MULTI-LEVEL ADAPTIVE NETWORKS IN TANDEM AND HYBRID ASR SYSTEMS

Peter Bell, Pawel Swietojanski, and Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK

ABSTRACT

In this paper we investigate the use of Multi-level adaptive networks (MLAN) to incorporate out-of-domain data when training large vocabulary speech recognition systems. In a set of experiments on multi-genre broadcast data and on TED lecture recordings we present results using of out-of-domain features in a hybrid DNN system and explore tandem systems using a variety of input acoustic features. Our experiments indicate using the MLAN approach in both hybrid and tandem systems results in consistent reductions in word error rate of 5–10% relative.

***Index Terms*—** deep neural networks, tandem, hybrid, MLAN, TED, BBC

1. INTRODUCTION

In an ASR system, neural networks may be used to directly compute HMM observation probabilities (the *hybrid* approach [1, 2]) or for feature extraction (the *tandem* approach [3]). Tandem systems are HMM-GMM systems which use features derived from neural networks trained as phone classifiers, concatenated with the original acoustic features. Tandem systems have successfully used both decorrelated and dimension-reduced phone posterior log probability features [4] and hidden layer bottleneck features [5]. In both formulations, neural networks have the advantage of being inherently discriminative — trained to optimise phone or state probabilities — and can incorporate wide acoustic contexts.

During the 1990s hybrid systems achieved good experimental results on some large vocabulary tasks [6, 7]. However, HMM-GMM systems became increasingly more accurate owing to the use various techniques such as context-dependent phone modelling [8] speaker adaptation using MLLR [9], sequence-level discriminative training techniques such as MMI and MPE [10], and high-dimension feature space transforms such as fMPE [11], which either took advantage of the GMM structure or were much more computationally feasible for GMM-based systems compared with hybrid systems. State-of-the-art HMM-GMM speech recognition systems employ these techniques in combination with the use of tandem features (e.g. [12]).

This research was supported by EPSRC Programme Grant grant EP/I031022/1, *Natural Speech Technology*.

Recent results have shown that hybrid systems can offer significantly improved accuracy on large vocabulary conversational speech recognition problems, compared with state-of-the-art HMM-GMM systems [13, 14]. The major differences compared with earlier work are: (1) The use of deep neural networks with many hidden layers [15]; (2) Context-dependent HMM states as output classes[1]; (3) Generative pre-training using restricted Boltzmann machines [18]. There are now a number of published comparisons between hybrid HMM-DNN and HMM-GMM systems [14, 19, 15, 20]. There are fewer comparisons between tandem and hybrid DNN systems: Sainath et al [21] found a tandem system using deep autoencoder bottleneck features was more accurate than a comparable hybrid system.

Posterior features obtained from neural networks have been used successfully in tandem systems for both cross-lingual adaptation [22, 23, 24] and cross-domain adaption [25, 26, 27, 28]. In these approaches, out-of-domain (OOD), or foreign language posteriors are combined using a merger MLP [27, 28, 24]; GMMs are retrained [26, 25]; or networks adapted by additional training [26]. In contrast to cross-lingual adaptation, when data from the target language is typically assumed to be sparse, domain adaptation has the potential to bring benefits even to resource-rich languages, not least because domains can be characterised with increased resolution as the amount of data increases.

We recently proposed a domain adaptation procedure based on DNNs called multi-level adaptive networks (MLAN) [29], described in Section 3. The central feature of MLAN is that a second DNN is trained on tandem features generated from out of domain nets. Thomas [22] has proposed a shallow neural network approach for spoken term detection that is rather similar to MLAN. A related multi-layer network structure, predating the current use of deep networks, was proposed by Schwarz et al [30], but was not employed for adaptation. We previously used MLAN exclusively in a tandem framework [29]; here we investigate the technique in a hybrid system, comparing MLAN hybrid systems with MLAN tandem systems, and presenting results on a larger scale task, the recognition of TED talks, in addition to the recognition of multi-genre broadcast recordings.

[1] Hybrid systems with limited context-dependence have been investigated previously [16, 17]

2. TRAINING TANDEM AND HYBRID DNNS

We trained DNNs to model frame posterior probabilities, using labels obtained from Viterbi forced alignment using an HMM-GMM system. The hidden layers use logistic sigmoid nonlinearities and a softmax is used for the output layer to provide posterior probability estimates for each output class. All nets were pre-trained in an unsupervised manner using layerwise restricted Bolzmann machine (RBM) pre-training [18]. The nets were further fine-tuned to minimise the negative log-posterior probability of the true class labels, using stochastic gradient descent. Training was performed using the Theano library [31] on NVIDIA GeForce GTX 690 GPUs.

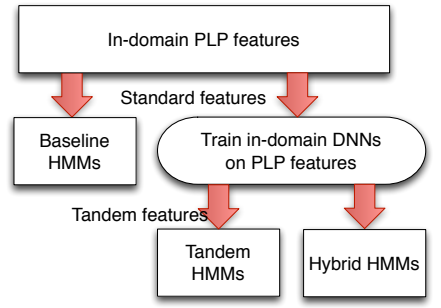
The tandem systems used monophone target classes. The output posterior probabilities were decorrelated and reduced to 30 dimensions using PCA, and then concatenated with the original acoustic features, which were used to train tandem HMM-GMMs from scratch. The DNN structure of the tandem systems was not optimised on word error rate (WER), but was set to minimise frame error rate on validation data. All tandem systems used 4-layer nets with 9 frames of acoustic context and 1024 units in each hidden layer.

The hybrid systems used output classes corresponding to tied context-dependent phone states. Scaled likelihood estimates were obtained by dividing the outputs by class priors estimated over the whole training data. The state tyings were obtained from an HMM-GMM system. The hybrid DNNs used 2048 hidden units in each layer and a acoustic context of 9 frames; we carried out experiments with 1–6 hidden layers. The context-dependent DNNs used in the hybrid approach offer higher resolution phonetic modelling, compared with tandem DNNs. However the tandem HMM-GMM system combines some of the advantages of DNNs — discriminative feature extraction, wide acoustic context — with the benefits of using generative statistical models, including model-space speaker adaptation and the use of HMM discriminative training algorithms.

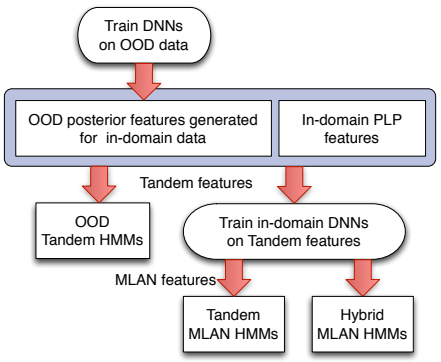
3. MLAN: MULTI-LEVEL ADAPTIVE NETWORKS

A network trained on data from one domain may be adapted to a new domain by further training using the new data [26]: the original data effectively provides a good initialisation for the weights of the final net. When RBM pre-training is used for initialisation, we observe only small gains from this approach. In a tandem setup, where the nets are trained on OOD data, adaptation may be performed by re-training or MAP-adapting the GMM part of the system on the target domain.

The MLAN approach [29] uses posterior features from a net trained on OOD data, augmented with in-domain acoustic features, as input to a further set of DNNs which are trained on in-domain data (Figure 1(b)). The advantage of the MLAN scheme, is that the second-level DNNs are able to select discriminatively the most important elements of the OOD features. Aside from Thomas et al [24], other schemes operate only on the outputs of first-level MLPs, without the original features, limiting the power of the second-level DNN; or they use shallow MLPs [30], so that the final structure is effectively one single DNN without pre-training. The second-level DNN can incorporate multiple sets of complementary OOD features (Section 4.4).



(a) Conventional DNN systems



(b) MLAN DNN systems

Fig. 1. A comparison of standard and MLAN DNN systems.

We investigated using the MLAN scheme in both a tandem and hybrid setup. We trained two sets of second-level DNNs on OOD tandem features: one set generates monophone posterior features, and was used for a tandem HMM-GMM system. The second set generate posteriors over context-dependent tied states which transformed to scaled likelihoods and used in a hybrid MLAN system. Figure 1 illustrates the six possible setups: (a) the shows the three systems in which no OOD data is used, and (b) shows the three systems utilising OOD network.

Figure 1(b) is in fact a simplification, as a further consideration for the hybrid MLAN system is the choice of GMM to obtain the triphone state tying. We chose to use the same state tying structure for tandem and hybrid systems trained on a given set of features. This implies that the tandem MLAN system is required to generate the tying structure used in the hybrid MLAN system. This minimises the influence of input features to which the DNNs are invariant.

4. EXPERIMENTS

4.1. ASR tasks

We carried out experiments on two diverse large vocabulary speech recognition tasks. The first, `bbc`, is a corpus of multi-genre television and radio shows from the British Broadcasting Corporation (BBC), kindly made available by BBC Research and Development, containing speech that is mostly British English. The radio portion includes a wide range of genres, including news, weather, on-location reports and parliamentary debates; the television portion consists of several episodes of a drama series, which includes dramatic, fast, emotional speech and high levels of background noise, making ASR particularly challenging. More detail of the corpus can be found in [29]. We divided the data at the show level into a training set of 20.7 hours and a test set of 2.3 hours.

Our second task is the TED English transcription task from the IWSLT evaluation campaign [32]. We present results on the `dev2010`, `tst2010` and `tst2011` sets, each containing 8-11 single-speaker talks of approximately 10 minutes' duration. The talks are pre-segmented by the IWSLT organisers [33]. Following the IWSLT rules, our in-domain acoustic model training data was derived from 813 publicly available TED talks dating prior to end of 2010. After automatic segmentation and lightly-supervised alignment, 143 hours of speech remained for acoustic model training.

4.2. Out-of-domain data

To investigate the MLAN technique, two varied sources of out-of-domain data were used: firstly 276 hours of US-English conversational telephone speech (CTS) taken from the Switchboard I, Switchboard II and CallHome corpora; secondly a set of multi-party meetings from the AMI corpus (AMI), described in [12]. Only the second of these was used in the TED system, because we felt the mismatch in style between the CTS and TED domains was particularly high – though we did not investigate this experimentally. Table 1 summarises the total quantities of training data available.

Domain	Training data (hrs)
BBC	20.8
TED	143.0
CTS	276.0
AMI	126.8

Table 1. Quantities of in-domain and OOD training data

4.3. Experimental setup

All systems used 3-state cross-word triphone HMMs. Models trained on the BBC data, used around 3,000 tied states; models trained on TED data used 12,000 states, reflecting the larger quantity of training data available. There were 16 Gaussians per state. The baseline acoustic features were were perceptual linear prediction (PLP) coefficients with first, second and third temporal derivatives, projected to 39 dimensions with an HLDA transform. All features were normalised for mean and variance at the speaker level.

The out-of-domain neural network features used varied with the corpus. For the AMI corpus, a stacked bottleneck architecture with a filterbank input was used following [12]. For the CTS corpus, tandem posterior features were computed using DNNs, after down-sampling the in-domain data to 8khz. In all cases the neural network features had to 30 dimensions, giving a total feature vector dimension of 69.

To make the tandem systems as competitive as possible, the HMM-GMM models were trained with MPE, which gave 1-2% improvement over ML-trained models. Due to the inaccuracy of speaker labels, we did not perform any speaker adaptation on the BBC data. For the TED system, we performed adaptation at the talk level. For the tandem setups, we used model-space CMLLR adaption with 32 block diagonal transforms per speaker, adapting the neural-network and acoustic features independently, and also performed speaker adaptive training (SAT). For the hybrid systems, initially no adaption was used; however, we later performed an adaptive training procedure by applying a feature-space MLLR transform derived the relevant tandem system to all input feature vectors. Unless otherwise noted, results were obtained using trigram language models, using HDecode.

4.4. Results

We firstly present result on the BBC task. Table 2 gives results for the baseline tandem and hybrid systems, and 3 gives results for the MLAN systems. Table 2 shows that the use of both in-domain and OOD tandem features provides substantial gains over the PLP baseline, confirming the domain-portability of the neural networks. Both AMI and CTS features improve performance, despite the relatively high domain mismatch[2] – the AMI features, in fact, perform better than the in-domain tandem system. The application of the MLAN technique yields additional improvements, for OOD features, the best system reducing WER by 2.9% absolute over the best tandem system. In particular, the results show that hybrid systems also benefit from training new DNNs on OOD tandem features, reducing WER from 33.7% for the baseline hybrid system to 29.6% when AMI and CTS tandem features were included. In this case, the hybrid MLAN systems give lower WER over equivalent tandem MLAN systems, with the same state-tying.

Secondly, we present results on the TED lecture task. Development results on the `dev2010` and `tst2010` sets are shown in Table 4. The results largely support the findings

[2] AMI and CTS baseline GMM systems used directly on the BBC task give WERs of 51.8% and 64.1% respectively.

Input features	WER (%)
PLP	36.1
BBC tandem	33.0
AMI tandem	32.5
CTS tandem	34.1
BBC hybrid	33.7

Table 2. WER (%) for baseline tandem and hybrid systems on the BBC test data.

MLAN system	Tandem	Hybrid
AMI	31.0	31.0
CTS	30.7	30.1
AMI + CTS	29.9	**29.6**

Table 3. WER (%) for tandem and hybrid MLAN systems on the BBC test data.

from the BBC experiments – even though much a larger quantity of in-domain training data is available – in that the use of AMI tandem features reduces WER compared to the PLP baseline[3], and both tandem and hybrid MLAN systems consistently improve performance further: by 1.5% and 1.2% on `tst2010` respectively, compared to equivalent systems without OOD features. We found that the use of SAT, using a single feature-space linear transform, was essential for the hybrid system to achieve competitive performance compared to the tandem system. In Table 5 we also show the results of our best-performing systems on the `tst2011` test set, additionally rescoring lattices with 4-gram LM. Here we found the tandem MLAN system to perform slightly better than the hybrid system with SAT. However, it should be noted that the tandem system benefits from more powerful adaptation transforms and the use of MPE training. Finally, Figure 2 illustrates the effect of increasing the number of layers of the hybrid MLAN DNN – we see that they continue to benefit from increased depth, as is well-known in the standard case.

System	dev2010	tst2010
PLP	21.0	20.3
TED tandem	19.4	17.9
AMI tandem	20.3	18.1
Baseline hybrid	21.0	20.3
+ SAT	18.6	17.6
Tandem MLAN	18.5	16.4
Hybrid MLAN	19.2	17.8
+ SAT	17.8	16.4

Table 4. A comparison of tandem systems on the TED lectures. Baseline and tandem systems are trained with SAT and MPE.

[3]An AMI baseline GMM with speaker-adaption on the TED task give WERs of 32.0% on `dev2010` and 30.7% on `tst2010`.

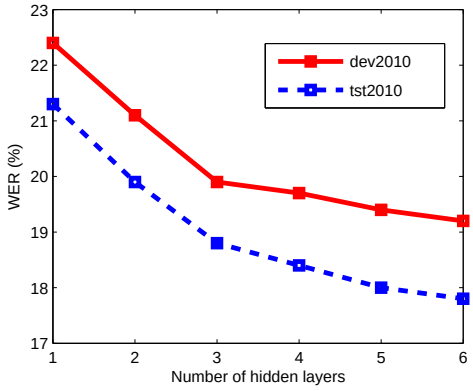


Fig. 2. The effect of increasing the number of DNN layers for the hybrid MLAN systems on TED lectures (systems without SAT)

System	WER
Tandem MLAN	15.1
+ SAT + MPE	12.8
+ 4gram LM	12.4
Hybrid MLAN	14.2
+ SAT	13.0
+ 4gram LM	12.6

Table 5. Results of MLAN systems on the `tst2011` test set

5. CONCLUSIONS

We have presented experiments using out-of-domain tandem features on two diverse ASR tasks. The results demonstrate that substantial gains in performance may be obtained by the use of multi-level adaptive networks, when both tandem and hybrid deep neural networks are trained on tandem features generated for in-domain data using out-of-domain network weights, even when the domain-mismatch is large. Our experiments comparing tandem and hybrid DNNs suggest that there is not a large difference in performance between the two.

There are, however, a number of directions to explore in future research. Firstly, the best form of OOD features for MLAN has not been investigated. We have currently used fixed 30-dimensional tandem features, but large vectors may be better when large quantities of OOD data are available. We also plan to compare bottleneck and posterior features as inputs to the MLAN DNNs.

The relatively large amount of data available for adaptation to each speaker – and the fact that speaker labels are known – leads to particularly large gains from speaker adaptation on the TED task, which benefits the tandem systems much more than the hybrid systems. For DNN hybrid systems to achieve maximum performance on tasks like this, it is clear that more effective methods of adaptation are required.

6. REFERENCES

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

[2] S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 161–174, 1994.

[3] H. Hermanksy, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1630.

[4] Q. Zhu, A. Stolcke, B.Y. Chen, and N. Morgan, "Using MLP features in SRIs conversational speech recognition system," in *Proc. Interspeech*, 2005.

[5] F. Grézl, M Karafiát, S. Kontar, and J. Černokcý, "Probabilistic and bottleneck features for LVCSR of meetings," in *Proc. ICASSP*, 2007.

[6] M. M. Hochberg, S. J. Renals, A. J. Robinson, and G. D. Cook, "Recent improvements to the ABBOT large vocabulary CSR system," in *Proc. IEEE ICASSP*, 1995, pp. 69–72.

[7] A. J. Robinson, G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams, "Connectionist speech recognition of broadcast news," *Speech Communication*, vol. 37, no. 1–2, pp. 27–45, 2002.

[8] S. J. Young and P. C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition," *Computer Speech & Language*, vol. 8, no. 4, pp. 369–383, 1994.

[9] "Maximum likelihood linear transforms for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 75-98, 1998.

[10] D. Povey and P.C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*. IEEE, 2002, vol. I, pp. 105–108.

[11] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "FMPE: Discriminatively trained features for speech recognition," in *Proc ICASSP*, 2005.

[12] T. Hain, L. Burget, J. Dines, P.N. Garner, F. Grezl, A.E. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[13] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[14] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, 2011.

[15] A. Mohammed, G.E. Dahl, and Hinton G., "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[16] H. Bourlard, N. Morgan, C. Wooters, and S. Renals, "CDNN: a context dependent neural network for continuous speech recognition," in *Proc. ICASSP*, 1992, vol. 2, pp. 349–352.

[17] D. Kershaw, T. Robinson, and S. Renals, "The 1995 ABBOT LVCSR system for multiple unknown microphones," in *Proc. ICSLP*, 1996, pp. 1325–1328.

[18] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[19] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, 2011.

[20] T. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A. Mohammed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Proc. ASRU*, 2011.

[21] T. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Proc ICASSP*, 2012.

[22] S. Thomas, S. Ganapathy, A. Jansen, and H. Hermansky, "Data-driven posterior features for low resource speech recognition applications," in *Proc. Interspeech*, 2012.

[23] F. Grézl, M. Karafiát, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. ASRU*, 2011.

[24] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. ICASSP*, 2012.

[25] S. Sivadas and H. Hermansky, "On use of task independent training data in tandem feature extraction," in *Proc. ICASSP*, 2004.

[26] A. Stolcke, F. Grézl, M.-Y. Hwang, X Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006.

[27] V.-B. Le, L. Lamel, and J.-L. Gauvain, "Multi-style MLP features for BN transcription," in *Proc. ICASSP*, 2010, pp. 4866–4869.

[28] J. Pinto, M. Magimai-Doss, and H. Bourlard, "MLP based hierachical system for task adaptation in ASR," in *Proc. ASRU*, 2009.

[29] P.J. Bell, M.J.F. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P.C. Woodland, "Transcription of multi-genre media archives using out-of-domain data," in *Proc. IEEE Workshop on Spoken Language Technology*, Dec. 2012.

[30] P. Schwarz, Matějka P., and J. Černokcý, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, 2006.

[31] J Bergstra, O Breuleux, F Bastien, P Lamblin, R Pascanu, G Desjardins, J Turian, D Warde-Farley, and Y Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. SciPy*, 2010.

[32] M Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[33] M. Cettolo, C. Girardi, and M Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16^{th} Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[34] E. Hasler, P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski, "The UEDIN systems for the IWSLT 2012 evaluation," in *Proc. IWSLT*, 2012.