Multilingual Multi-Task Learning for Low-Resource Acoustic Modeling

Josh Meyer[1]

[1]University of Arizona

joshua.richard.meyer@gmail.com

Abstract

Index Terms: speech recognition, multi-task learning, acoustic modeling

1. Introduction

Previous work has shown that performance for a low-resource language on speech recognition can be improved by adding training data from another, resource-rich language. Typically, data from another language is added as a separate task in the Multi-Task Learning framework [1] via an additional output layer. The targets for this addititional language have typically been states of context-dependent triphones, defined by some tree clustering algorithm.

This current research builds off the intuition that triphones encode information which is too fine-grained to be maximally useful for language-transfer. Using a higher-level of linguistic abstraction (eg. the monophone), we are able to better extract the kind of language-general information useful in training an acoustic model for some target language.

The intuition being that when adding a source language as an auxialilary task, it would be better to focus on source-language distinctions which are robust and will transfer well to a new, target language.

Each auxiliary task is created by redefining the parameters of the HMM-GMM system used to bootstrap the DNN-hybrid system, such that the phonetic decision tree is cut short.

The target language is Kyrgyz, and the source language is English. Both data come from audiobooks, English being from LibriSpeech and Kyrgyz from the Bizdin.kg project.

2. Background

The earliest examples of MTL with multiple languages can be found in [2] and [3], who both used triphones from each language as additional tasks. They were interested in improving performance on all languages, not just one target language. These two studies were then followed up in multiple other threads of research. More recently, [4] found that adding more triphones from a single, well-resourced language (English) actually leads to better performance, but this could be conflated with the fact that they did not use any weighting scheme, and as such, the source language with fewer states and more data would more quickly overfit, leaving less of a chance for the target language to exert influence during backprop.

In another direction, multiple works have investigated MTL for a single language, without any source language transfer. These approaches aim to find tasks which are phonetically relevant to the main task. Both [5] and later [6] looked at a very similar approach to what I explore here, with MTL on broader, more abstract phonetic categories for English. They both found improvement on TIMIT, but they didn't investigate multi-lingual transfer. With regards to low-resource languages, [7] and later [8] similarly looked at MTL for a single target language, using graphemes or a universal phoneset as extra targets.

3. DNN-Hybrid Training as Model Transfer

The standard DNN-Hybrid approach requires the GMM-HMM system to provide the labels for supervised training. This reliance of the DNN on GMM alignments is actually a form of model transfer, where the DNN is trained to perform the extact same classification as its GMM predecessor. The DNN not only learns the frame alignments from the individual GMMs, but also the DNN indirectly learns the structure of the phonetic decision tree used to define the tied-state system. This is because the output layer of the DNN is trained to predict targets which were defined via leaves of the decision tree.

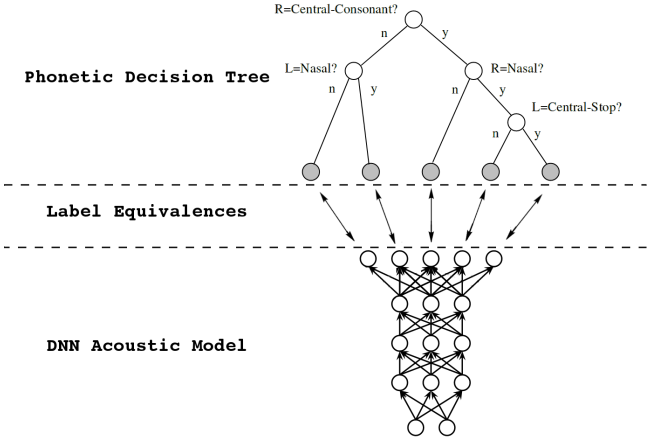

Figure 1: *GMM→DNN Model Transfer*

Given that standard triphones encode very fine-grained information which may not help performance on a target language, the following experiments investigate GMM→DNN model transfer at a higher level for an additional source language.

4. Experiments

This work investigates the application of MTL technique to low resource acoustic modeling. All experiments simulate a common development scenario: there exists little transcribed data for the target language, but lots of data in some source language.

The following experiments tease out (1) the level of detail at which the source language should be modeled and (2) the amount of weighting which should be given to the target language training examples.

The first point of interest is the level of detail at which the source language is modeled. This is investigated via addition of

multiple tasks to the TDNN during training. The experiements here are crafted to answer the question: *how much phonetic detail should the source language be modeled at to best transfer data to the target language?*.

We can model the source language with lots of contextual detail (ie. the triphone), with abstracted, context-independent detail (ie. the monophone), or somewhere in between. Building of the traditional hybrid DNN-HMM ASR training pipeline, investigating these levels of representation are easily acheived via the phonetic decision tree. We can merely assign labels to training data frames by moving up the tree, from leaves (triphones) to roots (monophones).

The second point of interest is the relative weighting of target vs. source language training data. It is clear that if we train two languages in parallel, the source language (with many more training samples) will dominate the target language in the fight for influence during backprop.

To my knowledge this has not been investigated in ASR acoustic modeling, although it was dedicated its own chapter in Caruana's 1997 dissertation. To investigate weighting further, I examine the following target vs. source weighting schemes: 1-to-2, 1-to-1, and 2-to-1 (all ratios are target-to-source).

These weights are instantiated during training via a weight to the target output label, where the label is a one-hot vector. For example, given 1000 hours of source language and 1 hour of target language, to acheive a 1-to-1 ratio in training, I would multiply the target labels from the target language by 1000, resulting in target vectors such as $[0, 0, 0, 0, 1000, 0, 0, \ldots]$ instead of $[0, 0, 0, 0, 1, 0, 0, \ldots]$. The final layer is a ReLU activation, so having a target value higher than 1 is not an issue as it would be with a traditional softmax layer.

4.1. Data

Two speech corpora are used in the following experiments:

1. $\approx$ 5 hours of LibriSpeech (4.86 hours)
2. $\approx$ 1.5 hours of Kyrgyz audiobook (1.59 hours)

4.2. Model Building

All models were build using the Kaldi `nnet3` approach. The main neural net run script used in this paper can be found at www.github.com/JRMeyer/kaldi-mirror/egs/kgz/kyrgyz-model/run_nnet3_multilingual.sh. The main GMM script used to create data alignments can be found at www.github.com/JRMeyer/kaldi-mirror/egs/kgz/kyrgyz-model/run_gmm.sh.

4.2.1. Decision Trees

In GMM training, monophones (for each language) were allotted 1,000 Gaussian components, and trained over 25 iterations of EM. These monophones were then expanded into context-dependent triphones via a phonetic decision tree, with a maximum of 2,000 leaves & 5,000 Gaussians (LibriSpeech reached 1584 leaves, and Kyrgyz reached 752). The resulting tied-state clusters (ie. leaves) are then trained as context-dedendent triphones over 25 iterations of EM.

4.2.2. Multi-Task Neural Net Acoustic Models

Given the alignments from the GMM-HMM models, a 5-layer, 500-dimensional TDNN is trained over 10 epochs of backprop on a single GPU instance.

Each auxiliary task is implemented as a separate output layer along with a separate, penultimate hidden layer. All other hidden layers of the TDNN are trained in parallel.

During testing, *only* the main task is used. The additional tasks are dropped and the baseline Kyrgyz triphones are used in decoding. This highlights the purpose of the extra tasks: to force the learning of robust representations in the hidden layers during training. The tasks may in fact not be the best option for final classification; they serve as "training wheels" which are then removed once the net is ready.

Baseline Model

All the following architectures will be compared to the performance of a baseline model of identical architecture (5 hidden layers, 500-dimentional layers, ReLU activations, same training algorithm). The output targets are standard context-dependent triphones trained on Kyrgyz audio.

To account for any advantage mutliple output layers may bring about, the baseline contains two output layers, where the tasks are identical. In this way, random initializations in the weights and biases for each task are accounted for.

Auxiliary Tasks

The auxiliary tasks all train on English language data from the LibriSpeech corpus. Investigating the intuition that labels generated by a standard triphone phonetic decision tree are not the best representation of data for transfer learning, the auxiliary tasks here investigate different levels in the decision tree's branches.

I split the LibriSpeech phonetic decision tree into three logical parts:

1. roots (standard monophones)
2. branches (what I dub, "half"-phones)
3. leaves (standard triphones)

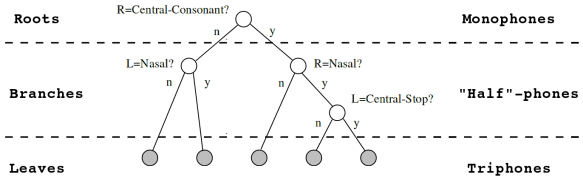

Figure 2: *Logical Tree Parts*

The "half"-phones were created by halving the optimal number of leaves from the triphone system (ie. 1584 leaves) and re-training a new GMM-HMM system with half the optimal number of leaves (1/2 * 1584 = 792 leaves). All the other parameters were left unchanged (number of Gaussian components, iterations of EM, etc.).

Table 1: *Auxiliary Tasks*

Logical Tree Part	Level of Phonetic Detail	№ of Tasks
Roots	Monophones	1
Branches	Half-phones	1
Leaves	Triphones	1
Lower Tree	Monophones + Half-phones	2
Upper Tree	Half-phones + Triphones	2
Whole Tree	Monophones + Half-phones + Triphones	3

Each of the above tasks were trained on the 5 hour section of LibriSpeech corpus. They are included as an extra output layer in the TDNN.

By forcing the neural net to recognize higher levels in the tree, we will learn representations which are more abstract, and therefore more likely to be relevant multi-lingually.

The addition of each above task adds approximately 5 hours of training data to the standard training of a Single Task Model on Kyrgyz. As such, a weighting procedure was used to balance the relative influence of source vs. target training data on back-prop. For example, to reach a one-to-one ratio, where one hour of Kyrgyz is equal to one hour of English, I multiplied every Kyrgyz target one-hot vector by 3.06.

Table 2: *Target:Source Data Weighting Scheme*

Target:Source Ratio	**Target Weight**
1:2	1.53x
1:1	3.06x
2:1	6.12x

4.3. Results

All results are performed on the same held-out section of Kyrgyz audiobook. The bigram language model, lexicon, and main-task decision tree are built into a standard decoding graph in the traditional Kaldi style. Decoding is performed with a bigram backoff language model trained on a Wikipedia Kyrgyz dump, and contains, 103,998 unigrams and 56,6871 bigrams.

1. Any amount of English beats out Kyrgyz-only baseline.
2. Triphones work better than monophones (except for the 2-to-1 weighting).
3. Both languages / tasks overfit (referencing frame-classification logs)
4. atai overfit slower with additional tasks
5. atai overfits with monophones earlier, I think because it's an easier task
6. only train on 30 minutes of Krygyz... more likely to find an effect

The below table shows Word Error Rates relative to the baseline, single-task model. They all show improvement over the baseline. However, the result that triphones work best is not new, and my "half"-phones don't beat them out strictly speaking.

I do see a trend with the +1 extra tasks: the more I increase the ratio of target-to-source data, the better abstract tasks perform. That is, the more data I have in the target language, the better it is to model the source language higher in the tree.

What's disappointing is that none of the +2 or +3 tasks outperform the more simple, +1 tasks. I would have hoped that modeling the whole tree at all levels would perform best, but that isn't the case.

The two last experiments are not done running yet, but I think they'll be worse than the other's in their column.

Table 3: *Word Error Rates (WER%) Relative to Baseline*

	Target:Source Weighting		
Auxiliary (Source Lang) Tasks	*1-to-2*	*1-to-1*	*2-to-1*
STL Baseline	50.54% WER		
Monophones	-3.13	-3.22	**-2.34**
Halfphones	-1.86	**-3.81**	-1.86
Triphones	**-3.81**	-3.42	-1.17
Monophones + Halfphones	-2.44	-2.05	-2.34
Halfphones + Triphones	-2.64	-2.54	
Monophones + Halfphones + Halfphones	-1.95	-2.34	
AVERAGE	-2.64	-2.90	

5. Discussion

6. Conclusions

7. Acknowledgements

8. References

[1] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[2] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[3] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8619–8623.

[4] F. Grézl and M. Karafiát, "Boosting performance on low-resource languages by standard corpora: An analysis," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 629–636.

[5] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6965–6969.

[6] Z. Huang, J. Li, S. M. Siniscalchi, I.-F. Chen, J. Wu, and C.-H. Lee, "Rapid adaptation for deep neural networks through multi-task learning," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] D. Chen, B. Mak, C.-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 *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5592–5596.

[8] D. Chen and B. K.-W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 7, pp. 1172–1183, Jul. 2015. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2015.2422573