REGULARIZATION OF CONTEXT-DEPENDENT DEEP NEURAL NETWORKS WITH CONTEXT-INDEPENDENT MULTI-TASK TRAINING

Peter Bell and Steve Renals

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

ABSTRACT

The use of context-dependent targets has become standard in hybrid DNN systems for automatic speech recognition. However, we argue that despite the use of state-tying, optimising to context-dependent targets can lead to over-fitting, and that discriminating between arbitrary tied context-dependent targets may not be optimal. We propose a multitask learning method where the network jointly predicts context-dependent and monophone targets. We evaluate the method on a large-vocabulary lecture recognition task and show that it yields relative improvements of 3-10% over baseline systems.

***Index Terms*—** deep neural networks, multitask learning, regularization

1. INTRODUCTION

This paper proposes a technique for managing the trade-off between modelling monophone and tied triphone states in deep neural network (DNN) based ASR. The use of state-clustered context dependent (CD) phone units (also known as *senones*) was an key advance in the performance of HMM-GMM continuous speech recognition systems in the 1990s [1], giving an important competitive advantage over hybrid neural network based approaches such as [2, 3] used at the time. Although some hybrid systems used a limited form of context-dependence [4, 5, 6], they were limited in the ability to fully model CD phone units partly by computational constraints and the limited capacity to scale to larger quantities of training data, and partly by the innovation of GMM-based state-clustering to deal with data sparsity issues, limiting the number of tied states modelled to far below the maximum number of possible CD units.

The benefit of CD units in GMM systems may be explained by two distinct effects: first, given sufficient training data, increasing the number of states modelled by using CD labels over context-independent (CI) monophone labels simply results in a larger, more powerful generative model of the acoustic space. This divide and conquer approach is similar to modelling different genders or background noise conditions using separate states, which may give benefits on test data even when the gender or noise is unknown at test time . Second, and perhaps more significantly, however, when computing the acoustic likelihood of observed acoustics x and a hypothesised phone sequence $(p_1, p_2, \dots)$ during decoding, the decoder replaces the joint probability $p(x, p_1, p_2, \dots)$ with $p(x, p_1{+}p_2, p_1{-}p_2{+}p_3, \dots)$, which may be interpreted as a context-sensitive adaptation of each phone model to its local acoustic context under the current hypothesis. This allows more effective discrimination: suppose that phones p_i and p_j are acoustically close. If there are contexts a, b, c, d where $p_a{-}p_i{+}p_b$ and $p_c{-}p_j{+}p_d$ overlap in acoustic space, this would present a problem for a context-independent decoder, but is more easily discriminated by a decoder using context-dependent units, especially if the contexts a, c and b, d are acoustically distinct.

After a period in which neural networks were primarily used to generate features for GMM systems (the *tandem* and *bottleneck* approaches [7, 8]), they have returned to widespread use in hybrid systems, generating posterior probabilities over tied states, which are scaled by state priors to obtain pseudo-likelihoods for use in a standard HMM decoder. Along with the use of deep network structures, a major factor in this revival was the switch to using networks with tied-state CD phone targets obtained from a GMM [9, 10], which demonstrated marked improvements over the use of monophone targets. To date, most DNN-based systems have followed [9] in using the GMM to obtain the CD state clustering and alignments for training [11], although there are recent approaches using purely DNNs for this task [12, 13], removing all requirement for GMMs.

In this paper, we question whether this approach to modelling CD units is optimal, regarding the requirements for both discrimination and generalisation. We propose a DNN acoustic model which jointly predicts both CD and CI units using multitask learning. In the following section we motivate the proposed method and discuss its relation to prior work.

2. CONTEXT MODELLING IN DNNS

We consider the advantages of modelling CD units with DNNs, with the respect to the benefits for GMMs outlined in the previous section. Clearly, the second effect of improved discrimination when decoding continues to apply: when computing the likelihood of a hypothesised phoneme sequence, it is useful for the decoding to take into account acoustic context. But does the first effect of more powerful modelling of the acoustic space still apply? Unlike the GMM, when trained with the usual cross-entropy criterion the DNN is an inherently discriminative model, learning weights to optimise the decision boundaries between classes. The functional form of these decision boundaries is determined by the network structure and class labels, rather than directly by the within-class model structure – and so the first effect is not applicable. In fact, we suggest, there is a disadvantage to using CD phones as DNN targets: there is no distinction between discrimination between different phones, and between different contexts of the same phone. The latter discrimination, as well as being rather arbitrary, varying according to the specific state clustering used, has a much more limited benefit to producing a more accurate phone hypothesis at test time – yet both are treated equally in cross-entropy DNN training. This may result in lower layers of the network learning discriminations that are not beneficial to the ultimate performance, essentially wasting free parameters. It may be noted that two standard methods of discriminative training for GMMs, using the MMI or MPE criteria [14] that operate

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

at the sequence level do not attempt to increase discrimination between different triphone contexts of the same phone. Also, the earliest proposal for context-dependent neural network (CDNN) acoustic modelling [4], was based on a conditional probability factorisation which resulted in a network trained to discriminate monophones, being combined with one or more networks trained to discriminate the context given the monophone.

There is of course, a second problem with modelling CD units compared to CI units, being the inherent data sparsity issue in having a large output layer, despite the use of state clustering. Increasing the number of output units obviously increases the number of weights to be trained between the output layer and final hidden layer, with fewer samples with fewer samples available to train each weight. Via backpropagation, this may lead to over-fitting in the lower layers .

We propose a novel use of multitask learning [15] to solve both problems described above. The aim of multitask learning is to improve the generalisation of the network by applying it to two related tasks, sharing the hidden layers, whilst alternating the output layer according to the task, with the aim of learning a better hidden representation for both. The technique was, to our knowledge, first applied to ASR in [16] to integrate classification and enhancement of noisy speech with RNNs. More recently it has been applied to joint learning of triphones and trigraphemes [17], and also cross-lingual applications [18, 19]. Specifically, we propose using multitask learning to jointly learn tied state CD targets and monophone CI targets. This seems somewhat counter-intuitive, since the CD targets already encode all information about the relevant monophone, suggesting that adding monophone prediction as a related task would have little effect. However, we hypothesise that the use of a reduced set of targets will reduce over-fitting in the lower layers, whilst explicitly encouraging discrimination between CI targets, rather than CD targets – all without diminishing the power of the DNN to model the CD targets themselves. In the terminology of [15], this could be viewed as "eavesdropping": here the hidden layer is useful to both tasks, but is difficult to learn for the larger task because it depends on the hidden layer in a more complex way, or because the residual error in the larger task is noisier.

Multitask learning has been proposed as a means of learning of phone context [20]. However, this differs substantially from our proposed technique in that the basic units modelled were monophones, applied to the TIMIT phone classification task (where the use of CD units is less common). Left and right phonetic contexts were used as the additional tasks, but there was no direct modelling of tied triphone states by the DNNs.

Although we believe that there are additional advantages to the proposed technique, as a form of regularisation, it should be compared to other regularisation methods proposed for DNNs with the aim of improving generalisation. These generally take the form of schemes for initialising the weights before backpropogation over the full deep structure. Perhaps the most well-known is the use of restricted Boltzmann machine (RBM) pre-training [21], where the each successive layer is initialised with the weights of an unsupervised generative model. A second, more closely-related method is *curriculum learning* [22]: the idea here is to first present the network with "easier" examples, illustrating simpler concepts, and gradually increasing the complexity: in other words, initially training on samples from a lower-entropy distribution and increasing entropy during training. Applied to ASR, this could be implemented by training first on a monophone targets for initialising the network in a pre-training phase, before performing full training on the CD units: this is subject of recently-proposed work [23]. We present experimental comparisons with the use of RBM pre-training and curriculum learning. Regularisation based on Kullback-Liebler (KL) divergence has been recently proposed for DNN adaptation [24]. This can be interpreted as a constrained case of multitask learning, where no weights are allowed to vary independently.

3. METHODOLOGY

3.1. Baseline DNN system

Our baseline system uses feed-forward DNNs in a standard hybrid configuration, modelling frame posterior probabilities over CD units using state clustering derived from a HMM-GMM system, also used to obtain state labels for each frame.

This system used three-state cross-word triphone HMMs, tied to give approximately 6,000 CD units in total. The GMM used 16 Gaussians per state. The DNNs used 6 hidden layers with 2048 units per layer. The hidden layers use logistic sigmoid non-linearities, with a softmax function used for the output layer. The DNNs were fine-tuned using back-propagation over the full structure using the framewise cross-entropy criterion. Learning rates were reduced using the "newbob" schedule, measuring the performance of the net using frame error rate on held-out validation data. Training was performed on NVIDIA GeForce GTX 690 GPUs using an in-house tool based on the Theano library [25]. For decoding, all systems used the tied state posterior probabilities generated by the DNN, scaled by state priors estimated over the whole training data.

The experimental task we chose benefits substantially from speaker adaptation due to the relatively large quantities of data available for each test speaker. Therefore, following a procedure we have used previously [26] we used speaker-adaptive training (SAT) for the DNNs: for each training speaker, we estimated a single constrained maximum likelihood linear regression (CMLLR) transform [27] with reference to a GMM trained on the same features. These transforms were used to generate speaker-normalised features as inputs for the DNN training. At test time, the output of a speaker-independent first pass was used to estimate a CMLLR transform for each test speaker in a similar manner, which was used to generate final features for decoding with the SAT-DNNs.

3.2. Regularisation and multitask learning

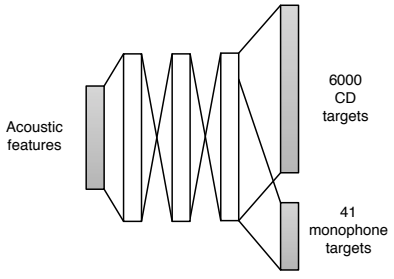

Fig. 1. Multitask network structure

For the proposed method of using the shared network to jointly predict both CD unit labels and CI labels, we used the network structure shown in Figure 1. All parameters are shared except for the weights and biases between the final hidden layer and the two output layers. This structure aimed to maximise the number of shared parameters. Note that since each CD unit can be mapped to a unique

monophone label, it would also have been theoretically possible to derive one set of output weights from the other, but we did not to do this, to avoid any over-fitting in the CD output layer from affecting the performance of the CI outputs. For the CI output layer, we used targets corresponding to the 41 monophones in the dictionary – we did not model monophone states separately. In multitask training of the network, samples for each task were presented randomly in an interleaved fashion at the minibatch level, so that over a complete epoch, each training sample was seen twice, with different targets. We did not weight the two tasks differently.

As mentioned above, we also investigated alternative, related methods of regularisation. We compared a random initialisation of the layers firstly with initialisation using layer-by-layer RBM pre-training, and secondly, with a complete round of discriminative pre-training using the monophone targets only. The latter may be viewed as a form of curriculum learning. These two techniques may also be combined with the multitask training scheme in the final finetuning.

3.3. Acoustic features

We experimented with two sets of input features to the nets. Initially, we used 13-dimensional PLP features, with first and second derivatives appended. These were normalised to zero mean and unit variance, then transformed with a per-speaker CMLLR transform as described in Section 3.1. PLP features were chosen over features such as filterbank coefficients for the ease of CMLLR adaptation. As input to the nets, we use features with 9 frames of acoustic context.

As a second set of input features, we used tandem features obtained using the multi-level adaptive networks (MLAN) scheme [28]. Following this scheme, a second DNN trained on out-of-domain (OOD) data in a conventional fashion was used to generate bottleneck features for the in-domain training data, which were appended to the original PLP features. The complete procedure is illustrated in Figure 2.

In this case, we used OOD DNNs trained on 300 hours of conversational telephone speech from the Switchboard corpus. For maximum diversity, we used filterbank coefficients here, generated using a filter to match the bandwidth of telephone speech. This proves to be a useful way of using narrowband speech data in wideband speech applications, which we will explore in more detail in future. SAT training was applied to the tandem feature vectors, and again, we use 9 frames of acoustic context.

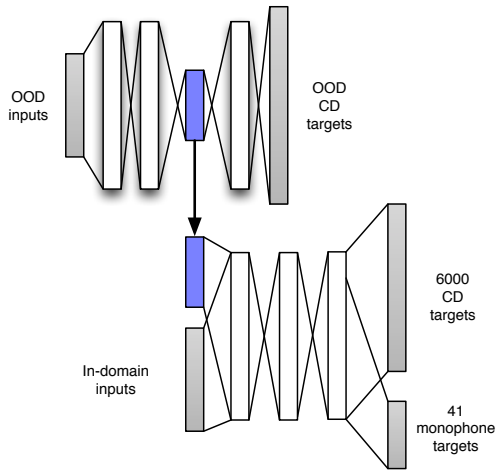

Fig. 2. Multitask network structure with MLAN features

System	Pretraining	dev2010	tst2010	tst2011	mean
Baseline	none	16.4	15.4	12.9	15.1
Baseline	RBM	16.9	15.6	13.0	15.4
Baseline	monophone	16.4	15.6	13.0	15.2
Multitask	none	16.2	15.0	12.2	14.8
Multitask	RBM	16.1	15.0	12.4	14.8
Multitask	monophone	15.9	14.9	12.3	**14.6**

Table 1. WER (%) across different test sets for systems trained on speaker-adapted PLP features

4. EXPERIMENTS

4.1. ASR task

We perform experiments on the TED English transcription task used in the IWSLT evaluation campaign [29]. 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 [30]. 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, giving 143 hours of speech for acoustic model training.

For the experiments presented here, we use a trigram LM trained on 312MW from data sources prescribed by the IWSLT evaluation, which include the corpus of TED talks, and the Europarl, News Crawl and Gigaword corpora. The vocabulary was limited to 64K words, selected according to unigram counts. Decoding was performed using HTK's HDecode, with customisations to allow the use of pseudo-likelihoods generated by DNNs. We did not perform any rescoring with a larger LM in this work.

4.2. Results

In Table 1 we present the results of various systems trained on speaker-adapted PLP features. We investigated systems with no pre-training, with RBM pre-training and with discriminative pre-training using monophone targets. For each pre-training method, we compare baseline systems trained only on CD targets with systems using the proposed multitask method. A clear trend is that multitask systems outperform the baseline, regardless of the use of pre-training, with a mean reduction of 0.5% absolute between the best baseline system and the best multitask system. The benefits of the various pre-training methods are less clear, as can be observed by the inconsistencies in performance across the different test sets. Generally, it seems that RBM pre-training is not beneficial for this size of training data. Monophone pre-training is shown to outperform RBM pre-training, but in the baseline case, no pre-training at all generally seems to be better.

In these experiments, we found that the performance of the baseline systems is quite sensitive to the initial learning rate used for the newbob schedule in finetuning, with rates towards the lower end our normal range performing particularly poorly. Following investigations on validation data, we generally set the initial learning rates of for these systems to 0.16, although exhaustive optimisation was not possible for all systems. By contrast, the multitask nets appear to be less affected by the initial learning rate, but we generally use 0.06 or 0.08 for these systems, which can be seen as compensating for the fact that training samples are presented twice during each epoch.

It is interesting to look at frame error rates (FER) on validation data held out during DNN training, shown in Table 2. With respect to the CD units, the FER actually increases when multitask training

System	Baseline CD	Multi CD	Multi Monophone
Monophone	-	-	31.9
No PT	54.3	55.4	31.5
RBM PT	54.5	55.1	30.9
Monophone PT	54.7	55.5	30.1

Table 2. Frame error rates (%) with respect to CD units and monophone units, computed on validation data for baseline and multitask ("Multi") systems trained on speaker-adapted PLP features

System	Pretraining	dev2010	tst2010	tst2011	mean
50% training data					
Baseline	none	18.9	18.3	15.3	17.8
Baseline	monophone	18.6	17.7	14.7	17.3
Multitask	none	16.9	15.9	13.1	**15.6**
Multitask	monophone	17.4	16.1	13.5	15.9
20% training data					
Baseline	none	21.4	21.0	17.4	20.3
Baseline	monophone	20.9	20.4	16.7	19.7
Multitask	none	19.0	17.9	14.7	**17.5**
Multitask	monophone	18.9	18.0	15.1	17.6

Table 3. WER (%) for systems trained on speaker-adapted PLP features with reduced data

is applied, despite the reduction in WER. This supports the theory that adding the prediction of monophone outputs as a second task as indeed improving the generalisation of the net, by preventing it overfitting to the CD targets. At the same time, the FER with respect to the monophone units is improved, compared to a system trained purely to the monophone targets, suggesting, as found in [20] that from this perspective, the use of context is a second task aids monophone prediction. Comparing with the results of Table 1 hints that monophone FER may be a better predictor of WER, even though CD outputs are used by the decoder.

Next, we investigate the regularising effects of the proposed technique by artificially reducing the quantity of training data used for all DNN training. We keep the network topology and state alignments the same. Table 2 shows the effects of reducing the quantity of data to 50% and 20% of the full set, both with and without monophone pre-training. The results are shown graphically in Figure 3. It may be seen that the proposed technique consistently outperforms the baseline here, and the gap widens significantly when less data is available. It appears that monophone pre-training yields some benefit to the baseline systems in the limited data cases, but is not as effective as using multitask training.

Finally, we present results combining multitask training with the MLAN scheme introduced in Section 3.3. We did not apply RBM pre-training on the MLAN features The use of OOD features gives further improvement, reducing the mean WER by around 1.2% in the full-data condition. The multitask training continues to yield benefits over the baseline: 0.4% on average when no pre-training is used, though this is reduced when monophone pre-training is used. It is not surprising that the difference between the systems is less in this case, since one of the effects of the MLAN scheme is to leverage much larger quantities of training data (in this case, 300 hours) via the bottleneck features input to the final DNNs. .

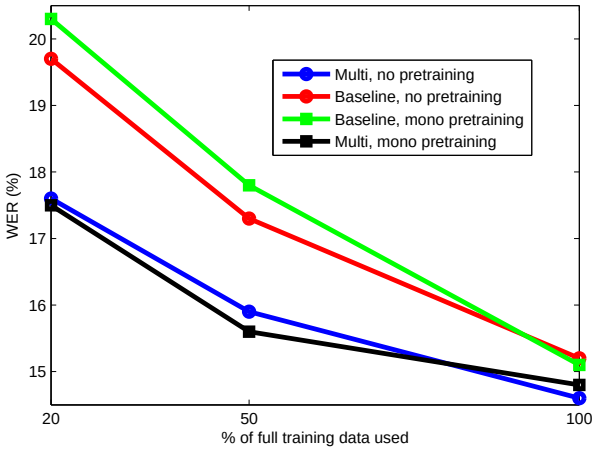

Fig. 3. WER (%) for baseline and multitask systems with and without pre-training, varying the quantity of training data used

System	Pretraining	dev2010	tst2010	tst2011	mean
Baseline	none	15.9	14.0	11.7	14.1
Baseline	monophone	15.8	13.7	11.7	13.9
Multitask	none	15.3	13.6	11.6	**13.7**
Multitask	monophone	15.6	13.5	11.5	**13.7**

Table 4. WER (%) for MLAN systems, using speaker-adapted tandem features from Switchboard DNNs

5. CONCLUSIONS

We have presented a simple, but effective of method of improving the performance of context-dependent hybrid DNN systems through the use of jointly optimising the classification performance of monophone units. This acts as a regulariser, and, we believe, encourages more relevant discrimination in the lower layers of the network. We have shown that the proposed technique yields relative performance improvements of 3%-10% over the baseline, depending on the quantity of training data available, even when the baseline networks are themselves initialised by finetuning to monophone targets.

In future, we will investigate the use of this technique with full-sequence training, where the targets used in DNN optimisation are more closely matched to the word error rate.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

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

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

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

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

[9] 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.

[10] 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.

[11] G Hinton, L. Deng, D. Yu, G.E. Dahl, A-R Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[12] Michiel Bacchiani and David Rybach, "Context dependent state tying for speech recognition using deep neural network acoustic models," in *Proc. ICASSP*, 2014, pp. 230–234.

[13] C. Zhang and P. C. Woodland, "Standalone training of context-dependent deep neural network acoustic models," in *Proc. ICASSP*, Florence, Italy, 2014.

[14] 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.

[15] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, pp. 41–75, 1997.

[16] S. Parveen and P. Green, "Multitask learning in connectionist robust ASR using recurrent neural networks," in *Proc. Interspeech*, 2003.

[17] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, "Joint acoustic modelling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proc. ICASSP*, 2014.

[18] Z. Tüske, R. Schlüter, and H. Ney, "Multi-lingual hierarchical MRASTA features for ASR," in *Proc. Interspeech*, 2013.

[19] P. Bell, J. Driesen, and S. Renals, "Cross-lingual adaptation with multi-task adaptive networks," in *Proc. Interspeech*, 2014.

[20] M. Selzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. ICASSP*, 2013.

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

[22] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, 2009.

[23] C. Zhang and P.C. Woodland, "Context independent discriminative pre-training," Unpublished work.

[24] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc ICASSP*, 2013.

[25] 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*, June 2010.

[26] P. Bell, H. Yamamoto, P. Swietojanski, Y. Wu, F. McInnes, C. Hori, and S. Renals, "A lecture transcription system combining neural network acoustic and language models," in *Proc. Interspeech*, Aug. 2013.

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

[28] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Proc. ICASSP*, 2013.

[29] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Frederico, "Report on the 10th IWSLT evaluation campaign," in *Proc. International Workshop on Spoken Language Translation*, Heidelberg, Germany, December 2013.

[30] 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.