. Further, the following HLDA matrix is also smaller.

7. REFERENCES

[1] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP 2000*, Turkey, 2000.

[2] H. Hermansky and S. Sharma, "TRAPs – classifiers of temporal patterns," in *5th International Conference on Spoken Language Processsing (ICSLP)*, Sydney, Nov 1998.

[3] B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks," in *Proc. ICSLP 2004*, Jeju Island, KR, Oct. 2004.

[4] B. Chen, Q. Zhu, and N. Morgan, "Tonotopic multi-layered perceptron: A neural network for learning long-term temporal features for speech recognition," in *Proc. ICASSP 2005*, Philadelphia, PA, USA, Mar. 2005.

[5] Q. Zhu, B. Chen, F. Grézl, and N. Morgan, "Improved MLP structures for data-driven feature extraction for ASR," in *Proc. INTERSPEECH 2005*, Lisbon, Portugal, Sept. 2005.

[6] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *Proceedings of 7th International Conference Text, Speech and Dialogue 2004*, 2004, p. 8.

[7] N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. thesis, John Hopkins University, Baltimore, 1997.

[8] M. Karafiát, F. Grézl, P. Schwarz, L. Burget, and J. Černocký, "Robust heteroscedastic linear discriminant analysis and lcrc posterior features in meeting data recognition," in *3nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Washington, USA, May 2006.

[9] H. Hermansky and P. Jain, "Band-independent speech-events categories for TRAP based ASR," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 1013–1016.

[10] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *Proc. ICASSP 1999*, Phoenix, Arizona, USA, Mar. 1999, pp. 1013–1016.

[11] V. Fontaine, Ch. Ris, and J. M. Boite, "Nonlinear discriminant analysis for improved speech recognition," in *Proc. EUROSPEECH 1997*, Rhodes, Greece, Sept. 1997.

[12] H. Hermansky, "Perceptual linear predictive (PLP) analysis for the speech," *J. Acous. Soc. Am.*, pp. 1738–1752, 1990.

[13] T. Hain et al., "The 2005 AMI system for the transcription of speech in meetings," in *Proc. Rich Transcription 2005 Spring Meeting Recognition Evaluation Workshop*, Edinburgh, UK, July 2005.

[14] S. Kontár, "Parallel training of neural networks for speech recognition," in *Proc. 12th International Conference on Soft Computing MENDEL'06*, 2006.