DEEP HIERARCHICAL BOTTLENECK MRASTA FEATURES FOR LVCSR

Zoltán Tüske[a], Ralf Schlüter[a], Hermann Ney[a,b]

[a]Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany
[b]Spoken Language Processing Group, LIMSI CNRS, Paris, France
{tuske, schlueter, ney}@cs.rwth-aachen.de

ABSTRACT

Hierarchical Multi Layer Perceptron (MLP) based long-term feature extraction is optimized for TANDEM connectionist large vocabulary continuous speech recognition (LVCSR) system within the QUAERO project. Training the bottleneck MLP on multi-resolutional RASTA filtered critical band energies, more than 20% relative word error rate (WER) reduction over standard MFCC system is observed after optimizing the number of target labels. Furthermore, introducing a deeper structure in the hierarchical bottleneck processing the relative gain increases to 25%. The final system based on deep bottleneck TANDEM features clearly outperforms the hybrid approach, even if the long-term features are also presented to the deep MLP acoustic model. The results are also verified on evaluation data of the year 2012, and about 20% relative WER improvement over classical cepstral system is measured even after speaker adaptive training.

Index Terms— LVCSR, MRASTA, MLP, bottleneck, hierarchical, deep neural network, hybrid, TANDEM

1. INTRODUCTION

Becoming a major component, neural networks (NN) are widely used in recent automatic speech recognition systems. Besides the probabilistic TANDEM approach proposed by [1] for Gaussian Mixture Models (GMM), the hybrid (e.g. MLP based) acoustic models have been also explored as an alternative approach within the Hidden Markov Model framework [2, 3]. Estimating class posterior probabilities, the neural network based feature extraction can be considered as a non-linear feature transformation technique. The bottleneck approach, which can be interpreted as dimension reduction method using non-linear discriminant analysis, was introduced for speech recognition in [4]. The bottleneck features consistently outperform the posterior features, and are usually concatenated with MFCC. Based on the recent success of deep neural networks in hybrid acoustic modeling, the first steps of investigation of deep bottleneck features have been already taken in [5, 6].

Using short-term features, we have shown in our previous work that the hybrid MLP-HMM acoustic model achieved better performance than GMM-HMM, even with speaker adapted features [7]. However, the experiments also revealed that the gain over GMM-HMM is mainly related to the longer context used during the training of the high performing MLP acoustic model.

In order to incorporate long-term speech information in GMM-HMM, hierarchical processing of multi-resolution RASTA filtering was proposed in [8]. As [9] has shown, the combination of the bottleneck and hierarchical concept resulted in better performance. Moreover, training the hierarchical MLPs on concatenated MRASTA and critical band energies led to lower WER in [10]. The previous works applied simple 5-layer bottleneck structures, where the NNs estimated phoneme posteriors.

Therefore, in this paper we generalize the hierarchical bottleneck feature extraction in the following two ways before extending our previous investigation in [7] on long-term features. First, the output layer is optimized; second, deeper structures are used for bottleneck MLP training. Besides the comparison of the TANDEM with baseline MFCC systems, the best results are contrasted with deep pretrained MLP hybrid systems.

The paper is organized as follows: Section 2 gives an overview of related works followed by a short description of the training and testing corpora in Section 3. Section 4 gives the details of the feature extraction. We describe the experimental setups in Section 5 followed by results (Section 6). The paper closes with conclusion in Section 7.

2. RELATION TO PRIOR WORK

While the previous studies about deep bottleneck features investigated only short-term features [5, 6], in this paper we apply the MRASTA based long-term speech representation of [11]. The state-of-the-art version of the MRASTA features [9] includes hierarchical processing introduced by [8], and the bottleneck concept of [4]. This work generalizes it further by introducing context-dependent targets and deep structures. The paper also extends the comparison of hybrid and TANDEM approaches on long-term features [7].

3. CORPUS DESCRIPTION

Within the QUAERO project about 300 hours of manually transcribed speech data was collected to train the acoustic models and the MLPs. The corpus is based on various web resources containing recordings of broadcast news and conversations. During the optimization of long-term hierarchical bottleneck structure the evaluation corpus of 2011 is used, whereas the final recognition performance is also measured on the evaluation set of 2012 (Eval12) to validate the improved BN features. Development sets

(Dev11/Dev12) are used only for tuning system parameters like language model scale and time distortion penalties. Table 1 shows the corpus statistics of the training and testing data.

Table 1. *Statistics of QUAERO French training and testing corpora*

	Training	Dev11	Eval11/ Dev12	Eval12
total data [h]	317	2.9	3.1	3.8
# running words	3.9M	36k	38k	45k

4. FEATURE EXTRACTION

4.1. Cepstral features

From the audio files, vocal tract length normalized (VTLN) cepstral features are extracted, where the warping factors are estimated using a text-independent Gaussian mixture classifier trained on the acoustic training corpus. The preemphasized power spectrum is computed every 10 ms over a Hanning window of 25 ms. After integration of the warped power spectrum by 20 triangular filters equally spaced on Mel-scale, their logarithm is taken. The 16 MFCCs are computed from the logarithmic critical band energies (CRBE) and then segmentwise mean and variance normalized. Finally, linear discriminant analysis (LDA) is applied by projecting MFCCs within a sliding window of length 9 to a 45 dimensional subspace.

4.2. Bottleneck MRASTA features

The original RASTA filters were introduced to extract features which are less sensitive to linear distortion [12]. Covering the modulation frequency range found relevant for speech perception, multi-resolutional smoothing of temporal trajectories of the CRBE was proposed in [11] by applying two-dimensional bandpass filters. In our setup one second trajectory of each critical band is filtered by first and second derivatives of the Gaussian functions, where the standard deviation varies between 8 and 60 ms resulting in 12 temporal filters per band. After the MRASTA filtering, the first frequency derivatives are calculated according to [11]. Since the 20 logarithmized CRBE are extracted from the MFCC pipeline, they are also vocal tract length normalized.

To incorporate the high dimensional MRASTA representation of the speech signal in the GMM-HMM, we follow the BN approach proposed by [9]. Tied-state posterior estimates are derived from a hierarchical processing of two bottleneck MLPs: the input of the first MLP in this hierarchy is based on the fast modulation frequencies of the MRASTA filtering, whereas the second MLP is trained on the slow modulation frequencies and the BN output of the first MLP. In both cases, the inputs are *augmented* with CRBEs (*A*MRASTA).

Fig. 1 shows the MLP structures applied in this study, where we introduced deeper structure (up to 3 hidden layer) before and after the BN) in each level of the hierarchy. The number of target classes is optimized experimentally, see Section 6. If it is not explicitly stated, the bottleneck layer contains 42 nodes. In the first experiment, where a classical 5-layer BN-MLP is used, the number of nodes in the hidden layers is fixed to 7000 in accordance with our previous investigations. Using deeper structures the size of the non-bottleneck layers is limited to 3000.

For better initialization of the deep BN-MLP the discriminative pretraining of [2] is modified in the following way for bottleneck structures. In the first step classical 3-hidden-layer BN network is trained. After that the bottleneck layer is replaced by three randomly initialized layers, where the middle one is again a bottleneck. In the next step, only the four weight matrices correspond to the new layers are trained. The weight update is followed again by inserting new layers. In each growing step the data is seen by the network only once. After the desired depth is achieved, the whole network is fine-tuned.

The linear output of the final bottleneck layer is reduced by Principal Component Analysis (PCA) retaining 95% of the total variability, and concatenated with the LDA transformed MFCC. All activations of the nodes within the output layer are transformed by the softmax function, whereas the sigmoid transfer function is applied in all other layers. All MLPs are trained using the cross-entropy criterion and approximate triphone tied-state posterior probabilities. The nets are trained using backpropagation algorithm in the mini-batch mode (512 frames). To prevent overfitting and for adjusting the learning rate parameter, 10% of the training corpus (chosen randomly) is used as cross-validation set.

5. EXPERIMENTAL SETUP

5.1. Acoustic Modeling

Instead of training the acoustic models from scratch, an initial alignment is generated by one of our previous best systems [7], and used to estimate the decision tree for the state-tying, the LDA transformation matrix for cepstral features, and the GMM and MLP parameters. The speaker adapted features are extracted using Constrained Maximum Likelihood Linear Regression (CMLLR) [13] with the simple target model approach [14]. No additional MLLR [15] model adaptation is applied to GMM-HMM in order to be able to compare the performance of GMM-HMM with MLP-HMM fairly. Furthermore, no discriminative training is performed due to computational benefits.

To apply MLP as a hybrid acoustic model in the HMM framework, the class posterior probabilities are converted to emission likelihoods where the state prior probabilities are estimated on the training corpus and scaled (optimized on the development set). In our experiments, the deep hybrid MLP models have 8 hidden layers with 3000 nodes each, and the networks are initialized by discriminative pre-training [16]. In order to perform speaker adapted recognition with hybrid model, the best available features (CMLLR transformed MFCC+$\text{BN}^{\text{3x3k}}_{\text{AMRASTA}}$) are taken.

As observed in [17], MLP can benefit from speaker normalized features. Therefore, not only the deep hybrid acoustic model but the deep BN features are also retrained on speaker normalized features in the final experiment. The block diagram of the modified two-pass recognition is shown in Fig. 2. In this case, both the TANDEM and hybrid neural networks are trained on the same (speaker normalized BN-reduced) features. This concept follows [5], where it was observed that BN features trained on speaker adapted features led to lower error rates than speaker adapted BN features.

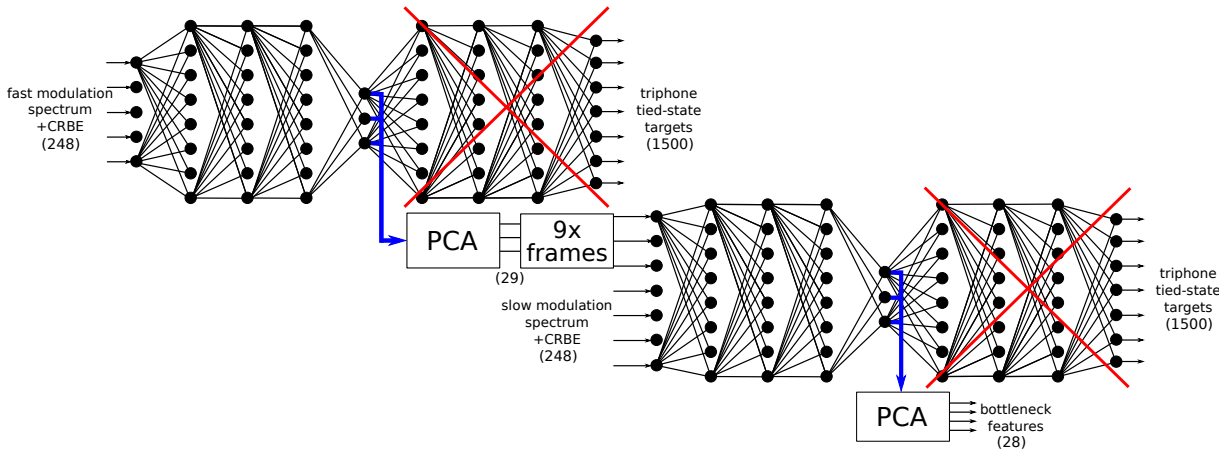

Fig. 1. *Deep hierarchical AMRASTA bottleneck features trained on context-dependent triphone tied-states. Feature dimensions are indicated in round brackets.*

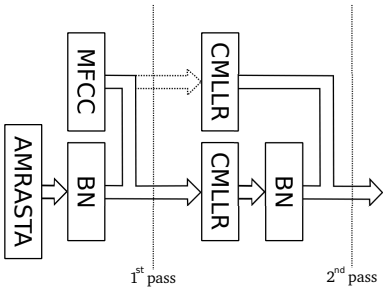

Fig. 2. *Two-pass recognition with bottleneck features retrained on speaker adapted features. For adaptation the output of the speaker independent system trained on concatenated cepstral and AMRASTA bottleneck features is used.*

5.2. Language Modeling

For the recognition of Eval11 set the same language model is used as in our previous work [7]. The recognition results on the Dev12 and Eval12 sets are achieved by applying a new LM estimated for the evaluation of 2012. The vocabulary contains 200k words, and the perplexity value for the smoothed, unpruned 4-gram LM is 122. All the recognition experiments are carried out with the freely available RASR decoder [18].

6. EXPERIMENTAL RESULTS

6.1. Optimization of BN MRASTA features

Based on our previous study [7] where we concluded that the MLP-HMMs outperformed the GMM-HMMs due to the longer temporal context the models incorporate, the long-term MRASTA features were optimized in the first experiment. According to our former investigation, 3-hidden-layer BN-MLPs are used with 7000 hidden nodes in the non-bottleneck layers. Furthermore, the short-term features based hybrid (MLP(MFCC)) and bottleneck TANDEM (BN^{1x7k}_{MFCC}) systems were also retrained on the extended training corpus. In this case, the input dimension of the MLPs is 297, resulting from nine frames of MFCC vectors, their full first-order, and part of the second-order time derivatives. As can be seen in Fig 3, the hierarchical long-term BN features clearly outperformed the short-term BN ones. The features reached their best performance when they were trained on 1500 tied-states, and even outperformed the hybrid system with very deep MLP structures. As a more fair comparison with the long-term BN TANDEM approach, a hybrid system was trained where the MRASTA filter outputs and CRBEs are also presented to the deep network (MLP(MFCC+AMRASTA)). The result indicates that the hybrid model can also benefit from the long-term features outperforming the GMM-HMM using simple, 5-layer BN structures. However, increasing the number of hidden layers in the bottleneck structures improved the BN AMRASTA features significantly, resulting in relative 4% lower error rate compared to our best hybrid system. Training a classical 3-hidden-layer BN features having the similar number of parameters as the deepest bottleneck structure (which corresponds to about 30k nodes in the non-bottleneck layers) hardly resulted in a better performance than $BN^{1x7k}_{AMRASTA}$. Therefore the performance gain originates mainly in the deeper structure. Futher comparison has been made between the hybrid and GMM acoustic modeling where one second of CRBEs were presented to the deep MLP and BN-MLP without any hierarchical or MRASTA processing (input vector dimension is 20·101=2020). The results (MLP(CRBE) and GMM(MFCC+BN^{3x3k}_{CRBE})) can indicate the necessity of the hierachical structure and MRASTA preprocessing for deep neural networks. Increasing the bottleneck up to 82 nodes resulted mainly in slight degradation. The initialization method described in Section 4.2 resulted in an (inconsistent) maximal improvement of 0.1% absolute. In summary, the total gain of our hierarchical deep BN TANDEM system over classical MFCC GMM-HMM is more than 25% relative. Our observations were also confirmed on the Eval12 test set (see Table 2).

6.2. Speaker adapted results

In the second set of experiments the optimized long-term features were investigated after speaker adaptation. Since the direct estima-

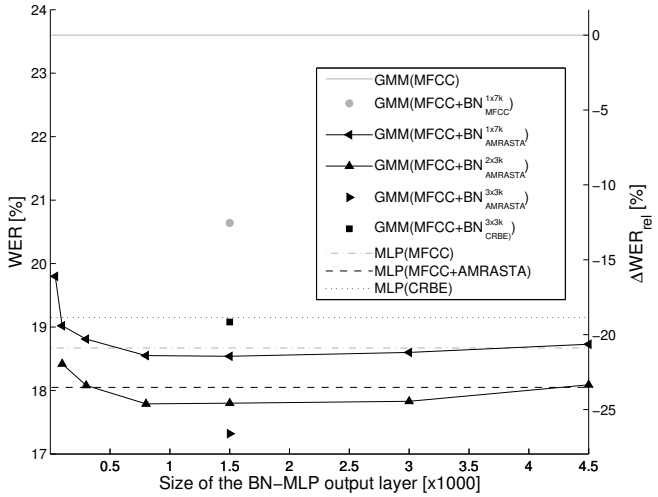

Fig. 3. *Optimization of augmented multi-resolution RASTA (AM-RASTA) based bottleneck (BN) features for GMM-HMM on Eval11 corpus. The number and the size of the hidden layer before and after the BN are indicated by superscript. The absolute performance in word error rate (WER) and the relative WER reduction (ΔWER_{rel}) over MFCC GMM-HMM is given on the left and right hand side, respectively. Color indicates whether the final feature vector contains 90ms (gray) or 1sec (black) of speech information.*

tion of CMLLR transformation matrices on the high dimensional MRASTA features can become inaccurate, the same features are used for both the hybrid MLP and GMM models to carry out the speaker normalized experiments. In another experiment we slightly modified the standard 2-pass recognition system (Fig. 2), and BN features were trained on speaker adapted features.

Table 2. *Speaker independent (SI) and adapted (SA) recognition results on Eval11 and Eval12 test sets using different features and acoustic models (AM)*

	AM	Features	Test set	
			Eval11	Eval12
SI	GMM	MFCC	23.6	24.4
		MFCC+$BN^{3x3k}_{AMRASTA}$	17.3	18.5
	MLP	MFCC+AMRASTA	18.1	19.3
SA	GMM	MFCC	21.9	22.7
		MFCC+BN^{3x3k}_{MRASTA}	16.6	17.6
		+deep BN retraining	**16.1**	**17.4**
	MLP	MFCC+BN^{3x3k}_{MRASTA}	16.9	17.8

In contrast to our previous study where only short-term features were used, the deep bottleneck TANDEM GMM-HMM based on long-term features slightly outperformed the hybrid MLP-HMM by about 1.5% relative (see Table 2). The performance gap increased further if the TANDEM features were retrained after the adaptation. The relative differences between the two systems are 4% and 2% on Eval11 and Eval12, respectively. It is worth mentioning that the GMM-HMMs are estimated using ML criterion only. Comparing the best NN based system with a classical GMM-HMM trained on stand-alone MFCC, more than 20% relative improvement are observed even after the 2nd pass.

7. CONCLUSIONS AND FUTURE WORK

Continuing our previous work, hybrid and TANDEM acoustic modeling approaches were compared using long-term features. According to our expectation the long-term TANDEM features improved the recognition rate substantialy compared to the BN features trained on short-term MFCC. Moreover, introducing the deep structure in the hierarchical bottleneck concept resulted in a significant gain, clearly outperforming the hybrid MLP approach. Our observation remained valid after speaker adaptation, although the performance gap decreased between the two modeling approaches.

This study aimed to compare GMM- and MLP-HMM using common features and feature adaptation methods. In the future we plan to analyze the effect of model adaptation techniques as well. The investigation also have to be extended on sequence-level discriminatively trained models.

Acknowledgement

This work has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. [213850]. 11, Speech Communication with Adaptive Learning - SCALE, and from the Quaero Programme funded by OSEO, French State agency for innovation. The study was partly supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

8. REFERENCES

[1] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 2000, pp. 1635–1638.

[2] F. Seide *et al.*, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," in *Proc of Interspeech*, 2011, pp. 437–440.

[3] T. N. Sainath *et al.*, "Making Deep Belief Networks Effective for Large Vocabulary Continuous Speech Recognition," in *Proc of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 30–35.

[4] F. Grézl *et al.*, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 757–760.

[5] T. N. Sainath *et al.*, "Auto-encoder bottleneck features using deep belief networks," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 4153–4156.

[6] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc of Interspeech*, no. August, 2011, pp. 237–240.

[7] Z. Tüske *et al.*, "Context-Dependent MLPs for LVCSR: TANDEM, Hybrid or Both?" in *Proc. of Interspeech*, 2012.

[8] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *ICASSP*, 2008, pp. 4165–4168.

[9] C. Plahl *et al.*, "Hierarchical Bottle Neck Features for LVCSR," in *Proc. of Interspeech*, 2010, pp. 1197–1200.

[10] F. Valente *et al.*, "Analysis and Comparison of Recent MLP Features for LVCSR Systems," in *Proc. of Interspeech*, 2011, pp. 1245–1248.

[11] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Interspeech*, 2005, pp. 361–364.

[12] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[13] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[14] G. Stemmer *et al.*, "Adaptive training using simple target models," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2005, pp. 997–1000.

[15] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[16] F. Seide *et al.*, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 24–29.

[17] Z. Tüske *et al.*, "A study on speaker normalized MLP features in LVCSR," in *Proc. of Interspeech*, 2011, pp. 1089–1092.

[18] D. Rybach *et al.*, "RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.