REVISITING HYBRID AND GMM-HMM SYSTEM COMBINATION TECHNIQUES

Pawel Swietojanski, Arnab Ghoshal, and Steve Renals

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

{p.swietojanski, a.ghoshal, s.renals}@ed.ac.uk

ABSTRACT

In this paper we investigate techniques to combine hybrid HMM-DNN (hidden Markov model – deep neural network) and tandem HMM-GMM (hidden Markov model – Gaussian mixture model) acoustic models using: (1) model averaging, and (2) lattice combination with Minimum Bayes Risk decoding. We have performed experiments on the "TED Talks" task following the protocol of the IWSLT-2012 evaluation. Our experimental results suggest that DNN-based and GMM-based acoustic models are complementary, with error rates being reduced by up to 8% relative when the DNN and GMM systems are combined at model-level in a multi-pass automatic speech recognition (ASR) system. Additionally, further gains were obtained by combining model-averaged lattices with the one obtained from baseline systems.

Index Terms— deep neural networks, tandem, hybrid, system combination, TED

1. INTRODUCTION

The automatic combination of multiple acoustic models is an important and commonly-used technique to improve the accuracy of automatic speech recognition (ASR) systems. Acoustic models may be combined at many different levels, from the feature to the final recognition output. *Explicit combination* operates directly in model space by averaging the likelihood scores produced by each of the particular models. *Implicit combination* operates in the space of generated sentence hypotheses and attempts to rescore or alter recognition hypotheses provided as either n-best lists or lattices.

One of the first attempts that allowed to combine multiple ASR outputs was the Recogniser Output Voting Error Reduction (ROVER) technique [1]. This approach attempts to combine 1-best recognition output obtained from two or more ASR systems into word-level networks, which may then be realigned and alternative word hypotheses at each time step are scored using voting or confidence measures. An alternative, explicit approach employed the Minimum Bayes Risk (MBR) criterion to minimise the expected word error rate with respect to the approximated hypothesis posterior distribution [2]. This latter approach led to the notions of confusion networks (CN) with consensus decoding [3] and CN combination (CNC) [4]. Both methods operate on lattices, rather than the best hypothesis or an N-best list, and use reliable confidence measure-based voting mechanisms.

Although different modelling techniques, feature extraction methods, and phonesets often provide some degree of complementarity, there is no theoretical guarantee. Additionally, the potential improvement from combining the given systems is difficult to estimate in advance of conducting experiments. Hence some effort has been put into incorporating complementarity into training criteria [5, 6], including approaches such as mixtures of experts [7] and products of experts [8].

In this paper we investigate the combination of hybrid HMM-DNN systems with tandem HMM-GMM systems. The tandem systems we explore combine stacked bottleneck features (obtained from a narrow hidden layer in feed-forward network) [9] with posteriogram features (obtained from neural network outputs) [10] in MLAN configuration [11]. The hybrid HMM-DNN system uses the DNN to estimate posterior probabilities of context dependent HMM states, which are transformed to scaled likelihoods and used directly as output probabilities in the HMM system [12, 13]. An additional motivation for this work is that constructing a context-dependent hybrid HMM-DNN system requires a trained HMM-GMM system in order to obtain the training alignment with context-dependent phone states. Indeed the hybrid system uses the set of context-dependent tied states defined by the HMM-GMM system. Furthermore, it is advantageous to perform feature space adaptation when training a hybrid system, and again such an adaptation can be conveniently obtained by constrained maximum likelihood linear regression (CMLLR) transforms [14], which enable the DNNs to be trained in a speaker-adaptive fashion.

We explore combining hybrid HMM-DNN systems with tandem HMM-GMM systems by performing experiments on the TED talks speech recognition task as used in the IWSLT evaluation campaign [15]. We have investigated combinations based on model averaging of likelihoods and on MBR-based combination of lattices.

This research was supported by EPSRC Programme Grant grant, no. EP/I031022/1 (Natural Speech Technology). We would also like to thank Daniel Povey for helpful discussion on MBR-based system combination, and Peter Bell for discussion on MLAN.

2. RELATION TO PRIOR WORK

There were a number of studies concerning the combination of GMM-based systems and hybrid HMM system in the 1990s. These investigations, mainly based on model averaging, showed some success when combing context-independent hybrid systems based on multi-layer perceptrons (MLPs) and recurrent neural networks (RNNs). Dugast et al. [16] combined posterior probability estimates obtained from a time-delay neural network with the likelihoods generated by an HMM system with state emissions modelled by a mixture of Laplacians. Similar approaches combining scaled-likelihoods produced by a two-layer MLP and HMM-GMM likelihoods were also investigated [17]. Hochberg et al. [18] smoothed together the outputs from RNNs using different features and running forward and backward in time. Segmental neural networks [19] may be viewed as an example of an implicit combination in which a phone segment-level MLP was used to rescore an n-best list produced using a GMM-HMM system. MLPs have also been used to estimate state-dependent weights for mixtures of Gaussians [20].

Although there has been intense interest recently in using DNN-based systems for ASR, there has been little reported work on model combination using DNNs. Sainath et al. [21] combined a strong PLP-based HMM-GMM system with a tandem HMM-GMM system that used autoencoder bottleneck features, and Jaitly et al. [22] combined PLP-GMM and hybrid systems using the SCARF framework [23].

3. MODEL COMBINATION

The experiments presented in this paper are based on the averaging of acoustic model likelihoods, and on lattice combination techniques. In this section we outline these two combination techniques.

3.1. Model averaging

Models could be combined in a number of ways at the frame or state likelihood level. Likelihood scores can be combined most simply in a frame-synchronous way. Alternatively sequences of frame-level likelihoods can be combined asynchronously using multi-stream approaches [24]. Here we perform frame-synchronous combination using a linear interpolation of the observation log-likelihoods under the two models:

$$\log(p(\mathbf{o}|q_j)) = \lambda \log \frac{P_{DNN}(q_j|\mathbf{o})}{P(q_j)} + (1-\lambda) \log p_{GMM}(\mathbf{o}|q_j), \quad (1)$$

where $P(q)$ is the probability of the tied triphone state q obtained from a state-level alignment of the training data, and λ is a scaling factor that is optimised on a development set. This approach assumes that the acoustic models share the same decision tree for the context-dependent tied states.

Combining likelihoods using (1) has limited theoretical justification in comparison with the logarithm of the averaged probabilities ($\log((p_1 + p_2)/2) \geq (\log p_1 + \log p_2)/2$), however, work on the combination of context-independent scaled likelihoods indicates that (1) offers better experimental results and does not require the likelihoods to be similarly scaled.

3.2. Lattice combination using MBR decoding

We compare the score-level combination done by model averaging with hypothesis-level combination using an MBR-based combination [25], which was shown to improve over the more traditional ROVER [1] and CNC [4]. The MBR combination finds the word sequence that minimises the expected word error rate across the different systems being combined:

$$W^* = \arg\min_W \left\{ \sum_{i=1}^{N} \lambda_i \sum_{W' \in \mathcal{L}_i} P(W'|\mathbf{O}; \mathcal{M}_i)\, L(W, W') \right\}, \quad (2)$$

where $L(W, W')$ is the Levenshtein distance between two word sequences, and $P(W|\mathbf{O}; \mathcal{M}_i)$ is the posterior probability of the word sequence W given the acoustic observation sequence $\mathbf{O}$ as computed under the i-th model $\mathcal{M}_i$. This posterior probability is approximated by that computed over the lattice $\mathcal{L}_i$ corresponding to the i-th system:

$$P(W|\mathbf{O}; \mathcal{M}_i) \approx \frac{P(W)\, p(\mathbf{O}|W)}{\sum_{W' \in \mathcal{L}_i} P(W')\, p(\mathbf{O}|W')}.$$

4. EXPERIMENTAL SETUPS

We have performed a set of experiments combining DNN-based hybrid and tandem systems using the publicly available TED talks corpus [26] according to the ASR evaluation protocol used in the IWSLT–2012 evaluation campaign [15]. The training data consists of 813 publicly available TED talks published before the end of 2010. After automatic segmentation and lightly-supervised alignment 143 hours of speech remained for training purposes [27]. The DNNs were trained on 131 hours of speech, the remaining 12 hours (45 talks) were used for cross-validation. We present results on two predefined test sets, referred to as dev2010 and tst2010, containing 8 and 11 talks of about 10 minutes duration, respectively.

Our HMM-GMM systems were built using the open source Kaldi speech recognition toolkit [28], and the DNNs code were trained with software that utilised the Theano library [29], which allows for transparent CPU and GPU computations. DNN training was carried out using NVIDIA GeForce GTX 690 GPUs.

4.1. HMM-GMM System

The GMM acoustic models use 13-dimensional perceptual linear prediction (PLP) features with first and second order

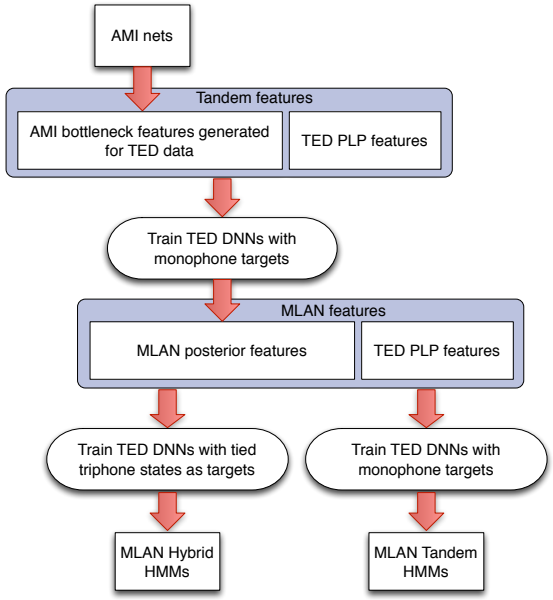

Fig. 1. MLAN features generation used for systems G1, H2 and H3.

differential coefficients. The models have 12,000 context-dependent tied states with around 16 Gaussians per state. Speaker adaptive training (SAT) is done using a single constrained (feature-space) MLLR transform per speaker. The SAT models are discriminatively trained using the boosted maximum mutual information (BMMI) [30] criterion. Results for the different HMM-GMM systems can be found in table 1 including contributions of each of the stages.

4.2. MLAN Features

Multi-Level Adaptive Networks (MLAN) [11] are a neural network based method to exploit out-of-domain training data. The fundamental idea, as depicted in figure 1, is that DNNs trained on out-of-domain data produce posterior or bottleneck features for in-domain data, and these neural network features are combined with the acoustic (PLP) features to train a second DNN on in-domain data. In this work we used bottleneck features obtained from a network trained on the AMI corpus [31] as inputs to a second network trained on the TED data. The resultant second level MLAN posteriors benefit from out-of-domain data and are adapted to the in-domain data. These posteriors were then concatenated with PLPs and used in both tandem fashion [10] to train the new GMM acoustic models (table 1) as well as for a context-dependent DNN to use in a hybrid HMM-DNN system (table 2).

4.3. Hybrid System

Our hybrid HMM-DNN system is similar to the one recently proposed by Dahl et al. [13], in which a pre-trained DNN estimates (scaled) likelihoods for context-dependent tied states of an HMM system. The tied states are extracted from a corresponding HMM-GMM system. Following the reported experience of Dahl et al. [13] and recommendations for gradient-based training of deep structures [32], as well as our previous experience with DNNs for ASR [33], we choose the network to have six hidden layers with 2048 units in each hidden layer.

The networks are initialised from stacked restricted Boltzmann machines (RBMs) that are pretrained in a layerwise fashion [34]. Finetuning was performed using stochastic gradient descent optimisation. The required hyperparameters were set to the same values as in our previous work [33], except the initial learning rate, which was increased to 0.16.

We present the ASR results for these hybrid systems in Table 2. We used both PLP and MLAN tandem features, and in both cases we trained an additional speaker adaptive training (SAT) variant, in which a constrained (feature-space) MLLR transform obtained from an HMM-GMM system was used to adapt the input features to the DNN. Although the term SAT is used to describe both tandem and hybrid systems, there is a difference: SAT for DNNs is effectively doing a cross-adaptation, in which the CMLLR transforms are estimated using a HMM-GMM system and applied to an HMM-DNN system.

Table 1. Word error rates for the baseline HMM-GMM systems.

System		WER(%)	
		dev2010	tst2010
ML PLP GMM		32.0	36.9
+SAT		23.5	23.0
+BMMI	(G0)	21.0	20.3
ML MLAN GMM		23.4	22.2
+SAT		19.7	17.8
+BMMI	(G1)	18.3	17.3

Table 2. HMM-DNN hybrid results for different features types

Features		WER(%)	
		dev2010	tst2010
PLP	(H0)	20.0	18.9
+SAT	(H1)	18.0	16.2
MLAN	(H2)	19.1	17.5
+SAT	(H3)	17.9	15.8

5. RESULTS

Before discussing our model combination results, we first discuss the baseline results, focusing on the tst2010 set to avoid complicating the exposition. First, based on the WERs in table 1 we can see how the application of SAT transforms in the

Table 3. Model averaging and corresponding lattice combination results. Relative WER changes are given in parenthesis w.r.t. the better of the two systems. λ scales hybrid systems and is optimised on devset (GMMs are scaled by 1-λ)

Systems combined	WER(%)	
	dev2010	tst2010
Model averaged experiments		
G0 $\oplus$ H0, $\lambda = .6$	18.7 (-6.5)	**17.3 (-8.0)**
G0 $\oplus$ H1, $\lambda = .7$	17.8 (-2.2)	**16.0 (-1.2)**
G1 $\oplus$ H2, $\lambda = .6$	17.6 (-3.8)	**16.5 (-4.6)**
G1 $\oplus$ H3, $\lambda = .7$	17.1 (-4.5)	15.8 (0)
Lattice combination experiments		
G0 $\oplus$ H0, $\lambda = .6$	18.0 (-10.0)	17.9 (-5.3)
G0 $\oplus$ H1, $\lambda = .6$	17.9 (-0.5)	17.1 (+5.5)
G1 $\oplus$ H2, $\lambda = .4$	18.0 (-1.6)	17.1 (-1.2)
G1 $\oplus$ H3, $\lambda = .7$	17.5 (-2.2)	**15.6 (-1.3)**

HMM-GMM system results in a very large reduction in WER, since a sufficiently large amount of speech ($\sim$10 minutes) is available per speaker. For the ML PLP variant, the WER drops from 36.9% to 23% after CMLLR adaptation. Using MLAN features results in a much lower WER for the unadapted system (22.2%), which is reduced by a further 4.4% absolute after performing SAT. Applying a CMLLR transform to the HMM-DNN hybrid system also gave significant WER reductions of 2.7% and 1.7% absolute, when trained on PLP and MLAN features respectively. This compares well to recently reported results on different tasks, conversational telephone speech [35] and North American English broadcast news transcription [21].

The first part of Table 3 contains the results of linear combinations of likelihood scores using eq. (1). The results indicate that the hybrid system becomes less complementary to the HMM-GMM when both were trained on features adapted using CMLLR. This is not unexpected since the CMLLR transform, used for the features on which the hybrid system DNN is trained, was obtained using the HMM-GMM system. The second part presents the same combination but obtained at the lattice level by MBR decoding.

Operating in the hypotheses-space allows combination of systems with different decision trees — results of which are presented in Table 4 where we cross-combine the systems trained on MLAN and PLP features. Finally, Table 5 shows that model averaging and lattice combinations are complementary to each other: for example, comparing the last lines of Table 4 and 5, we see that first combining the best MLAN GMM system (G1) with the best MLAN hybrid system (H3) using model averaging followed by an MBR combination with the best PLP hybrid system gives better results than a 3-way MBR combination of the same systems. This may be due to the fact that in the first case the MBR decoding starts with a better set of hypotheses to choose from.

Table 4. MLAN/PLP lattice combination results.

Systems combined	WER(%)	
	dev2010	tst2010
G1 $\oplus$ H0	17.9 (-2.2)	16.9 (-7.6)
G1 $\oplus$ H1	17.3 (-3.9)	15.8 (-2.5)
H2 $\oplus$ H0	18.6 (-2.1)	17.2 (-1.7)
H2 $\oplus$ H1	17.7 (-1.7)	16.4 (+1.2)
H3 $\oplus$ H0	17.7 (-1.1)	17.3 (+9.5)
H3 $\oplus$ H1	17.4 (-2.8)	15.6 (-1.26)
G1 $\oplus$ H3 $\oplus$ H1	17.4	15.2 (-3.8)

Table 5. Selected lattice combination results with lattices obtained from model-averaged decodes (Table 3)

Systems combined	WER(%)	
	dev2010	tst2010
G1H3 $\oplus$ G0H1	16.8 (-1.7)	15.0 (-5.1)
G1H3 $\oplus$ G1H2	17.2 (+0.6)	15.8 (0)
G1H3 $\oplus$ H1	16.9 (-1.2)	**14.9 (-5.7)**

6. CONCLUSIONS AND FUTURE WORK

In the paper we investigated model- and lattice-based system combination techniques for HMM-DNN and HMM-GMM systems trained using the most recent advances in both domains. We showed that model-averaging HMM-GMM and HMM-DNN systems improves the final accuracy and that applying the same CMLLR transforms reduces the complementarity between the combined HMM-GMM and HMM-DNN systems. We also showed that combination of model-averaged systems with each other, or with baseline systems, using MBR combination may bring further gains in accuracy. Finally, small reductions in WER could be obtained by training DNNs on MLAN features.

A promising future direction involves the exploration of filterbank features, which were found to be a good choice for DNNs [36], and may prove to be more complementary to models trained on conventional cepstral features. It is also possible to have a hybrid HMM-DNN version of the products of experts [37], whereby different DNNs become responsible for different subsets of tied-states obtained from HMM-GMM decision tree clustering.

7. REFERENCES

[1] JG Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE ASRU*, 1997, pp. 347–352.

[2] A Stolcke, Y Konig, and M Weintraub, "Explicit word error minimization in n-best list rescoring.," in *EUROSPEECH*, 1997.

[3] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other appli-

cations of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373 – 400, 2000.

[4] G Evermann and PC Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. NIST Speech Transcription Workshop*, 2000.

[5] C Breslin and MJF Gales, "Generating complementary systems for speech recognition.," in *INTERSPEECH*, 2006.

[6] C. Breslin and M. J. F. Gales, "Directed decision trees for generating complementary systems," *Speech Communication*, vol. 51, no. 3, pp. 284–295, 2009.

[7] J. Fritsch, M. Finke, and A. Waibel, "Adaptively growing hierarchical mixtures of experts," in *Advances in Neural Information Processing Systems*, 1997, pp. 459–465.

[8] MJF Gales and SS Airey, "Product of gaussians for speech recognition," *Computer Speech and Language*, vol. 20, 2006.

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

[10] H Hermansky, DPW Ellis, and S Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, 2000.

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

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

[13] GE 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 & Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[14] MJF Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, April 1998.

[15] 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.

[16] C. Dugast, L. Devillers, and X. Aubert, "Combining TDNN and HMM in a hybrid system for improved continuous-speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 217 –223, jan 1994.

[17] 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.

[18] M. Hochberg, S. Renals, T. Robinson, and D. Kershaw, "Large vocabulary continuous speech recognition using a hybrid connectionist/HMM system," in *Proc. ICSLP*, Yokohama, 1994.

[19] G. Zavaliagkos, Y. Zhao, R. Schwartz, and J. Makhoul, "A hybrid segmental neural net/hidden Markov model system for continuous speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 151 –160, jan 1994.

[20] Y. J. Chung and C. K. Un, "Multilayer perceptrons for state-dependent weightings of HMM likelihoods," *Speech Communication*, vol. 18, no. 1, pp. 79 – 89, 1996.

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

[22] N Jaitly, P Nguyen, A Senior, and V Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Interspeech*, 2012.

[23] G Zweig and P Nguyen, "Scarf: a segmental conditional random field toolkit for speech recognition," in *Interspeech*, 2010, pp. 2858–2861.

[24] A Morris, A Hagen, H Glotin, and H Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, vol. 34, no. 1–2, pp. 25–40, 2001.

[25] H Xu, D Povey, L Mangu, and J Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, vol. 25, no. 4, pp. 802—828, October 2011.

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

[27] 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.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, December 2011.

[29] 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.

[30] D Povey, D Kanevsky, B Kingsbury, B Ramabhadran, G Saon, and K Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. IEEE ICASSP*, 2008, pp. 4057–4060.

[31] 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.

[32] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," arXiv.1206.5533, 2012.

[33] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. IEEE Workshop on Spoken Language Technology*, Miami, 2012.

[34] GE Hinton, S Osindero, and Y Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, 2006.

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

[36] A Mohamed, GE Dahl, and GE Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, 2012.

[37] GE Hinton, "Products of experts," in *Proc. Int. Conf. Artificial Neural Networks (ICANN)*, 1999, vol. 1, pp. 1–6.