Joint Sequence Training of Phone and Grapheme Acoustic Model based on Multi-task Learning Deep Neural Networks

Dongpeng Chen[1], *Brian Mak*[1], *Sunil Sivadas*[2]

[1]Department of Computer Science & Engineering
Hong Kong University of Science & Technology
[2]Institute for Infocomm Research, A*STAR, Singapore
{dpchen,mak}@cse.ust.hk, sivadass@i2r.a-star.edu.sg

Abstract

Multi-task learning (MTL) can be an effective way to improve the generalization performance of singly learning tasks if the tasks are related, especially when the amount of training data is small. Our previous work applied MTL to the joint training of triphone and trigrapheme acoustic models using deep neural networks (DNNs) for low-resource speech recognition. Significant recognition improvement over the performance of their DNNs trained by single-task learning (STL) was obtained. In that work, both STL-DNNs and MTL-DNNs were trained by minimizing the total frame-wise cross entropies. Since phoneme and grapheme recognition are inherently sequence classification tasks, here we study the effect of sequence-discriminative training on their joint estimation using MTL-DNNs. Experimental evaluation on TIMIT phoneme recognition shows that joint sequence training outperforms frame-wise training of phone and grapheme MTL-DNNs significantly.

Index Terms: sequence training, phone modeling, grapheme modeling, multi-task learning, deep neural networks

1. Introduction

To address the problem of limited speech and language resources in low-resource automatic speech recognition (ASR), a multi-task learning (MTL) approach was taken in our previous work [1]. Unlike other popular approaches that make use of cross-lingual [2, 3] or multi-lingual [4] information to improve acoustic modeling of a low-resource language, our MTL approach does not require resources from languages other than the target language, nor a good mapping between its phonemes and phonemes from other languages which is sometimes not easy to find. In [1], we make use of the fact that phone modeling and grapheme modeling are highly related learning tasks, and estimate triphone acoustic models and trigrapheme acoustic models of the same language together using a single deep neural network (DNN) [5]; we call the resulting DNN, MTL-DNN. During MTL estimation of the phoneme and grapheme models, only the orthographic transcriptions of the training speech and a phonetic dictionary of the target language (which phonetic acoustic modeling already uses) are required. The MTL-DNN is trained by minimizing the total frame-wise cross entropy. Experimental evaluation of our MTL-DNN approach on three low-resource South African languages shows that their MTL-DNN outperforms both of their triphone DNN and trigrapheme DNN that are singly learned — STL-DNN, and even the ROVER combination of the two STL-DNNs.

In [1], the MTL-DNNs are trained by minimizing the total frame-wise cross entropy criterion. However, speech recognition is essentially a sequential labeling problem. The frame-wise criterion does not capture the long term correlation among the target classes in an utterance. On the other hand, sequence-discriminative training has been an indispensable step in building state-of-the-art ASR systems that are based on hidden Markov models (HMMs) with state output probability distributions estimated using Gaussian mixture model (GMMs). Recently, sequence-discriminative training has been extended to DNN training using different training criteria, such as minimum Bayes risk (MBR) [6], minimum phone error (MPE) [7], maximum mutual information (MMI) [8] and boosted MMI (BMMI) [9]. Consistent improvements are reported on both phoneme recognition [10] and large-vocabulary ASR [11, 12, 13]. In this paper, we further explore joint sequence-discriminative training of both phone and grapheme acoustic models under the MTL-DNN framework. That is, for each training utterance, we have to produce both a phone lattice as well as a grapheme lattice, compute the sequence-discriminative training error from each of them, and propagate these error signals back to the MTL-DNN to its weights under the MTL framework.

The rest of this paper is organized as follows. In the next section, the concepts of multi-task learning deep neural network and joint phone and grapheme acoustic modeling are reviewed. Then in Section 3, we describe the proposed joint sequence training of phone and grapheme acoustic models using a DNN in the MTL framework. Experimental evaluation are presented in Section 4, followed by concluding remarks in Section 5.

2. Joint phone and grapheme acoustic modeling using MTL-DNN

2.1. Multi-task learning deep neural network (MTL-DNN)

Multi-task learning (MTL) [14] or learning to learn [15] aims at improving the generalization performance of a learning task by jointly learning multiple *related tasks*. The multiple tasks share some internal representation, so that their learned knowledge can be transfered among each other. In fact, multi-task learning is effectively a regularization method that may alleviate overfitting, and is more effective when the amount of training data is small. MTL can be readily implemented by artificial neural networks (ANN) in which the weights are used as the common representation of learned knowledge shared across multiple tasks. In fact, MTL has been applied successfully to the training of ANNs in many learning tasks in fields of speech, language, and

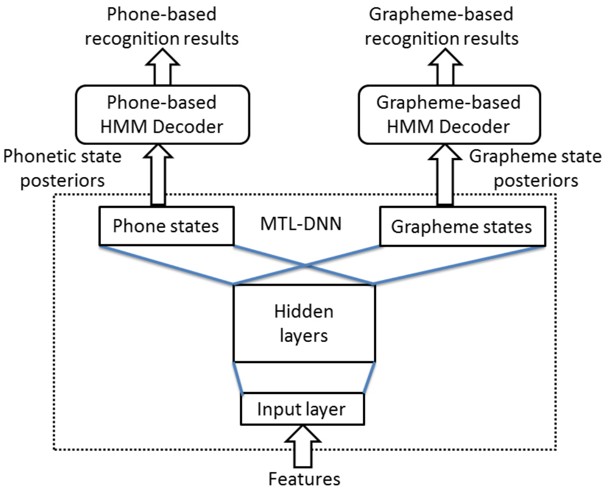
Phone-based recognition results
Grapheme-based recognition results
Phone-based HMM Decoder
Grapheme-based HMM Decoder
Phonetic state posteriors
Grapheme state posteriors
Phone states
MTL-DNN
Grapheme states
Hidden layers
Input layer
Features

Figure 1: An MTL-DNN system for the joint training of phone and grapheme acoustic models.

image/vision. For example, in ASR, MTL is used to improve ASR robustness using recurrent neural networks in [16]. In language applications, [17] applies MTL on a single convolutional neural network to produce state-of-the-art performance for several language processing predictions; [18] improves intent classification in goal-oriented human-machine spoken dialog systems especially when the amount of labeled training data is limited. In [19], the MTL approach is used to perform multi-label learning in an image annotation application.

MTL has been extended to training the popular deep neural networks (DNNs) to further improve learning performance. Related works in the area of ASR include the use of MTL-DNN for TIMIT phoneme recognition [20] which learns posteriors of monophone states together with a secondary task that can be learning phone labels, state contexts, or phone contexts. MTL-DNN is also used in multi-lingual ASR to transfer cross-lingual knowledge [21, 22].

2.2. Joint phone and grapheme acoustic modeling

Fig.1 shows an overview of the MTL-DNN system for joint training of phone and grapheme acoustic models in our previous work [1]. Essentially two single-task learning DNNs (STL-DNNs), one for training the posterior probabilities of phone states and the other for training the posterior probabilities of grapheme states are merged so that their input and hidden layers are shared, while each of them keeps its own output layer. Although the DNN architecture looks similar to the one used in multi-lingual speech recognition works [21, 22] mentioned above, there is a subtle difference between our MTL procedure and theirs. In these works, each of the multiple languages has its own output layer (for its own tied states); when the training samples of language, say, L are presented to the DNN, only the output layer of language L is trained but not the output layers of the other co-training languages. On the other hand, in our work, for each input training sample, it is propagated through all the hidden layers to the output layers of both phone states and grapheme states. More specifically, given an input vector $\mathbf{x}$, the posterior probability of the phone output layer's ith phone state s_{ip} is computed using the softmax function as follows:

$$P(s_{ip}|\mathbf{x}) = \frac{\exp(y_{ip})}{\sum_{i'=1}^{N_p}\exp(y_{i'p})}\,, \quad \forall i = 1, \ldots, N_p\,,$$

where y_{ip} is the activation of the state, and N_p is the total number of phone states. A similar formula may be derived for the posterior probabilities $P(s_{ig}|\mathbf{x})$ of the N_g grapheme states at the grapheme output layer. Finally, the whole MTL-DNN is trained by minimizing the sum of cross-entropies from the two tasks over all frames:

$$F_{ce} = \sum_{\mathbf{x}}\left(\sum_{i=1}^{N_p} d_{ip}\log P(s_{ip}|\mathbf{x}) + \sum_{i=1}^{N_g} d_{ig}\log P(s_{ig}|\mathbf{x})\right)\,,$$

where d_{ip} and d_{ig} are the target values of the ith phone state and the ith grapheme state respectively.

Before the joint training of phone and grapheme acoustic models, one first trains the conventional GMM-HMMs for the phones and graphemes. The phone and grapheme states in the output layers of the MTL-DNN are obtained from their corresponding GMM-HMM systems. The phone and grapheme GMM-HMMs are also utilized to obtain the initial frame labels of the training speech by forced alignment. During MTL-DNN training, the target values of exactly one phone state in the phone output layer and one grapheme state in the grapheme output layer will be set to 1.0, while the target values of all the remaining output units will be zero. During recognition, the MTL-DNN posterior probabilities of the phone states or grapheme states are fed into their respective decoders and afterward, Viterbi decoding is performed on their respective MTL-DNN-HMMs. In addition, one may combine the recognition results from the phone-based decoder and the grapheme-based decoder using, e.g., ROVER [23], to obtain a better performance.

3. Joint sequence training of phone and grapheme acoustic model

The joint training of phone and grapheme acoustic models using an MTL-DNN described in the last Section is found effective [1]. Nevertheless, the optimization criterion of minimizing the total frame-wise cross-entropies does not take into account the correlation between neighboring frames. Since sequence-discriminative training has been applied successfully to STL-DNN [10, 11], we would like to further investigate the effectiveness of joint sequence-discriminative training of both phone and grapheme acoustic models using an MTL-DNN. Moreover, since it has been shown in [11] that the various discriminative training criteria give similar performance, we simply choose the minimum phone error (MPE) criterion for the phone-based decoder, and the minimum grapheme error (MGE) criterion for the grapheme-based decoder. Hence, the joint sequence-discriminative training criterion of our MTL-DNN is to minimize the sum of phone errors and grapheme errors as follows:

$$\begin{aligned} F_{mpge} &= F_{mpe} + F_{mge} \\ &= \sum_u \left(\frac{\sum_{W_p} P(\mathbf{O}^{(u)}|W_p)^{\kappa_p} P(W_p) A(W_p, W_p^{(u)})}{\sum_{W_p'} P(\mathbf{O}^{(u)}|W_p')^{\kappa_p} P(W_p')} \right. \\ &\quad \left. + \frac{\sum_{W_g} P(\mathbf{O}^{(u)}|W_g^{(u)})^{\kappa_g} P(W_g) A(W_g, W_g^{(u)})}{\sum_{W_g'} P(\mathbf{O}^{(u)}|W_g')^{\kappa_g} P(W_g')} \right)\,, \end{aligned}$$

where $W_p^{(u)}$ and $W_g^{(u)}$ are the true phonetic and graphemic transcriptions of the utterance u; $\mathbf{O}^{(u)} = \{\mathbf{o}_1^{(u)}, \mathbf{o}_2^{(u)}, ..., \mathbf{o}_{T_u}^{(u)}\}$ is its acoustic observation sequence; $A(W_p, W_p^{(u)})$ is the phonetic transcription accuracy of the utterance defined as the num-

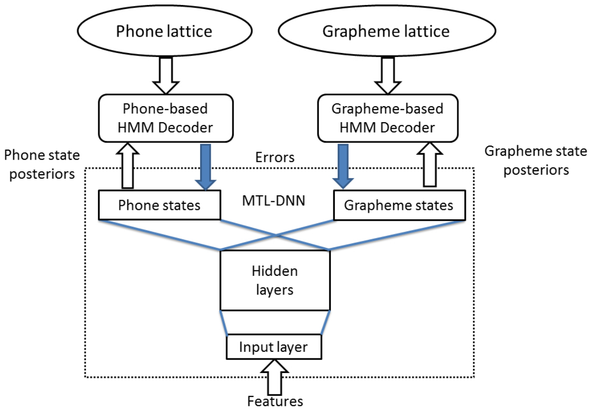

Figure 2: Joint sequence training of phone and grapheme MTL-DNNs.

ber of correct phone labels in $W_p^{(u)}$ minus the number of errors in the hypothesis W_p; $P(W_p)$ is the probability of W_p given by the lattice. The graphemic transcription accuracy $A(W_g, W_g^{(u)})$ is defined in a similar way. κ_p and κ_g are the likelihood scales used in MPE and MGE training respectively.

Taking the derivative of F_{mpge} w.r.t. $\log p(\mathbf{o}_t|s)$, we obtain, for the phone state s in phone a,

$$\frac{\partial F_{mpge}}{\partial \log P(\mathbf{o}_t^{(u)}|s)} = \kappa_p \gamma_{p,t}^{den(u)}(s)\left(\bar{A}_p^{(u)}(s(t) \in \mathbf{S}_a) - \bar{A}_p^{(u)}(*)\right)$$

where $\mathbf{S}_a$ is the set of states of phone a; $\bar{A}_p^{(u)}(*)$ is the average accuracy of all the paths in the lattice of utterance u; $\bar{A}_p^{(u)}(s(t) \in \mathbf{S}_a)$ is the average accuracy of those paths going through phone a at time t in the phone lattice; $\gamma_{p,t}^{den(u)}(s)$ is the posterior probability that at time t the utterance u reaches state s, and is calculated by the extended Baum-Welch algorithm using the phone denominator lattice. Similarly,

$$\frac{\partial F_{mpge}}{\partial \log P(\mathbf{o}_t^{(u)}|s)} = \kappa_g \gamma_{g,t}^{den(u)}(s)\left(\bar{A}_g^{(u)}(s(t) \in \mathbf{S}_b) - \bar{A}_g^{(u)}(*)\right)$$

for grapheme state s in grapheme b. Note that the phone lattice and grapheme lattice of the same utterance are disjoint.

An overview of the sequence training procedure is shown in Fig. 2. Firstly, an MTL-DNN is trained by minimizing the total frame-wise cross-entropies. Then the well-trained MTL-DNN is used to produce both the phone and the grapheme state posteriors of each training utterance. The phone posteriors are used by the phone-based decoder to generate the phone denominator and numerator lattices for the utterance, while the grapheme state posteriors are used by the grapheme-based decoder to generate the grapheme denominator and numerator lattices separately. Finally, the following procedure is repeated for each utterance u in the data set:

STEP 1 : Acoustic features of the whole utterance are again fed into the MTL-DNN to produce the posteriors of the phone and grapheme states.

STEP 2 : The two phone-based and grapheme-based decoders take in the corresponding state posteriors and compute the respective MPE and MGE statistics and the required gradients using the extended Baum-Welch algorithm.

STEP 3 : The weights of the MTL-DNN are updated by back-propagating the combined MPE and MGE errors from the two decoders through the hidden layers to the bottom layer.

4. Experimental evaluation

4.1. The TIMIT speech corpus

The standard NIST training set which consists of 3,696 utterances from 462 speakers was used to train the various models, whereas the standard core test set which consists of 192 utterances spoken by 24 speakers was used for evaluation. The development set is part of the complete test set, consisting of 192 utterances spoken by 24 speakers. Speakers in the training, development, and test sets do not overlap.

We followed the standard experimentation on TIMIT, and collapsed the original 61 phonetic labels in the corpus into a set of 48 phones for acoustic modeling; the latter were further collapsed into the standard set of 39 phones for error reporting. Moreover, the glottal stop [q] was ignored. At the end, there are altogether 15,546 cross-word triphone HMMs based on 48 base phones. Phone recognition was performed using Viterbi decoding with a phone bigram language model (LM) that was trained from the TIMIT training transcriptions using the SRILM language modeling toolkit. The phone bigram LM has a perplexity of 16.44 on the core test set.

A grapheme recognition task is designed as the secondary task. The 26 English alphabets are used as labels and word transcriptions in the data set are expanded to their grapheme sequences. We estimated a grapheme bigram LM again from the transcriptions of the training data; it has a perplexity of 22.79 on the core test set.

4.2. Feature extraction and system configurations

4.2.1. GMM-HMM baselines

39-dimensional acoustic feature vectors consisting of the first 13 MFCC coefficients, including c0, and their first and second order derivatives were extracted at every 10ms over a window of 25ms from each utterance. Then, conventional strictly left-to-right 3-state continuous-density hidden Markov models were trained by maximum-likelihood estimation. State output probability densities were modeled by Gaussian mixture models with at most 16 components.

4.2.2. STL-DNN training by minimizing frame-wise cross-entropy

Deep neural network (DNN) systems were built using 40-dimensional log filter-bank features and the energy coefficient as well as their first- and second-order derivatives. Single-task learning (STL) DNNs were trained to classify the central frame of each 15-frame acoustic context window. Feature vectors in the window were concatenated and then normalized to have zero mean and unit variance over the whole training set. All DNNs in our experiments had 4 hidden layers with 2048 nodes per layer. During pre-training, the mini-batch size was kept at 128, and a momentum of 0.5 was employed at the beginning which was then grown to 0.9 after 5 iterations. For Gaussian-Bernoulli restricted Boltzmann machines (RBMs), training kept going for 220 epochs with a learning rate of 0.002, while Bernoulli-Bernoulli RBMs were trained for 100 iterations with a learning rate of 0.02. After pre-training, a softmax layer was added on top of the deep belief network (DBN). The targets were derived from the tied states of the respective GMM-HMM baseline models. The whole network was fine-tuned by minimizing the frame-wise cross-entropy with a learning rate starting at 0.02 which was subsequently halved when performance gain on the validation set was less than 0.5%. Training contin-

Table 1: Recognition performance of various phone- and grapheme-based ASR systems in terms of phone error rate (PER) and grapheme error rate (GER).

MODEL	**PER** (%)	**GER** (%)
GMM	28.20	42.64
STL-DNNs (CE)	22.22	38.42
STL-DNNs (MPE / MGE)	21.68	37.79
MTL-DNN (CE)	21.59	36.93
MTL-DNN (MPGE)	21.01	36.52

ued for at least 10 iterations and was stopped when the classification error rate on the development set started to increase.

4.2.3. MTL-DNN training by minimizing frame-wise cross-entropy

An MTL-DNN was initialized by the same DBN used to initialize the training of STL-DNNs. However, the single softmax output layer in STL-DNNs was now replaced by two separate softmax layers, one for the primary phoneme recognition task, and the other one for the grapheme recognition secondary task. During training, two targets, one for each of the two tasks, were activated at the same time. We used the same global learning rate for the output layer, but since there were two tasks now, the learning rate for the hidden layers were halved. Otherwise, the training procedure of MTL-DNN is the same as that of STL-DNN.

4.2.4. Sequence-discriminative training of DNNs

STL-DNN or MTL-DNN trained by minimizing the total frame-wise cross-entropies was employed to generate the numerator and denominator lattices for its own sequence training. The denominator lattice were obtained by performing 30-best recognition using the HTK toolkit. Afterwards, sequence training was performed on top of the well-trained STL-DNN or MTL-DNN by following the procedure described in Section 3. It was empirically found that sequence training of STL-DNN might well be started with a small global learning rate of 1e-5, but sequence training of MTL-DNN required a larger learning rate of 1e-4 to start. This may indicate that the parameter update of joint sequence training of MTL-DNN is more stable so that a larger learning rate may be used. Training continued for at least 5 iterations with learning rate halving, and stopped if no further improvement was observed. In joint sequence training, the likelihood scales and insertion penalties of both tasks were tuned to obtain the least phone error rate on the development set.

During decoding, the insertion penalty was fixed to 0 and the grammar factor was fixed to 1 for all DNN systems.

4.3. Experimental results

The recognition performance of various acoustic models on TIMIT phonemes and graphemes are listed in Table 1. We have the following observations:

- Compared to English phoneme recognition, English grapheme recognition is much more difficult. Although in the English grapheme recognition task, there are only 26 graphemes/letters to distinguish, the grapheme bigram LM has a higher perplexity of 22.79! As a result, all the grapheme-based recognition systems have high GERs of around 40%. This is expected as there is a very complicated relationship between English pronunciation and its written form.
- The hybrid DNN-HMM systems greatly reduce the PER or GER of their GMM-HMM counterparts. For example, the phone STL-DNN trained by minimizing the total frame-wise cross-entropies reduces the PER by 21% relative, while a similarly trained grapheme STL-DNN reduces the GER by 10% relative.
- Both STL-DNNs are further improved by sequence-discriminative training. MPE training reduces the PER by 0.54% absolute, which is close to the results of MMI training in [10].
- The STL-DNNs can also be improved by multi-task learning. Regardless of the use of frame-wise cross-entropy criterion or sequence-discriminative training criterion, MTL-DNNs can reduce the PER of their STL-DNN counterparts by about 0.6% absolute, which is even greater than the PER reduction obtained by sequence training of STL-DNNs.
- Although MTL-DNN training was stopped according to its phoneme recognition performance on a separate development set, one can see that multi-task learning not only benefits the phone models, but also the grapheme models. The evidence comes from the improved GER of the MTL-DNNs over the corresponding STL-DNNs.
- Joint sequence-discriminative training of MTL-DNN gives the best phoneme recognition performance. The absolute gain is 1.21% (or relatively 5.5%) when compared to the STL-DNN baseline, and 0.58% (or relatively 2.6%) when compared to the MTL-DNN trained on minimizing the frame-wise cross-entropy.

5. Conclusions

Although graphemic acoustic models do not give good recognition performance in English due to the highly complicated relationship between English pronunciation and its writing, we show that they still can be utilized to improve the estimation of phonetic acoustic models in the multi-task learning framework. We further study the effect of joint sequence-discriminative training on MTL-DNN. The MTL-DNN is trained with error signals from multiple sequential labeling tasks. Experiment results show that sequence-discriminative training is able to further improve frame-wise cross-entropy training of MTL-DNNs. We will analyze how the auxiliary grapheme knowledge alleviates the confusion among phonemes and how the phoneme knowledge is able to resolve some of the complicated mappings from acoustic features to graphemes.

6. Acknowledgments

We would like to thank Karel Vesely of Brno University of Technology for his help with the use of TNet[1] and example MPE scripts for sequence training in this paper and Cheung-Chi Leung of Institute for Infocomm Research, A*STAR for his comments. This work was supported by the Research Grants Council of the Hong Kong SAR under the grant numbers HKUST616513 and HKUST16206714.

[1] http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet.

7. References

[1] D. Chen, B. Mak, C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014.

[2] K. U. Ogbureke and J. Carson-Berndsen, "Framework for cross-language automatic phonetic segmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 5266–5269.

[3] V. Le and L. Besacier, "Automatic speech recognition for under-resourced languages: Application to Vietnamese language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 1471–1482, 2009.

[4] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proceedings of the International Conference on Spoken Language Processing*, 1996.

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

[6] J. Kaiser, B. Horvat, and Z. Kacic, "A novel loss function for the overall risk criterion based discriminative training of HMM models," in *Proceedings of the International Conference on Spoken Language Processing*, 2000.

[7] D. Povey, "Discriminative training for large vocabulary speech recognition," *Cambridge, UK: Cambridge University*, vol. 79, 2004.

[8] L. Bahl, P. Brown, P. V. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 11. IEEE, 1986, pp. 49–52.

[9] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2008, pp. 4057–4060.

[10] A.-r. Mohamed, D. Yu, and L. Deng, "Investigation of full-sequence training of deep belief networks for speech recognition." in *Proceedings of Interspeech*, 2010, pp. 2846–2849.

[11] K. Vesel̀y, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of Interspeech*, 2013, pp. 2345–2349.

[12] H. Su, G. Li, D. Yu, and F. Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription." in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 6664–6668.

[13] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2009, pp. 3761–3764.

[14] R. Caruana, "Multitask learning," Ph.D. dissertation, Carnegie Mellon University, USA, 1997.

[15] S. Thrun and L. Pratt, *Learning to Learn*. Kluwer Academic Publishers, November 1997.

[16] S. Parveen and P. D. Green, "Multitask learning in connectionist ASR using recurrent neural networks," in *Proceedings of the European Conference on Speech Communication and Technology*, 2003, pp. 1813–1816.

[17] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the International Conference on Machine Learning*. ACM, 2008, pp. 160–167.

[18] G. Tur, "Multitask learning for spoken language understanding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. 585–588.

[19] Y. Huang, W. Wang, L. Wang, and T. Tan, "Multi-task deep neural network for multi-label learning," in *Proceedings of the IEEE International Conference on Image Processing*, 2013, pp. 2897–2900.

[20] M. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 6965–6968.

[21] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013, pp. 7304 – 7308.

[22] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep-neural networks," in *Proc. ICASSP*, 2013, pp. 7319–7323.

[23] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.