INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Sequence Student-Teacher Training of Deep Neural Networks

Jeremy H. M. Wong and Mark J. F. Gales

Department of Engineering, University of Cambridge
Trumpington Street, CB2 1PZ Cambridge, England
jhmw2@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

The performance of automatic speech recognition can often be significantly improved by combining multiple systems together. Though beneficial, ensemble methods can be computationally expensive, often requiring multiple decoding runs. An alternative approach, appropriate for deep learning schemes, is to adopt student-teacher training. Here, a student model is trained to reproduce the outputs of a teacher model, or ensemble of teachers. The standard approach is to train the student model on the frame posterior outputs of the teacher. This paper examines the interaction between student-teacher training schemes and sequence training criteria, which have been shown to yield significant performance gains over frame-level criteria. There are several possible options for integrating sequence training, including training of the ensemble and further training of the student. This paper also proposes an extension to the student-teacher framework, where the student is trained to emulate the hypothesis posterior distribution of the teacher, or ensemble of teachers. This sequence student-teacher training approach allows the benefit of student-teacher training to be directly combined with sequence training schemes. These approaches are evaluated on two speech recognition tasks: a Wall Street Journal based task and a low-resource Tok Pisin conversational telephone speech task from the IARPA Babel programme.
Index Terms: speech recognition, student-teacher training, sequence training, combination, ensemble methods

1. Introduction

Ensemble combinations have often been found to outperform single systems in Automatic Speech Recognition (ASR) [1–4]. The performance gains are attributed to the possibility of correcting errors that occur in each system [1], reducing the likelihood of selecting a poor model, and increasing the space of possible models [5]. Ensemble methods are expected to be especially helpful when the quantity of training data is limited [5]. However, if implemented as a hypothesis-level combination, such as ROVER [1] and confusion network combination [2], the computational demand of decoding through the ensemble scales linearly with the number of systems. To address this problem in regression tasks, [6] proposes the idea of training a single student model to emulate the function learned by a teacher ensemble. This approach will be referred to as student-teacher training.

Applying student-teacher training to ASR tasks, [7] uses the KL-divergence between the student and teacher ensemble frame posteriors as the training criterion, with the introduction of a temperature to soften the distribution. Raising the temperature enhances the likelihoods of competing classes and eases the learning of the teacher posteriors. It is shown in [7] that training the student to emulate the frame posteriors of a teacher ensemble can give performance gains over standard Cross-Entropy (CE) training on the forced alignment hard targets. Frame-level student-teacher training can also be used as a form of pretraining, before fine-tuning with the standard CE criterion, and has been shown to lead to improved generalisation [8]. A related idea to student-teacher training jointly trains an ensemble with a diversity penalising term in the criterion, encouraging each system to converge toward the function of the ensemble [9].

Current student-teacher training methods in ASR have so far only trained a student to emulate the teacher's frame posteriors, and the teachers themselves have only been trained using the CE criterion [7, 8, 10, 11]. Sequence training methods [12] have been shown to produce significant gains over CE for single systems [13, 14].

This paper investigates incorporating sequence training into ensemble and student-teacher methods. Firstly, by building upon the work in [3] that investigates frame-level combinations of CE-trained ensembles, the experiments in this paper will investigate whether frame-level combination methods are able to benefit from sequence training of the ensemble. Then, an incorporation of sequence training into student-teacher training shall be assessed by first training a student to emulate the frame posteriors of a sequence-trained teacher ensemble, and then further refining the student with standard sequence training. Finally, hypothesis-level student-teacher training will be proposed, to train the student to directly emulate the hypothesis posteriors of the teacher ensemble.

2. Ensemble Student-Teacher Training

To achieve large combination gains, the systems in the ensemble need to be both diverse and individually accurate [5]. Some of the common methods to introduce diversity within the ensemble are to use different random weight initialisations, bagging [15], and random decision trees [16]. The ensemble can then be explicitly trained to encourage diversity, using methods such as Adaboost [17] and negative correlation learning [18].

In this paper, the teacher ensemble is made diverse solely by using a different random seed for the Deep Neural Network (DNN) weight initialisation of each teacher, without any explicit diversity encouragement during training. Using additional diversity methods can be expected to improve the ensemble performance. However, using just the simple technique of different random DNN weight initialisations should provide sufficient diversity to demonstrate the student-teacher training methods in this paper, particularly when working with limited quantities of training data.

Each teacher in the ensemble is first pretrained with layer-wise discriminative pretraining, and then fine-tuned with the CE

http://dx.doi.org/10.21437/Interspeech.2016-911

criterion,

$$\mathcal{F}_{\text{CE}} = -\sum_r \sum_t \log P\left(s^*_{rt}|\boldsymbol{o}_{rt}, \boldsymbol{\Phi}_m\right), \quad (1)$$

where s^*_{rt} are the state-level forced alignment hard targets, $\boldsymbol{o}_{rt}$ are the observations, r is the utterance index, t is the time step, $\boldsymbol{\Phi}_m$ are the teacher models, and m is the model index. The teachers can then be sequence-trained with either the Maximum Mutual Information (MMI) criterion,

$$\mathcal{F}_{\text{MMI}} = -\sum_r \log P\left(h^*_r|\mathbf{O}_r, \boldsymbol{\Phi}_m\right) \quad (2)$$

or the Minimum Bayes' Risk (MBR) criterion,

$$\mathcal{F}_{\text{MBR}} = \sum_r \sum_{h_r} \mathcal{L}\left(h_r, h^*_r\right) P\left(h_r|\mathbf{O}_r, \boldsymbol{\Phi}_m\right), \quad (3)$$

where h_r are the sentence hypotheses, h^*_r are the manual transcriptions, and $\mathcal{L}$ is a loss function that is taken to be the state-level minimum edit distance in this paper. Training with the state-level loss function is known as the state-level MBR (sMBR) criterion [19, 20].

An important aspect of ensemble methods is to establish a baseline. Having trained the teacher models, the ensembles in this paper can then be combined either at the frame or hypothesis level. At the frame level, the combination used is a weighted linear average of the frame posteriors,

$$P\left(s_{rt}|\boldsymbol{o}_{rt}, \boldsymbol{\Phi}\right) = \sum_{m=1}^{M} \alpha_m P\left(s_{rt}|\boldsymbol{o}_{rt}, \boldsymbol{\Phi}_m\right), \quad (4)$$

where s_{rt} are the DNN output classes, M is the ensemble size, and α_m are the teacher ensemble mixture weights, such that $\sum_m \alpha_m = 1$ and $\alpha_m \geq 0$. This is similar to using a linear ensemble with diagonal matrices in [3]. These combined frame posteriors are then passed up to the Hidden Markov Model (HMM) as observation likelihoods, and used with standard decoding. At the hypothesis level, the ensemble is decoded using MBR combination decoding [4],

$$h^*_r = \arg\min_{h'_r} \sum_{h_r} \mathcal{L}\left(h_r, h'_r\right) \sum_{m=1}^{M} \beta_m P\left(h_r|\mathbf{O}_r, \boldsymbol{\Phi}_m\right), \quad (5)$$

where here the loss function, $\mathcal{L}$, is the word-level minimum edit distance and β_m are the teacher ensemble mixture weights, such that $\sum_m \beta_m = 1$ and $\beta_m \geq 0$. This can be realised by first normalising the lattices generated by each teacher, then taking their union, and finally performing MBR decoding on the merged lattice. These combination methods are chosen, because they produce valid posterior distributions that the student model can be trained to emulate, and because they are related to the target posterior distributions in the student-teacher training methods described in sections 2.1 and 2.2.

Decoding through a frame-level combination requires one forward pass through each DNN within the ensemble. In addition to this, decoding through a hypothesis-level combination also requires generating and processing one decoding lattice for each system within the ensemble, and is therefore more computationally expensive. Decoding through a single student model greatly reduces this computational demand.

The student model used in this paper has the same architecture as one teacher within the ensemble. However, it is possible to use a student model that has a different complexity than the teachers [10], or is even of a different model type [8, 11].

2.1. Frame-level student-teacher training

Having obtained a teacher ensemble, a single student model can then be trained to emulate its attributes. The existing methods for student-teacher training aim at emulating the frame-level posteriors, by for example, minimising a weighted average of the standard CE criterion with hard targets and the KL-divergence between the student and teacher frame posteriors [7, 8, 11],

$$\mathcal{C}_{\text{CE}} = -\sum_r \sum_t \sum_{s_{rt}} P^*_{\text{CE}}\left(s_{rt}\right) \log P\left(s_{rt}|\boldsymbol{o}_{rt}, \boldsymbol{\Theta}\right), \quad (6)$$

where $\boldsymbol{\Theta}$ are the student model parameters. The target frame posteriors are

$$P^*_{\text{CE}}\left(s_{rt}\right) = (1-\lambda)\,\delta\left(s_{rt}, s^*_{rt}\right) + \lambda \sum_{m=1}^{M} \alpha_m P\left(s_{rt}|\boldsymbol{o}_{rt}, \boldsymbol{\Phi}_m\right). \quad (7)$$

The variable λ is the frame-level ensemble target weight, which sets the relative contribution of the teacher posteriors and the hard targets. This is the distribution that the student will tend toward given a sufficiently powerful model and enough training data. Therefore, in addition to the hard targets, the student model has the opportunity to learn attributes of the teacher ensemble frame posteriors. Setting $\lambda = 0$ leads to the standard CE criterion.

2.2. Hypothesis-level student-teacher training

Just as sequence training has been shown to provide performance gains over CE training for a single system [13], it is also reasonable to expect better performance from student-teacher training at the hypothesis level than the frame level. This paper therefore proposes a new student-teacher training criterion, to minimise a weighted average of the MMI criterion and the KL-divergence between the student and teacher hypothesis posteriors,

$$\mathcal{C}_{\text{MMI}} = -\sum_r \sum_{h_r} P^*_{\text{MMI}}\left(h_r\right) \log P\left(h_r|\mathbf{O}_r, \boldsymbol{\Theta}\right). \quad (8)$$

The target hypothesis posteriors are

$$P^*_{\text{MMI}}\left(h_r\right) = (1-\eta)\,\delta\left(h_r, h^*_r\right) + \eta \sum_{m=1}^{M} \beta_m P\left(h_r|\mathbf{O}_r, \boldsymbol{\Phi}_m\right). \quad (9)$$

The relative contribution between the teacher posteriors and the manual transcriptions is adjusted using the hypothesis-level ensemble target weight, η. This is again the distribution that the student will tend toward given a sufficiently powerful model and enough training data. Setting $\eta = 0$ leads to the standard MMI criterion. This proposed hypothesis-level student-teacher criterion is related to MMI, and sMBR variants may also be used, but are not investigated here.

The dynamic range of the posterior distribution has been shown to be an important aspect for sequence training [21]. To adjust the dynamic range in the hybrid ASR architecture, the hypothesis posteriors may be factorised into language and acoustic models,

$$P\left(h_r|\mathbf{O}_r, \boldsymbol{\Phi}_m\right) = \frac{P^{\kappa}\left(h_r\right) p^{\gamma}\left(\mathbf{O}_r|h_r, \boldsymbol{\Phi}_m\right)}{\sum_{h'_r} P^{\kappa}\left(h'_r\right) p^{\gamma}\left(\mathbf{O}_r|h'_r, \boldsymbol{\Phi}_m\right)}, \quad (10)$$

where κ and γ are the language and acoustic scaling factors respectively. These scaling factors can be used to alter the dynamic range of the distributions, and play a similar role to the temperature in [7].

The derivative of the criterion in Equation (8) with respect to the pre-softmax activations, $a_{s_{rt}}$, is

$$\frac{\partial \mathcal{C}_{\text{MMI}}}{\partial a_{s_{rt}}} = \gamma \sum_{h_r} \Big[P\left(h_r|\mathbf{O}_r, \boldsymbol{\Theta}\right) - P^*_{\text{MMI}}\left(h_r\right) \Big] P\left(s_{rt}|h_r, \mathbf{O}_r, \boldsymbol{\Theta}\right) \quad (11)$$

$$= \gamma \Bigg[P\left(s_{rt}|\mathbf{O}_r, \boldsymbol{\Theta}\right) - (1-\eta)\, P\left(s_{rt}|h^*_r, \mathbf{O}_r, \boldsymbol{\Theta}\right) - \eta \sum_{h_r} \sum_{m=1}^{M} \beta_m P\left(h_r|\mathbf{O}_r, \boldsymbol{\Phi}_m\right) P\left(s_{rt}|h_r, \mathbf{O}_r, \boldsymbol{\Theta}\right) \Bigg]. \quad (12)$$

As expected, Equation (11) shows that the derivative is zero when the student hypothesis posteriors are equal to the target hypothesis posteriors. To retain computational tractability, the sum needs to be limited to only over the hypotheses present in a pruned lattice. Unigram lattices are used for training, as a weaker language model allows deficiencies in the student's acoustic model to be more apparent and thus more effectively corrected [22]. Preliminary tests have also shown better performance when using unigram, rather than trigram, teacher lattices. An issue with using pruned lattices is that for any hypothesis that exists within the teacher but not the student lattice, $P\left(s_{rt}|h_r, \mathbf{O}_r, \boldsymbol{\Theta}\right) = 0$ for all s_{rt}, which is not a valid distribution, as it does not sum to one. To address this, the approximation is taken in this paper that

$$P\left(s_{rt}|h_r, \mathbf{O}_r, \boldsymbol{\Theta}\right) \approx P\left(s_{rt}|h_r, \mathbf{O}_r, \boldsymbol{\Phi}\right), \quad (13)$$

when h_r is not contained within the pruned student lattice. The combined lattice of the teacher ensemble is used to compute $P\left(s_{rt}|h_r, \mathbf{O}_r, \boldsymbol{\Phi}\right)$. This approximation does however present a mismatch between the student model and the gradient. The degree of mismatch is dependent on the probability mass allocated to the hypotheses that are disjoint between the teacher and student lattices. This can be minimised by using wider lattices. Alternatively, this issue of missing hypotheses within the pruned student lattice can be addressed by acoustically rescoring a union of the student and teacher lattices with the student observation likelihoods. This method will require further investigation.

In the current implementation of Equation (12), $P\left(s_{rt}|\mathbf{O}_r, \boldsymbol{\Theta}\right)$ and $P\left(s_{rt}|h^*_r, \mathbf{O}_r, \boldsymbol{\Theta}\right)$ are computed through standard forward-backward passes over the lattices, while $\sum_{h_r} \sum_m \beta_m P\left(h_r|\mathbf{O}_r, \boldsymbol{\Phi}_m\right) P\left(s_{rt}|h_r, \mathbf{O}_r, \boldsymbol{\Theta}\right)$ is computed using n-best lists. Using large n-best lists can include all hypotheses within the pruned lattices into the sum, but maintaining a finite list size safeguards against long computation times. There may be more efficient methods of computing the gradient, which will require further investigation. If the student lattice is determinised and not regenerated, then it is possible to pre-compute and store $\sum_{h_r} \sum_m \beta_m P\left(h_r|\mathbf{O}_r, \boldsymbol{\Phi}_m\right) P\left(s_{rt}|h_r, \mathbf{O}_r, \boldsymbol{\Theta}\right)$, thereby avoiding having to re-compute it at every training iteration. The memory requirement to store this pre-computed term is the same as that needed to store the teacher frame posteriors in frame-level student-teacher training.

3. Experiments

All experiments are realised using the Kaldi speech recognition toolkit [23], and operate on the Babel Tok Pisin (IARPA-babel207b-v1.0e) [24] and WSJ [25] datasets. The Tok Pisin Very Limited Language Pack (VLLP) is used, comprising approximately 3 hours of conversational telephone speech. This dataset contains a fairly limited quantity of training data, and should therefore benefit much from system combination. A graphemic lexicon [26] is used. The standard 10 hour development set is used for decoding with a trigram language model trained on the VLLP manual transcriptions. For WSJ, the 14 hour *si-84* training set is used, and the 64K words open vocabulary *eval92* test set is used for decoding, with the Kaldi big dictionary trigram language model, which adds additional words within the training data to the standard language models. Experimenting on these two datasets will allow an investigation of student-teacher training over different performance ranges.

Frame alignment hard targets are obtained from a Gaussian Mixture Model (GMM)-HMM. The WSJ GMM-HMM is trained following the Kaldi s5 recipe, up to tri4b. For Tok Pisin, the GMM alignments are used to train a DNN with CE, which is then used to refine the alignments. Initial tests show that realigning with the DNN is important for Tok Pisin, due to the poor quality of the GMM model. The Tok Pisin feature vectors consist of 107-dimensional tandem features [27]. The WSJ DNN uses 40-dimensional filter-bank features, appended with first and second order temporal derivatives, and a 15 frame splice context. After state clustering, the Tok Pisin and WSJ DNN outputs have 949 and 3444 targets respectively. The DNN architecture used for Tok Pisin has 4 layers with 1000 nodes per layer, while that for WSJ has 6 layers with 2000 nodes per layer. All DNNs are first initialised with layerwise discriminative pretraining, and then fine-tuned with the CE or frame-level student-teacher criterion. Sequence training is then performed, without lattice regeneration. All evaluation is done using MBR decoding, with language scaling factors of 10 for Tok Pisin and 14 for WSJ.

3.1. Ensemble training criteria and combinations

In this first experiment on Tok Pisin, an ensemble of 10 teachers are trained up to different criteria, and combined at either the frame or hypothesis level, following equations (4) and (5) respectively. Both combinations use equally weighted averages, with $\alpha_m = \frac{1}{M}$ and $\beta_m = \frac{1}{M}$ for all m. The results in Table 1 show that even with layerwise discriminative pretraining, initialising each DNN with a different random seed is still able to provide significant combination gains. As expected, sequence training improves the performance of individual teachers. These single system performance gains also result in improvements to the combined performance with both combination methods.

Table 1: WER (%) of ensemble combinations of 10 teachers trained with various criteria for Tok Pisin

Ensemble criterion	Single system WER (%) mean	best	worst	std dev	Frame combine	Hypothesis combine
CE	51.4	51.3	51.5	0.1	50.8	50.5
MMI	49.3	49.1	49.4	0.1	48.7	48.4
sMBR	48.2	48.1	48.4	0.1	47.3	47.0

For WSJ, an ensemble of 4 sMBR-trained teachers is used, with a mean single system WER of 5.09 %. Frame-level com-

bination of the WSJ ensemble gives a WER of 4.89 %, while hypothesis-level combination gives 4.84 %.

The results show that greater gains can be achieved by combining at the hypothesis level, rather than the frame level. The reason for this may be that hypothesis combination leads directly to more accurate hypothesis posteriors, while frame combination may not. Decoding depends directly on the hypothesis posteriors. The hypothesis-combined ensemble is therefore a better performing teacher ensemble to training the student on. However, it is more computationally expensive to decode a hypothesis combination than a frame combination.

3.2. Frame-level student-teacher training

This experiment examines student models trained through frame-level student-teacher training to emulate the frame posteriors of teacher ensembles trained with different criteria, and with various frame-level target weights, λ. Figure 1 shows that the gains obtained by the teacher ensembles through sequence training do transfer over to the students, improving their performance. It is interesting that as the performance of the teacher ensemble improves, the optimal λ increases. This is expected, as the system is "backing-off" to a CE criterion as λ decreases.

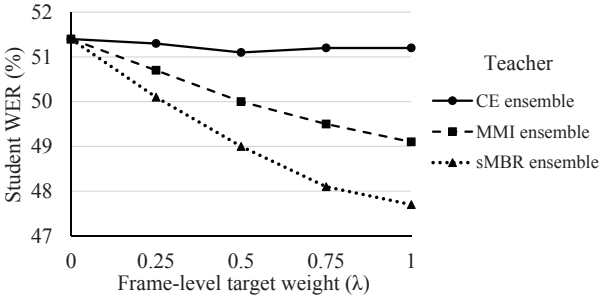

Figure 1: WER (%) of frame-level students trained with various frame-level teacher ensembles and target weights for Tok Pisin

Comparing the results in Table 1 and Figure 1, it can be seen that although the students have only been trained at the frame level, their $\lambda = 1$ WERs of 49.1 % and 47.7 %, with MMI and sMBR-trained teacher ensembles respectively, are able to outperform standard MMI and sMBR-trained models with mean WERs of 49.3 % and 48.2 %. However, all of the students in Figure 1 are still unable to match the performance of their respective teacher ensemble frame-level combinations in Table 1.

Table 2: WER (%) of sequence training on $\lambda = 1$ frame-level students initialised with frame-level sMBR teacher ensembles

Student-teacher training	Tok Pisin	WSJ
frame level	47.7	5.07
frame level + MMI	47.6	5.09
frame level + sMBR	47.2	4.94
sMBR ensemble frame combination	47.3	4.89
sMBR ensemble hypothesis combination	47.0	4.84

Although the teacher ensembles may have been sequence-trained, the students have only been trained to emulate the teachers at the frame level. To address this shortcoming, sequence training is used to refine student models that are initially trained with frame-level student-teacher training using the sMBR teacher ensembles and $\lambda = 1$. The results in Table 2 show that sequence training on the frame-level student models can provide additional gains. Training the frame-level students with the MMI criterion does not yield any significant gains, as the initial students have already been trained toward sMBR teacher ensembles, and the sMBR criterion has been shown to perform better than the MMI criterion. Further sMBR training of the frame-level students does result in improved performance, and for Tok Pisin, brings the WER to match that of the teacher ensemble frame-level combination. This suggests that the full gains of sequence training have not been carried through to the student models in frame-level student-teacher training.

3.3. Hypothesis-level student-teacher training

Table 3: WER (%) of hypothesis-level student-teacher training with the hypothesis-level sMBR teacher ensembles

Student-teacher training	η	Tok Pisin	WSJ
frame level	-	47.7	5.07
frame level + MMI	0.0	47.6	5.09
hypothesis level	0.5	47.0	4.91
hypothesis level	1.0	47.4	4.94

This final experiment compares training using the proposed MMI-based hypothesis-level student-teacher criterion with frame-level student-teacher training. Before sequence training, the student models are initialised using the same frame-level student-teacher training as in the previous experiment. The results in Table 3 show that for both $\eta = 0.5$ and 1.0, hypothesis-level student-teacher training is able to outperform frame-level student-teacher training on both datasets. The proposed hypothesis-level student-teacher training even outperforms MMI training of the frame-level student models. As with the frame level, the learned hypothesis posteriors of the teacher ensembles do aid in training the students. The proposed hypothesis-level student-teacher training is able to match the performance of sMBR training of the frame-level students in Table 2. For Tok Pisin, the $\eta = 0.5$ student model is able to match the WER of the teacher ensemble hypothesis-level combination of 47.0 %. However, it may still be possible to obtain further improvements using a hypothesis-level student-teacher criterion that is based on sMBR, rather than MMI.

4. Conclusion

This paper has investigated the interactions between student-teacher training and sequence training, and has presented a novel student-teacher training method to emulate the hypothesis posteriors, improving upon previous methods of only emulating the frame posteriors. This training method allows the teacher ensemble to be constructed through hypothesis-level combination, which has been shown to perform better than frame-level combination. The experiments demonstrate that the gains from sequence training of the teacher ensemble can be emulated by the frame-level student. Further sequence training of the frame-level student can bring additional gains. Training the student to match the teacher posteriors at the hypothesis level has been shown to perform better than at the frame level, even with further MMI training of the frame-level student.

The future work will investigate methods to incorporate the sMBR criterion into student-teacher training.

5. References

[1] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recogniser output voting error reduction (ROVER)," in *ASRU*, Santa Barbara, USA, Dec 1997, pp. 347–354.

[2] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Speech Transcription Workshop*, vol. 27, 2000.

[3] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *INTERSPEECH*, Singapore, Sep 2014, pp. 1915–1919.

[4] 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, Oct 2011.

[5] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, Cagliari, Italy, Jun 2000, pp. 1–15.

[6] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, Philadelphia, USA, Aug 2006, pp. 535–541.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Deep Learning and Representation Learning Workshop, NIPS*, Montréal, Canada, Dec 2014.

[8] Z. Tang, D. Wang, and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," in *ICASSP*, Shanghai, China, Mar 2016, pp. 5900–5904.

[9] X. Zhang, D. Povey, and S. Khudanpur, "A diversity-penalizing ensemble training method for deep learning," in *INTERSPEECH*, Dresden, Germany, Sep 2015, pp. 3590–3594.

[10] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *INTERSPEECH*, Singapore, Sep 2014, pp. 1910–1914.

[11] K. J. Geras, A.-R. Mohamed, R. Caruana, G. Urban, S. Wang, O. Aslan, M. Philipose, M. Richardson, and C. Sutton, "Blending LSTMs into CNNs," in *ICLR (workshop track), preprint*, San Juan, Puerto Rico, May 2016.

[12] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 14–36, Sep 2008.

[13] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *ICASSP*, Taipei, Apr 2009, pp. 3761–3764.

[14] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *INTERSPEECH*, Lyon, France, Aug 2013, pp. 2345–2349.

[15] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug 1996.

[16] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *ICASSP*, Philadelphia, USA, Mar 2005, pp. 197–200.

[17] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," AT&T Bell Laboratories, Tech. Rep., 1995.

[18] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural Networks*, vol. 12, no. 10, pp. 1399–1404, Dec 1999.

[19] M. Gibson and T. Hain, "Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition," in *INTERSPEECH*, Pittsburgh, USA, Sep 2006, pp. 2406–2409.

[20] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to MPE for large scale discriminative training," in *ICASSP*, Honolulu, USA, Apr 2007, pp. IV321–IV324.

[21] D. Povey and P. C. Woodland, "Improved discriminative training techniques for large vocabulary continuous speech recognition," in *ICASSP*, Salt Lake City, USA, May 2001, pp. 45–48.

[22] R. Schlüter, B. Müller, F. Wessel, and H. Ney, "Interdependence of language models and discriminative training," in *ASRU*, Keystone, USA, Dec 1999, pp. 119–122.

[23] 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 *ASRU*, Hawaii, USA, Dec 2011.

[24] M. P. Harper, "IARPA babel program," http://www.iarpa.gov/index.php/research-programs/babel, 2011.

[25] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Workshop on Speech and Natural Language*, Harriman, USA, Feb 1992, pp. 357–362.

[26] M. J. F. Gales, K. M. Knill, and A. Ragni, "Unicode-based graphemic systems for limited resource languages," in *ICASSP*, South Brisbane, Australia, Apr 2015, pp. 5186–5190.

[27] H. Wang, A. Ragni, M. J. F. Gales, K. M. Knill, P. C. Woodland, and C. Zhang, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *INTERSPEECH*, Dresden, Germany, Sep 2015, pp. 3660–3664.