ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Context-Dependent MLPs for LVCSR: TANDEM, Hybrid or Both?

Zoltán Tüske, Martin Sundermeyer, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany
{tuske,sundermeyer,schluter,ney}@cs.rwth-aachen.de

10.21437/Interspeech.2012-5

Abstract

Gaussian Mixture Model (GMM) and Multi Layer Perceptron (MLP) based acoustic models are compared on a French large vocabulary continuous speech recognition (LVCSR) task. In addition to optimizing the output layer size of the MLP, the effect of the deep neural network structure is also investigated. Moreover, using different linear transformations (time derivatives, LDA, CMLLR) on conventional MFCC, the study is also extended to MLP based probabilistic and bottle-neck TANDEM features. Results show that using either the hybrid or bottle-neck TANDEM approach leads to similar recognition performance. However, the best performance is achieved when deep MLP acoustic models are trained on concatenated cepstral and context-dependent bottle-neck features. Further experiments reveal the importance of the neighbouring frames in case of MLP based modeling, and that its gain over GMM acoustic models is strongly reduced by more complex features.

Index Terms: HMM, GMM, MLP, bottle-neck, hybrid, ASR, TANDEM

1. Introduction

Nowadays, neural networks (NN) are widely used in LVCSR systems. However, due to the efficiency of the back-propagation algorithm, it is a special class of feed-forward NNs, MLP, which is applied in practice, and can be trained on hundreds of hours of speech data. Estimating the class (phone, phone state, tied triphone state) posterior probabilities, MLPs are applied in two different ways in current Hidden Markov Model (HMM) based automatic speech recognition (ASR) systems.

First, the probabilistic TANDEM [1] approach — after the logarithmic and feature reduction (PCA or LDA) transformation — uses the posteriors as features in a GMM based system. In the more specific bottle-neck TANDEM concept proposed by [2], an MLP consisting of at least 5 layers with a narrow one in the middle is trained, whereas the linear output of the middle (bottle-neck) layer is taken as output instead of the posteriors. The biggest advantage of the TANDEM approach is that all techniques developed for GMMs in the previous decades remain applicable, e.g. speaker adaptive (SAT) and discriminative training.

Second, like in classical artificial neural network based HMMs proposed in the early 90's [3], the GMM based emission probabilities are replaced by the class posterior probabilities estimated by an MLP. As an alternative to GMM in LVCSR, this hybrid approach recently became popular since it had been discovered that the MLP based estimation of posterior probabilities of thousands of tied triphone states is feasible. Moreover, the introduction of deep structures (many hidden layers) further increased the effectiveness of the MLP as acoustic model [4]. In order to remedy the convergence difficulties arising from the deep structure, different MLP pre-training algorithms [5, 6, 7] have been proposed.

Comparing speaker-independent, discriminatively trained GMMs and deep MLP acoustic models, more than 24% relative gain was reported on a conversational speech transcription task [4]. Even the MLP with a single but wide enough hidden layer was able to outperform the GMM based system. However, according to [8] the use of speaker adapted and discriminatively trained cepstral features decreased the performance gap significantly between the GMM and MLP models, and only MLPs with deep structures were able to outperform the traditional GMM on English broadcast news task.

Since in the state-of-the-art ASR systems [9, 10] MLPs are mainly used in TANDEM approach as features, in this paper our main goal is to compare the two different acoustic modeling techniques not only on MFCC, but also on concatenated cepstral and posterior features. Therefore, evaluating several feature transformation techniques (SAT, LDA) developed for GMM, the study is also extended to context-dependent MLP features.

The paper is organized as follows. Section 2 describes the corpus used in our experiments. The details of our experimental setups are given in Section 3. Section 4 reports the results. The study finishes with conclusions in Section 5.

2. Corpus description

The comparison of the different acoustic models is carried out on a French broadcast news and conversations task. Focusing on transcription of web data, different speech types (news, comedy, cooking sessions, interviews, and talk-shows, etc.) are collected within the QUAERO project. At the current state of the project about 250 hours of transcribed French speech data are available to train the acoustic models and neural networks (Train). While the feature extraction and system parameters are tuned on the development corpus (Dev10) of 2010, the evaluation set (Eval10) of 2010 is used for measuring the recognition performance. To mitigate the "training-on-testing-data" problem our final conclusions are drawn from the recognition per-

formance achieved on the evaluation data from 2011 (Eval11). For the evaluation of 2011 a new language model (LM) is estimated, where the LM scale parameter is tuned on Dev11. Table 1 summarizes the corpus statistics of training and testing data.

Table 1: *Training and testing corpora*

	total data [h]	# running words
Train	257	9,800k
Dev10	3.7	41k
Eval10/Dev11	2.9	36k
Eval11	3.1	38k

3. Experimental setups

3.1. Features

3.1.1. Cepstral features

From the audio files, vocal tract length normalized cepstral features are extracted. The pre-emphasized power spectrum is computed every 10 ms over a window of 25 ms. After integration of the warped power spectrum — 20 triangular filters are used, equally spaced on Mel-scale — the features are logarithmized. Finally, we compute the 16 MFCCs from the logarithmic critical band energies and apply mean and variance normalization. When linear discriminant analysis (LDA) is applied, features within a sliding window of length 9 are projected to a 45 dimensional subspace.

3.1.2. TANDEM MLP features

To select the best posterior features for our experiments two types of MLP features are tested: classical 3-layer MLP based posteriors which are further transformed by logarithm [1], and 5-layer MLP based bottle-neck features [2], where the linear output of the bottle-neck layer was taken as features. Both feature sets are reduced by Principal Component Analysis (PCA) according to 95% of the variability. All MLPs are trained using the cross-entropy criterion and approximate class posterior probabilities. The size of the output layer — from 42 phones up to 4500 tied triphone states using the decision tree estimated for GMM — is optimized experimentally, see Section 4.2. All activations of the nodes within the output layer are transformed by the softmax function — all outputs sum up to 1, whereas the sigmoid transfer function is applied in all other layers. Nine consecutive frames of acoustic feature vectors are fed to the MLP. The number of nodes in the hidden layer is fixed to 7000, except the bottle-neck layer which has 42 nodes. Furthermore, the net is trained using back-propagation algorithm in mini-batch (512 frames) mode. Ten percent of the training set (chosen randomly) is used as cross validation set for adjusting the learning rate and to prevent overfitting.

3.2. Acoustic models

3.2.1. GMM-HMM

In order to obtain the GMM acoustic models, the steps from [11] are followed. Instead of training the acoustic models from scratch, an initial alignment is generated by the previous best system [10], and used to estimate the decision tree of the state-tying, the LDA matrix for cepstral features, and the mixture parameters in the first iteration steps. To extract the speaker normalized (SAT) features, the Constrained Maximum Likelihood Linear Regression (CMLLR) speaker normalization is applied using the simple target model approach [12].

3.2.2. Hybrid MLP-HMM

To use an MLP as acoustic model in the HMM system instead of GMM the class posterior probabilities are converted to emission likelihoods $p(x|s)$:

$$p(x|s) \sim \frac{P(s|x)}{P(s)^{\alpha}}, \tag{1}$$

where x denotes the observation vector (e.g. the 9 consecutive MFCC frames), and s refers to the HMM state. The state prior probabilities, $P(s)$, are estimated on the training corpus, and its scale, α, is optimized on the development set. For single hidden layer MLPs the nets are the same as in Section 3.1.2. In our experiments with deep MLP structures, MLPs with three hidden layers each having 3072 nodes are trained. In order to train deep structures we followed [6], where the layers are initialized by discriminative pre-training (DPT). As it was shown, DPT initialized NNs slightly outperformed the nets pre-trained with Restricted Boltzmann Machines.

3.3. Language model

In order to be able to compare the performance of the hybrid systems with our previous results, the same LM is used as in [11] during the recognition experiments on Dev10 and Eval10. The details about this LM are available in [10]. Moreover, the recognition results obtained on the Dev11 and Eval11 sets are achieved by a new LM estimated on the available data collected in the QUAERO project (4.2 billion words), where the vocabulary contains 200k words. To smooth the 4-gram LM, the discount parameters are estimated on held-out data according to [13]. The perplexity value of the new unpruned LM is 111.2. As speech recognizer the RASR [14] is used.

4. Experimental results

4.1. Comparison of GMM and MLP acoustic modeling using cepstral features

In the first two experiments the different acoustic modeling are compared using the cepstral (MFCC) and speaker normalized cepstral features (MFCC_{SAT}). The previously optimized GMMs with 4500 tied states estimated on LDA transformed (9 frames) MFCCs are considered as baseline acoustic model [11]. The results in word error rate (WER) on Eval10 are shown in Fig. 1. Investigating the optimal size of the MLP output, we can conclude that the hybrid acoustic model reaches the lowest WER,

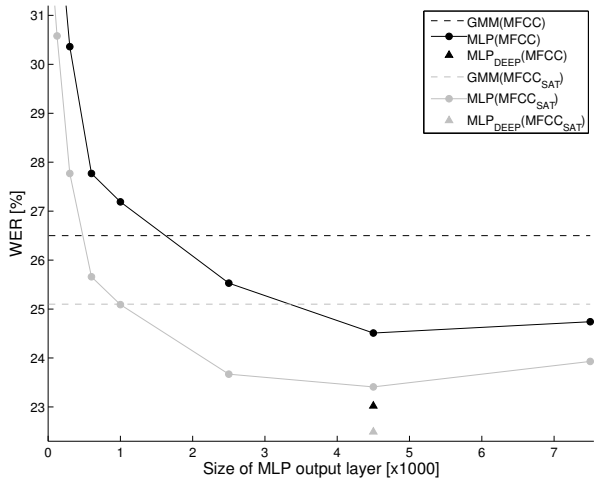

Figure 1: *Effect of the output layer size on Word Error Rate (WER) obtained on Eval10 test set using hybrid MLP acoustic models and cepstral features*

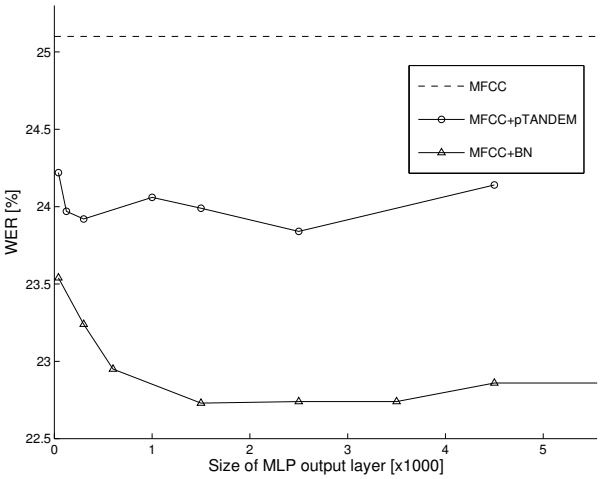

Figure 2: *Effect of the output layer size on Word Error Rate (WER) obtained on Eval10 test set after SAT using concatenated cepstral and probabilistic TANDEM (pTANDEM) or bottle-neck (BN) features*

when the 3-layer MLP are trained using 4500 tied triphones states as output labels. Therefore, for the further experiments the output layer of the hybrid model is fixed to this size. As can be seen, the speaker independent hybrid MLP system performs more than 1.5% absolute better than the GMM based. In addition, it achieves an even lower WER than the GMM system trained on speaker normalized data. Doubling the hidden layer size up to 14000, we did not observe any significant improvement. Furthermore, repeating the experiment with deep MLP, the performance gap increased further between the GMM and MLP based acoustic models. The relative improvement is over 10% in the speaker independent case, whereas the relative gain reduces to 8% using more complex SAT features. The improvement achieved by the MLP over the GMM are consistent with the results reported previously in the literature [4, 8].

4.2. Selection of context-dependent MLP features

In order to compare the two acoustic modeling techniques on TANDEM features, first the MLP features trained on 9 frames context are optimized. Since in our previous work [10] the MLP features are extracted from phone class posteriors, in this study the effect of the output layer size is investigated for both the classical probabilistic and bottle-neck TANDEM features, as well. The MLP features are always concatenated with MFCCs in these experiments, and the results are compared after speaker adaptive training and two-pass GMM based recognition. As Figure 2 shows, increasing the output layer could lead to better TANDEM features. The bottle-neck features (BN) outperform the probabilistic TANDEM (pTANDEM) features in all cases. Furthermore, the bottle-neck features also benefit from the wider output layer, although the bottle-neck layer size is not changed. Considering also the fact that the dimension of the probabilistic TANDEM features is increasing with the growing output size of the MLP, the context-dependent bottle-neck features seem to be more attractive. For subsequent experiments, the BN features are chosen, where the output layer is fixed to 2500. Exploiting context-dependent bottle-neck features lead to more than 3% relative gain over the phone posterior based BN features and to 9% relative improvement over the stand-alone MFCC$_{SAT}$ features. Moreover, BN features and GMM acoustic modeling shows slightly better result on Eval11 set after the second pass than the hybrid acoustic model trained on MFCC$_{SAT}$ (Table 2).

4.3. Comparison of GMM and MLP acoustic modeling using MFCC+BN features

Using the optimized MLP features, in the next step the different acoustic modeling techniques are compared on the more sophisticated speaker normalized, concatenated cepstral and bottle-neck features, (MFCC$_{LDA}$+BN)$_{SAT}$. As can be seen in the last row of the Table 2, better performance is achieved with the deep MLP acoustic modeling. Using context-dependent MLP features and deep MLP acoustic model together results in the best recognition rate. However, training a single hidden layer MLP as acoustic model (result not reported) or using the same number of input feature frames as in case of GMM, no improvement over GMM is observed.

4.4. Effect of the number of input frames

As a more fair comparison with GMMs, the hybrid acoustic models are also trained on single frame of state-of-the-art features. Nevertheless, a single feature vector can implicitly contain information over longer context, e.g. LDA transformation is trained on 9 consecutive MFCC frames. Investigating the effect of the number of input frames to train the MLP, we observed that the MLP acoustic model could be even less effective than the GMM. As can be seen in Table 2, using the same feature context the MLP acoustic models perform slightly better than GMM. Based on the results on Eval11 set, the improvement over GMM is about 3% relative when MFCC$_{LDA}$ or

Table 2: *Comparison of GMM (baseline) and deep MLP based acoustic modeling (AM) with different features. Results are given as word error rate (WER).*

	Test set	Dev10			Eval10			Dev11			Eval11		
	AM	GMM	MLP		GMM	MLP		GMM	MLP		GMM	MLP	
	# input frames	1	1	9	1	1	9	1	1	9	1	1	9
Features	MFCC+Δ+$\Delta\Delta$	27.4	27.3	21.9	29.8	29.3	23.0	28.5	27.8	21.8	26.7	27.2	20.5
	MFCC_{LDA}	24.6	23.6	22.1	26.5	25.3	23.2	25.3	24.0	21.9	23.6	22.8	20.8
	MFCC_{SAT}	23.8	23.5	22.0	25.1	24.5	22.5	23.8	23.1	21.4	21.6	21.1	19.4
	$(\text{MFCC}_{\text{LDA}}\text{+BN})_{\text{SAT}}$	21.6	21.8	21.4	22.7	22.7	21.9	21.6	21.4	20.6	**19.0**	19.1	**18.4**

MFCC_{SAT} features are applied. However, using the time derivatives of the cepstral features (MFCC+Δ+$\Delta\Delta$) or the complex $(\text{MFCC}_{\text{LDA}}\text{+BN})_{\text{SAT}}$ features the GMM acoustic model outperformed the hybrid MLP. Therefore, the results in Table 2 confirm an essential property of MLP based modeling: despite the strong correlation between the neighbouring frames, MLP could directly benefit from the longer context of 9 frames which usually leads to performance degradation in case of GMM without additional dimension reduction. E.g. GMM takes longer context into account using LDA on 9 consecutive frames of MFCC.

5. Conclusions

Using French broadcast news and conversation LVCSR task, the present study has investigated several MFCC based features in GMM and MLP acoustic modeling. From the results in Table 2 we can conclude that MLP based acoustic modeling outperforms the GMM based one. The difference between the two acoustic modeling method is over 10% relative when linear transformed (derivatives, LDA, CMLLR) MFCC features are applied. Nevertheless, the performance gap decreases using more general transformation. Moreover, a non-linear transformation of MFCC, the bottle-neck features, was also investigated. Using either hybrid modeling or bottle-neck TANDEM approach similar recognition performance was observed. However, the best performance was achieved when deep MLP was trained on the speaker normalized concatenated MFCC and context-dependent bottle-neck features (bottle-neck-TANDEM hybrid MLP acoustic model). Further experiments also revealed the importance of the neighbouring frames to train high performing MLP acoustic model.

Since our research was limited to short-term TANDEM features, we intend to carry out further experiments with long-term features (e.g. MRASTA) and even deeper MLPs, as well. Furthermore, other discriminative features (e.g. fMMI) should be also part of the future work.

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. Furthermore, the work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

6. 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. 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.

[3] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

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

[5] A. Mohamed *et al.*, "Acoustic Modeling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14 –22, jan. 2012.

[6] 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.

[7] C. Plahl *et al.*, "Improved Pre-training of Deep Belief Networks using Sparse Encoding Symmetric Machines," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Mar. 2012.

[8] 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.

[9] F. Valente *et al.*, "A comparative large scale study of MLP features for Mandarin ASR," in *Proc. of Interspeech*, 2010, pp. 2630–2633.

[10] M. Sundermeyer *et al.*, "The RWTH 2010 QUAERO ASR Evaluation System for English, French, and German," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 2212–2215.

[11] Z. Tüske *et al.*, "Comparison and combination of different CRBE based MLP features for LVCSR," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Mar. 2012.

[12] G. Stemmer *et al.*, "Adaptive training using simple target models," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 997–1000.

[13] M. Sundermeyer *et al.*, "On the Estimation of Discount Parameters for Language Model Smoothing," in *Proc. of Interspeech*, 2011, pp. 1433–1436.

[14] 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.