Interspeech 2018
2-6 September 2018, Hyderabad

Multilingual Deep Neural Network Training using Cyclical Learning Rate

Andreas Kirkedal, Yeon-Jun Kim

Interactions LLC, USA
{akirkedal,ykim}@interactions.com

Abstract

Deep Neural Network (DNN) acoustic models are an essential component in automatic speech recognition (ASR). The main sources of accuracy improvements in ASR involve training DNN models that require large amounts of supervised data and computational resources. While the availability of sufficient monolingual data is a challenge for low-resource languages, the computational requirements for resource rich languages increases significantly with the availability of large data sets.

In this work, we provide novel solutions for these two challenges in the context of training a feed-forward DNN acoustic model (AM) for mobile voice search. To address the data-sparsity challenge, we bootstrap our multilingual AM using data from languages in the same language family. To reduce training time, we use cyclical learning rate (CLR) which has demonstrated fast convergence with competitive or better performance when training neural networks on tasks related to text and images.

We reduce training time for our Mandarin Chinese AM with 81.4% token accuracy from 40 to 21.3 hours and increase the word accuracy on three romance languages by 2-5% with multilingual AMs compared to monolingual DNN baselines.

Index Terms: speech recognition, multilingual, cyclical learning rate

1. Introduction

State-of-the-art hybrid ASR systems consists of a neural network AM, a pronunciation model (PM), and a language model (LM) which are built or trained separately. The AM is typically a feed-forward (FF), long short-term memory (LSTM), convolutional neural network (CNN) or a mixture. Hybrid systems with LSTM AMs tend to provide state-of-the art performance on public benchmarks, but FF AMs are still used in production systems because they provide good performance and are easier to train.

Since DNN-HMM systems are complex and rely on hand-crafted lexicons, all-neural end-to-end (E2E) and lexicon-free approaches to ASR have received a lot of attention in the research community because they simplify the ASR pipeline. While interesting, these approaches suffer from a few drawbacks that make them impractical. Attention-based models such as ”Listen-Attend-Spell” require the entire sequence as input which makes it impossible to stream audio to the ASR system and show incremental output [1]. E2E ASR systems trained with connectionist temporal classification (CTC) do not reach the same performance and require orders of magnitude more data to train reliably if the E2E systems directly outputs words [2]. For these reasons, hybrid DNN-HMM systems are still widely used despite being more complex.

Multilingual ASR has been investigated in low-resource settings for decades and generally fall into two categories: data augmentation (multilingual) and model adaptation (cross-lingual) [3, 4, 5, 6]. In [7, 8], data from related languages improve phone modelling by providing better coverage of phonetic contexts. However, [9] shows multilingual data can have adverse effects on performance compared to monolingual training if there is sufficient data available. We show that data from the same language family can improve ASR accuracy in resource-rich languages and minimize negative effects from the differences between the training languages by carefully constructing a universal phoneme mapping to guide tree-clustering.

Language-independent data augmentation techniques that perturb your data set or apply different types of noise have been proposed to create more training data [10, 11, 12, 13, 14]. We do not use these techniques in our work, but they could be applied as a pre-processing step.

A cyclical learning rate schedule similar to CLR is SGD with warm restarts (SGDR) [15]. In SGDR, the learning rate starts at a high rate and anneals to a low learning rate at the end of a cycle which means the learning rate jumps directly from low to high learning rate. SGDR should force SGD to jump out of one minimum and converge to a new minimum that potentially improve accuracy. Because SGDR should discover a number of different minima, it is possible to ensemble the models at each minimum and realise better performance in some cases [16]. In our experiments, we did not achieve good results with SGDR and [17] observe that the jump from a small learning rate directly to a large learning rate can cause training error to spike and make convergence more difficult especially in the beginning of training when the network weights change rapidly between updates. They suggest that a *warm-up* phase is necessary and this is built into CLR, but not SGDR.

Adaptive learning rates like ADAM [18], Adagrad [19] and Adadelta [20] have been proposed as an alternative to annealing schedules, but [21] show that adaptive learning rates can converge to sharp minima that do not generalise well. The high learning rates used in CLR helps SGD to skip over sharp minima and converge to a wide minimum.

[22] investigates how to reduce training time for multilingual ASR and achieve a training time reduction of 46-65% with same or slightly improved word error rate (WER). The reduction comes from multilingual initialisation and using a bottle-neck feature extractor. The accelerated training speed is gained from bottle-neck features which is orthogonal to our work and could potentially be combined.

Our contributions in this paper are

- A method to use multilingual data to improve ASR performance in resource-rich languages with data from the same language family.
- We show that CLR can reduce DNN AM training time when computational resources are limited and data sets are large.

10.21437/Interspeech.2018-1891

2. Multilingual DNN AMs

Multilingual DNN AM training is a good example of multi-task learning (MTL) [23, 24] which aims to train a model with data from related tasks to improve performance on a target task. Figure 1 shows a typical multilingual DNN architecture using hard parameter sharing in the hidden layers and language-specific softmax outputs.

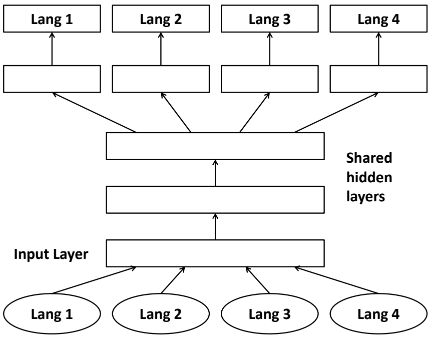

Figure 1: *Typical multilingual DNN AM with hard parameter sharing*

In our work, we adopt an architecture with an embedding layer which allows sharing a single softmax output layer across all the languages. To train our multilingual DNN AM, we first bootstrap monolingual DNN AMs from GMM AMs in each language. We then relabel the training data and use tree-based clustering and state-tying to create a new label set. The language-specific phoneme sets use symbols from SAMPA which we use to create a cross-lingual phoneme mapping. We use this phoneme mapping and phonetic features in addition to filterbank coefficients to guide the clustering. The state labels become the multilingual softmax output nodes and the phonetic context-tree is used to estimate language-specific PMs and LMs. Because we can use a single softmax output layer, we can train as we do in the monolingual case and the linear transform resembles a cross-lingual label embedding layer as shown in Figure 2. The number of output nodes serving each language varies depending on the number of phonemes in the language and the size of training data. We randomise training utterances across languages before feeding them into DNN training.

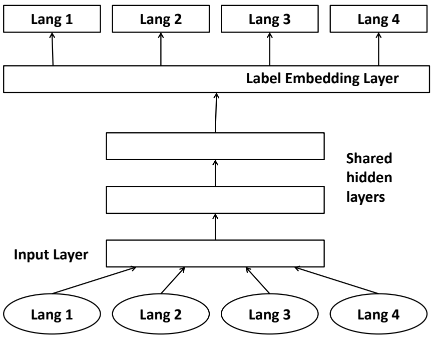

Figure 2: *Multilingual DNN AM with embedding layer*

We observe that our multilingual AMs consistently improve word accuracy by 2-5% compared to monolingual DNN AMs, but the training time is also increased by the number of training utterances [25]. It took a month to train a multilingual DNN AM on 1000 hours of speech on a NVIDIA K20 GPU and 10 days on a Pascal 1080 GPU in our experiments. Therefore, we explore the possibility to accelerate multilingual DNN training with cyclic learning rate schedule.

3. Cyclical Learning Rate

When we train a DNN model, we want our model to converge to a good solution quickly and the most important hyper-parameter in this respect is the learning rate (LR). If the LR is too small, it will take a long time for the model to converge which is a challenge if we wish to update a model frequently as we acquire new data that can improve ASR performance. If the LR is too high, the model can fail to converge. Choosing the right LR requires skill, experience and a lot of experiments due to effects from regularisation, annealing and momentum [26, 27].

A simple LR schedule that offers a method to set the global LR is the cyclical learning rate proposed in [28]. This approach requires you to

1. Choose *stepsize* and *cycle length*
2. Find the upper and lower bound with the *LR range test*
3. Train for a number of cycles

A cycle goes from a low LR to a high LR and back to a low LR and usually has a triangular shape. A step is half a cycle and the step size is between 2-10 epochs so the LR reaches minimum and maximum values when an epoch is finished and we decrease or increase the LR over several epochs.

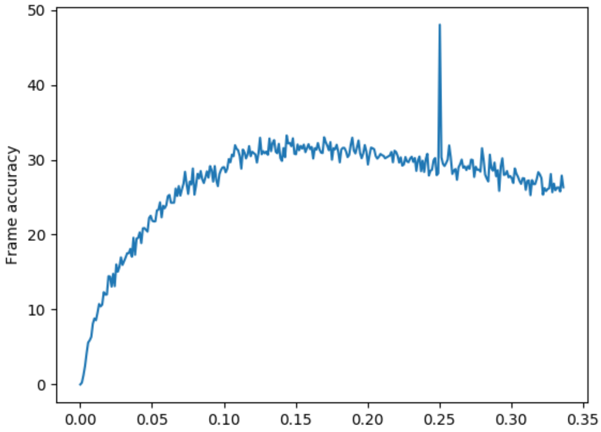

Figure 3: *LR range test on our internal Mandarin Chinese data set.*

To find the upper and lower bound on CLR for a new model or data set, we train the model for a single step from a very low bound to a very high bound e.g. 0.000001 to 1. We find a suitable upper bound just before accuracy stops increasing. In Figure 3, a suitable lower bound might be 0.00001 (or 0) and the upper bound 0.055 or 0.1.

We then train for a number of cycles and evaluate. [29] suggests that in some cases we may only need to train for a single cycle which will greatly speed up training and also corroborate the findings in [30] that high LRs regularise training and converge to wide minima that generalise better to unseen data than the sharp minima adaptive learning rates are prone to find.

CLR and the LR range test provides a principled way to find a good LR when we work with a new model or new data set so we do not need to run many trial and error experiments to find a good LR.

4. Experiments

4.1. Data

Monolingual DNNs were trained on Mandarin Chinese (ZH-CN) and multilingual DNNs on European French (FR-FR), Italian (IT), European Spanish (ES-ES) and American Spanish (ES-US). We use DNNs trained with our previous optimal LR schedules as baselines and compare to CLR. The domain of our data sets is mainly mobile voice search and Table 1 clearly indicates that we are not in a low-resource scenario. Tokens are words except for ZH-CN where tokens are symbols.

Table 1: *Training data statistics*

Model	Language	Hours	Types	Tokens
Monolingual	ZH-CN	379	5799	3365426
Multilingual	FR-FR	290	51915	1515568
	IT	345	56731	1422268
	ES-ES	400	58887	1822116
	ES-US	160	35281	1638681

4.2. CLR on Mandarin Chinese

We train a 5-layer FF-DNN with 1280 nodes in each layer followed by a linear transform and a softmax layer with 31,661 output nodes and initialise all affine layers with a diagonal matrix which gives a small constant accuracy boost in our experiments. We use ReLU activations, cross-entropy as loss function and standard SGD without momentum or regularisation. The input is 80-dimensional log filterbank coefficients plus energy stacked with ±8 frames of context (1377 dimensions in total). Energy-based voice activity detection has separated speech and non-speech frames before training.

The baseline trains with a LR of 0.01 for 6 epochs, 0.001 for 2 epochs and 0.0001 for several epochs, but no improvement was observed after 4 epochs. Like [28], we will denote this schedule as *piece-wise constant* LR (PC-LR). The upper and lower bound on CLR are 0.0001 and 0.055 and were found based on Figure 3. We estimate a trigram LM on the training transcripts to use in decoding.

Our training code uses an intermediary data partitioning called a *chunk* that is larger than a mini-batch and smaller than an epoch. We set the chunk size to 2002 utterances and define CLR per chunk rather than per mini-batch because the batch size is dynamically resized depending on the data. Between epochs, we randomly shuffle the order of chunks and shuffle samples within chunks.

Table 2: *ZH-CN validation accuracy and training time*

Model	Frame	Token	Train time	Epochs
PC-LR	35.08%	81.3%	40h	10
CLR4	34.86%	81.4%	**21.3h**	**4**
CLR8	34.81%	81.3%	40.3h	8

The frame accuracy curves in Figure 4 show that when the LR decreases, the training accuracy drops, but at the same time validation accuracy peaks. CLR peaks after 17 hours and after 36 hours which is faster than the baseline which achieves the highest validation accuracy after 40 hours. The frame and token accuracies in Table 2 show that we do not lose performance and we have added the LR range test to the training time for fair comparison (4.3h). On a single Pascal 1080 GPU, we can cut training time in half and CLR has the potential to significantly reduce the training time of our large scale DNN AM training.

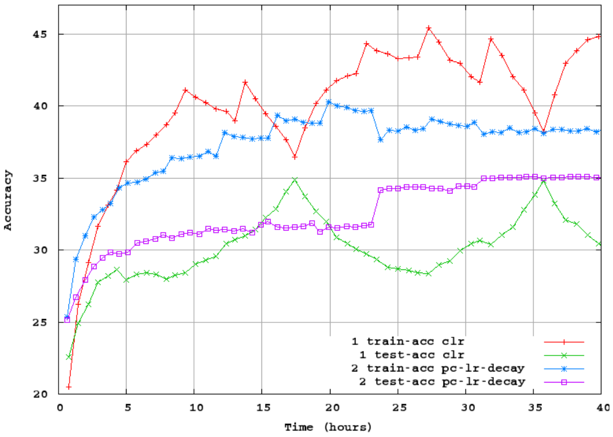

Figure 4: *Frame accuracy on the validation and training set using CLR and PC-LR as baseline.*

4.3. CLR on multilingual DNN

The multilingual DNN is a 7-layer ReLU FF-DNN, the chunk size is 4000 utterances, the input uses a context window of ±9 frames and the upper and lower bound on CLR are 0.0001 and 0.024 but otherwise identical to the monolingual ZH-CN AM. Our results in Table 3 confirm that we can reduce training time for large scale multilingual training with CLR with little loss of performance. We can train competitive multilingual AMs with CLR more than twice as fast as the baselines. On Spanish, the CLR-trained multilingual AM underperforms by 0.7% absolute compared to the previous best multilingual AM but all CLR models trained for 8 epochs outperform the monolingual baselines.

Table 3: *Comparison of DNNs in word accuracy*

DNN type	Epochs	FR-FR	IT	ES-ES
Monolingual	20+	77.7%	80.2%	80.5%
Multilingual	20+	82.9%	84.1%	82.0%
with CLR,	4	81.8%	83.1%	80.5%
cycle-len=4	8	82.6%	83.9%	81.3%
with CLR,	2	80.8%	81.7%	79.4%
cycle-len=2	4	81.6%	82.8%	80.5%
	8	82.1%	83.7%	81.2%

4.3.1. *Training time*

To further reduce training time, we cut the cycle length into 2 epochs. Figure 5 shows the frame accuracy curves using two different CLR schedules, one with 2 epoch cycle (green line)

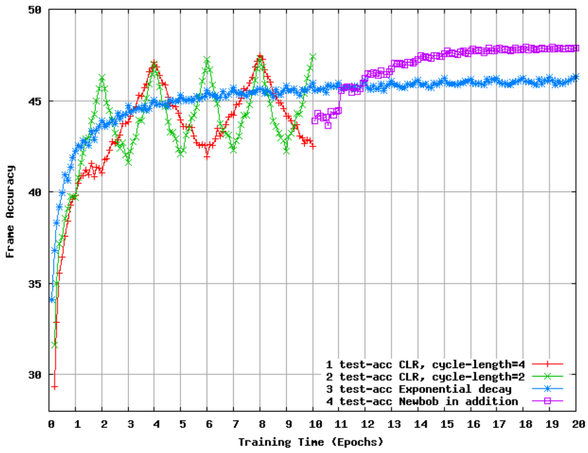

Figure 5: *Frame accuracy comparison on multilingual DNN training using 1) CLR and 2) exponential decay plus newbob*

and 4 epoch cycle (red line) as well as exponential decay learning rate then newbob schedule. All models obtained after 4 and 8 epochs in Table 3 have similar word accuracies, but the models obtained after 2 epochs may outperform the monolingual baseline for some languages.

4.3.2. *Performance impact*

In addition to the triangular learning rate schedule, we trained multilingual DNNs using the *triangular2* policy described in [28] which cut the maximum learning rate in half at the end of each cycle while maintaining the same minimum learning rate. As shown in Figure 6, DNN training with *triangular2* achieves better frame accuracy on validation than standard triangular in our experimental results. Table 4 also shows that we could achieve the same word accuracies with CLR while reducing training time from 20+ epochs to 12 epochs in Italian data.

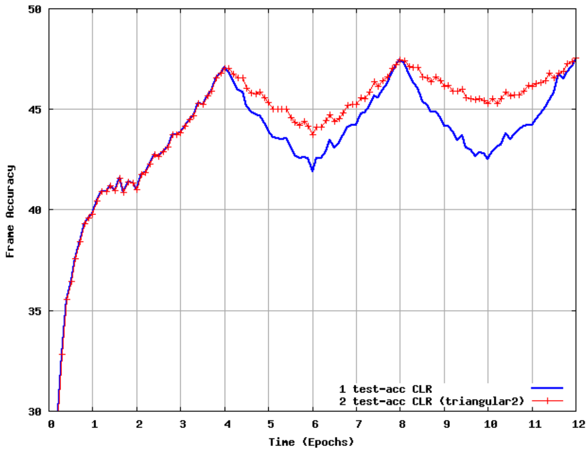

Figure 6: *Frame accuracy comparison on multilingual DNN training using 1) CLR and 2) triangular2*

5. Discussion

The techniques we use to create multilingual AMs have been used in speech research for a long time and the improvements in word accuracy make our approach a useful alternative to semi-/unsupervised training. An additional benefit is that our approach can be combined with semi-/unsupervised training and other data augmentation techniques.

Table 4: *Comparison of triangular and triangular2 on Italian word accuracy*

LR type	Number of Epochs		
	4	8	12
triangular	83.1%	83.9%	84.1%
triangular2	–	83.5%	84.2%

Our multilingual AM training relies on both data and linguistic knowledge. The linguistic knowledge is encoded in the SAMPA labels, the phonetic features, and the choice of data which we believe has a large influence on the success of our multilingual training because the cross-lingual phonetic difference is reduced compared to mixing data from different language families. We will continue this research and investigate the sensitivity of our approach to the acoustic input and the amount of linguistic information. If we can reduce the amount of linguistic knowledge or add data from different language families, the approach becomes more scalable. Auto-encoder techniques like those in [31] could reduce the phonetic difference and improve the cross-lingual tied-state clusters with less linguistic knowledge and diverse acoustic data. The next research direction will be to decode languages not used in training.

CLR effectively speeds up AM training with a large data set from various sources and training time could be even further reduced by using momentum optimisers which were not introduced in our work. We see no reason why similar time reductions should not be realised when training is distributed on multiple GPUs. The LR range test removes the need to set the learning rate and annealing schedule manually or by extensive search, but is currently an automation bottle-neck. The LR range test needs to be interpreted by a human and we are investigating an automatic way to determine the bounds on CLR and how the bounds may change when data are added in portions.

6. Conclusion

We have demonstrated that multilingual data from the same language family can improve word accuracy on languages in the training data with a novel approach that uses standard ASR techniques. We have also applied a simple cyclical learning rate schedule to FF-DNN AM training and achieved training time reduction with little or no loss of word accuracy in both monolingual and multilingual training.

This work suggests that CLR effectively speeds up DNN training while requiring minimal computation. It is easy to implement unlike other adaptive learning rate approaches, but provides significant improvement in DNN AM training. Though CLR finds minima with lower training accuracy at the end of a cycle, the solution generalises well for the validation data suggesting convergence towards near optimal results.

7. Acknowledgements

We would like to thank Andrej Ljolje, Vincent Goffin and Ryan Price for helpful comments on the development of the multilingual DNN training and CLR implementation.

8. References

[1] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art Speech Recognition with Sequence-to-Sequence Models," *CoRR*, vol. abs/1712.01769, 2017.

[2] K. Audhkhasi, B. Kingsbury, B. Ramabhadran, G. Saon, and M. Picheny, "Building competitive direct Acoustics-to-word Models for English Conversational Speech Recognition," *CoRR*, vol. abs/1712.03133, 2017.

[3] T. Schultz, M. Westphal, and A. Waibel, "The GlobalPhone Project: Multilingual LVCSR with JANUS-3," in *Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop*. Citeseer, 1997, pp. 20–27.

[4] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. Elsevier, 2006.

[5] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas, "Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models," *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4334–4337, 2010.

[6] N. T. Vu, F. Kraus, and T. Schultz, "Rapid Building of an ASR System for Under-Resourced Languages based on Multilingual Unsupervised Training," in *INTERSPEECH*, 2011.

[7] H. Lin, L. Deng, J. Droppo, D. Yu, and A. Acero, "Learning Methods in Multilingual Speech Recognition," 2008.

[8] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C.-H. Lee, "A study on Multilingual Acoustic Modeling for Large Vocabulary ASR," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4333–4336, 2009.

[9] M. Müller, S. Stüker, and A. H. Waibel, "Multilingual Adaptation of RNN Based ASR Systems," *CoRR*, vol. abs/1711.04569, 2017.

[10] A. Ragni, K. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for Low Resource Languages," in *INTERSPEECH*, 2014.

[11] V. Peddinti, G. Chen, D. Povey, and S. Khudanpur, "Reverberation robust Acoustic Modeling using i-Vectors with Time Delay Neural Networks," in *INTERSPEECH*, 2015.

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *INTERSPEECH*, 2015.

[13] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic Spectral Distortion for Low Resource Speech Recognition with Deep Neural Networks," *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 309–314, 2013.

[14] N. Jaitly and G. E. Hinton, "Vocal Tract Length Perturbation (VTLP) improves Speech Recognition," 2013.

[15] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Restarts," *CoRR*, vol. abs/1608.03983, 2016.

[16] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot Ensembles: Train 1, get M for free," *CoRR*, vol. abs/1704.00109, 2017.

[17] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: training Imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[19] J. C. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2010.

[20] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *CoRR*, vol. abs/1212.5701, 2012.

[21] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The Marginal Value of Adaptive Gradient Methods in Machine Learning," in *NIPS*, 2017.

[22] S. Stüker, M. Müller, Q. B. Nguyen, and A. H. Waibel, "Training Time Reduction and Performance Improvements from Multilingual Techniques on the BABEL ASR task," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6374–6378, 2014.

[23] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[24] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *CoRR*, vol. abs/1706.05098, 2017.

[25] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual Acoustic Models using Distributed Deep Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[26] B. T. Polyak, "Some Methods of speeding up the Convergence of Iteration Methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[27] Y. Nesterov, "A method of solving a Convex Programming Problem with convergence rate O (1/k2)," in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1983, pp. 372–376.

[28] L. N. Smith, "Cyclical Learning Rates for training Neural Networks," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, 2017.

[29] L. N. Smith and N. Topin, "Super-Convergence: Very fast training of Residual Networks using Large Learning Rates," *CoRR*, vol. abs/1708.07120, 2017.

[30] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. J. Storkey, "Three Factors Influencing Minima in SGD," *CoRR*, vol. abs/1711.04623, 2017.

[31] D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater, "A comparison of Neural Network Methods for Unsupervised Representation Learning on the Zero Resource Speech Challenge," in *INTERSPEECH*, 2015.