# MULTILINGUAL TRAINING OF DEEP NEURAL NETWORKS

*Arnab Ghoshal, Pawel Swietojanski, Steve Renals*

Centre for Speech Technology Research, University of Edinburgh, UK
{a.ghoshal,p.swietojanski,s.renals}@ed.ac.uk

## ABSTRACT

We investigate multilingual modeling in the context of a deep neural network (DNN) – hidden Markov model (HMM) hybrid, where the DNN outputs are used as the HMM state likelihoods. By viewing neural networks as a cascade of feature extractors followed by a logistic regression classifier, we hypothesise that the hidden layers, which act as feature extractors, will be transferable between languages. As a corollary, we propose that training the hidden layers on multiple languages makes them more suitable for such cross-lingual transfer. We experimentally confirm these hypotheses on the GlobalPhone corpus using seven languages from three different language families: Germanic, Romance, and Slavic. The experiments demonstrate substantial improvements over a monolingual DNN-HMM hybrid baseline, and hint at avenues of further exploration.

***Index Terms*—** Speech recognition, deep learning, neural networks, multilingual modeling

## 1. INTRODUCTION

Recent years have seen a renewed interest in developing speech technologies for a broad range of languages and domains, encouraged by commercial and other forces and enabled by newer techniques and improved understanding. This naturally leads to techniques that attempt to transfer knowledge between languages, since the major obstacle for developing speech technology in a new language is the lack of linguistic resources in form of (manually) transcribed audio, pronunciation lexicon, and in-domain text for language modeling. While the constraints on resources is a very practical reason to look at multilingual techniques[1], a more pragmatic motivation is that by engineering such techniques we may improve our understanding of the algorithms used as well as the commonalities between different languages.

In this work we focus on multilingual acoustic modeling, and assume that we have access to a pronunciation dictionary and language model. Previous approaches in this direction have included: constructing a universal phone set [1, 2, 3]; modeling a set of universal speech attributes, such as voicing, nasality and frication [4]; and mapping between phones of different languages using some automated method that relies on some predefined distance measure [5, 6, 7] or by manually creating a mapping using acoustic phonetic knowledge [5].

The subspace Gaussian mixture model (SGMM) [8] has been shown to be suitable for cross-lingual modeling without requiring an explicit mapping between phone units in different languages [9, 10]. In an SGMM, the emission densities of a hidden Markov model (HMM) are modeled as mixtures of Gaussians, whose parameters are factorized into a globally-shared set that does not depend on the HMM states, and a state-specific set. The global parameters may be thought of as a model for the overall acoustic space, while the state-specific parameters provide the correspondence between different regions of the acoustic space and individual speech sounds. It is the decoupling of these two aspects of modeling speech data that makes the SGMM global parameters, which do not directly depend on the phone units, suitable for sharing between different languages. This line of argument naturally implies that training the SGMM global parameters using multiple languages makes them more suitable for transferring to a new language, since multilingually-trained subspace parameters will necessarily provide better coverage of the acoustic space. This is, in fact, what is observed in practice [10].

Deep neural networks (DNNs) also provide a decoupling of intermediate representations of data and the correspondence between the representations and the categories of interest, albeit one that is very different from the decoupling in SGMMs. The hidden layers of a DNN act as a cascade of feature extractors [11, 12] and only the output layer provides the direct correspondence to the classes of interest. We hypothesise that the hidden layers of a DNN may be transferable between languages, in an analogous way to the SGMM global parameters. This hypothesis is further supported by our earlier work [13], in which we showed that layerwise pretraining of deep networks using stacked restricted Boltzmann machines (RBMs) [14] is not sensitive to the choice of language.

In this work, we start with a network that has been fine-tuned on one language, and whose output layer corresponds to tied context-dependent phone states in that language. We then replace the output layer with that corresponding to another language, borrowing the rest of the (hidden) layers, and

[1] Here we do not use the term "multilingual" in the sense of code-switching, but purely for techniques that can learn from multiple languages.

finetune the whole network on the new language. This process is repeated for several different languages. The networks are used in a hybrid DNN-HMM setup, in which the DNN outputs are used to provide scaled likelihood estimates for the states of an HMM [15]. We see consistent gains in recognition accuracy from training the hidden layers using multiple languages.

Previous uses of neural networks in cross-lingual acoustic modeling have mainly focused on tandem approaches [16], which use neural network outputs as discriminative features for a GMM-HMM acoustic model. In such approaches, networks trained on a source language are used to provide features for a target language [17, 18]. These features typically improve on a competitive target language baseline when the amount of transcribed audio in the target language is small [19, 20].

Of these, the method of Thomas, et al. [20] is the closest in spirit to our current work. That paper proposes a multilingual tandem system where a 3-layer network with a narrow hidden layer — the bottleneck layer — is multilingually trained by using a different output layer (corresponding to context-independent phones) for each language. Two separate networks, one using spectral (PLP) features and the other using modulation (FDLP) features, are trained and their combination is shown to improve over a baseline system trained on 1 hour of speech using PLP features. In the current work, we use deeper networks with 7 layers that do not contain a bottleneck layer; the network outputs correspond to tied triphone states; the networks are trained on the same (MFCC) features as the standard acoustic model; and we show improvements over a monolingual hybrid DNN-HMM system.

## 2. DNNS FOR SPEECH RECOGNITION

We use the convention that a layer in a feedforward neural network correspond to a matrix of connection weights between two sets of neurons. We denote the input to the $l$-th layer by $\mathbf{u}_{l-1}$, with $\mathbf{u}_0 = \mathbf{o}_t$, the acoustic observation at time $t$. The output of the $l$-th layer, $\mathbf{u}_l$ is obtained as:

$$\mathbf{u}_l = \sigma(\mathbf{W}_l \mathbf{u}_{l-1} + \mathbf{b}_l), \qquad \text{for } 1 \leq l < L,$$

where $\mathbf{W}_l$ is the weight matrix and $\mathbf{b}_l$ is the additive bias vector at the $l$-th layer; $\sigma(x) = 1/(1+\exp(-x))$ is a sigmoid non-linearity, also known as the activation function. The $L$-th layer, also called the output layer, uses a softmax function to obtain the posterior probability $P_\theta(y|\mathbf{o}_t)$ of each tied triphone state $y$ given the acoustic observation $\mathbf{o}_t$ at time $t$:

$$P_\theta(y|\mathbf{o}_t) = \frac{\exp(\mathbf{w}_{Ly}^\top \mathbf{u}_{L-1} + b_{Ly})}{\sum_{\tilde{y}} \exp(\mathbf{w}_{L\tilde{y}}^\top \mathbf{u}_{L-1} + b_{L\tilde{y}})},$$

where $\mathbf{w}_{Ly}^\top$ is the $y$-th row of the weight matrix $\mathbf{W}_L$. To obtain scaled likelihoods, the posterior probability estimates produced by the network are divided by the prior probabilities of the states $y$ [15].

We use stochastic gradient descent to train DNNs, minimizing a negative log posterior probability cost function over the set of training examples $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$:

$$\theta^* = \arg\min_\theta - \sum_{t=1}^{T} \log P_\theta(y_t|\mathbf{o}_t),$$

where $\theta = \{\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L\}$ is the set of parameters of the network, and $y_t$ is the most likely state at time $t$ obtained by a forced alignment of the acoustics with the transcript.

While the basic idea was used in speech recognition in the early 1990s [15], earlier uses of hybrid systems were mainly limited to estimating scaled likelihoods for monophone states using feedforward network with two layers [21] and recurrent networks [22], owing to computational constraints. More recently DNNs with up to 9 layers have been used with outputs corresponding to both monophone states [23] and tied context-dependent states [24]. While training deep networks directly results in a difficult optimization problem, an unsupervised pretraining phase using greedy layer-wise training of restricted Boltzmann machines [14] has been shown to give good results. Following our earlier work [13], we use RBM-based pretraining to initialize the DNN models.

## 3. MULTILINGUAL DNN TRAINING

One may view multilayered neural networks as a cascaded sequence of feature extractors followed by a logistic regression classifier at the output layer. From this perspective, it is reasonable to argue that the hidden layer feature extractors ought to be transferable across domains and languages. It is this simple idea that motivates the exploration in this paper.

A schematic for the training procedure of the multilingual DNNs is shown in Figure 1. We initialize our networks with stacked RBMs that are pretrained on a single language. We found in our earlier work [13] that the choice of the language used for RBM-based pretraining did not make a significant difference to the final result. In fact, all experiments reported in this paper use DNNs that are initialized from stacked RBMs trained on Polish. Starting from this, a softmax layer, with randomly initialized weights and classes corresponding to the tied triphone states in a given language, is added as the final layer and the whole structure is finetuned on the chosen language. Afterwards, the softmax layer is replaced by one corresponding to a different language, with randomly initialized weights, and finetuning is done for the next language. This process is repeated for multiple languages.

A potential problem with the protocol followed here is that this form of language-sequential training may lead to more biased estimates. We see some indication of that in our results. An alternative would be to train all the languages simultaneously. Language-sequential training, nevertheless, has a practical advantage: training a model for a new language

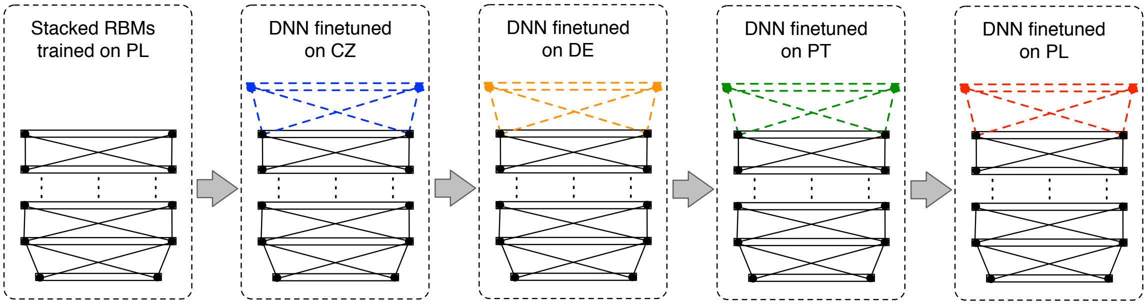

**Fig. 1**. Multilingual training of deep neural networks.

does not require retraining any previously trained models for other languages. Ideally, one would like the hidden layers to converge to an optimized set of feature extractors that can be reused across domains and languages. However, such a study is inherently empirical, and variations of the techniques reported here are currently under investigation.

## 4. EXPERIMENTS

We used the GlobalPhone corpus [25] for our experiments. The corpus consists of recordings of speakers reading newspapers in their native language. There are 19 languages from a variety of geographical locations: Asia (Chinese, Japanese, Korean), Middle East (Arabic, Turkish), Africa (Hausa), Europe (French, German, Polish), and Americas (Costa Rican Spanish, Brazilian Portuguese). Recordings are made under relatively quiet conditions using close-talking microphones; however acoustic conditions may vary within a language and between languages.

In this work we use seven languages from three different language families: Germanic, Romance, and Slavic. The languages used are: Czech, French, German, Polish, Brazilian Portuguese, Russian and Costa Rican Spanish. Each language has roughly 20 hours of speech for training and two hours each for development and evaluation sets, from a total of about 100 speakers. The detailed statistics for each of the languages is shown in Table 1.

**Table 1**. Statistics of the subset of GlobalPhone languages used in this work: the amounts of speech data for training, development, and evaluation sets are in hours.

| Language | #Phones | #Spkrs | Train | Dev | Eval |
|---|---|---|---|---|---|
| Czech (CZ) | 41 | 102 | 26.8 | 2.4 | 2.7 |
| French (FR) | 38 | 100 | 22.8 | 2.1 | 2.0 |
| German (DE) | 41 | 77 | 14.9 | 2.0 | 1.5 |
| Polish (PL) | 36 | 99 | 19.4 | 2.9 | 2.3 |
| Portuguese (PT) | 45 | 101 | 22.8 | 1.6 | 1.8 |
| Russian (RU) | 48 | 115 | 19.8 | 2.5 | 2.4 |
| Spanish (SP) | 40 | 100 | 17.6 | 2.0 | 1.7 |

### 4.1. Baseline systems

For each language, we built standard maximum-likelihood (ML) trained GMM-HMM systems, using 39-dimensional MFCC features (C0-C12, with delta and acceleration coefficients), using the Kaldi speech recognition toolkit [26]. The number of context-dependent triphone states for each language is 3100 with a total of 50K Gaussians (an average of roughly 16 Gaussians per state). The development set word error rates (WER) for the different languages are presented in Table 2. The results reported here are better than those in our earlier work [13] because we used better LMs obtained from the authors of [3, 27]. We must stress that the ML baseline results are presented here to serve as a point of reference, and not for direct comparison with the DNN results. The scripts needed to replicate the GMM-HMM results are publicly available as a part of the Kaldi toolkit[2].

### 4.2. DNN configuration and results

For training DNNs, our tools utilize the Theano library [28], which supports transparent computation using both CPUs and GPUs. We train the networks on the same 39-dimensional MFCCs as the GMM-HMM baseline. The features are globally normalised to zero mean and unit variance, and 9 frames (4 on each side of the current frame) are used as the input to the networks. All the networks used here are 7 layers deep, with 2000 neurons per hidden layer. The initial weights for the softmax layer were chosen uniformly at random: $w \sim U[-r, r]$, where $r = 4\sqrt{6/(n_{l-1} + n_l)}$ and $n_l$ is the number of units in layer $l$. Fine-tuning is done using stochastic gradient descent on 256-frame mini-batches and an exponentially decaying schedule, learning at a fixed rate (0.08) until improvement in accuracy on cross-validation set between two successive epochs falls below 0.5%. The learning rate is then halved at each epoch until the overall accuracy fails to increase by 0.5% or more, at which point the algorithm terminates. While learning, the gradients were smoothed with

[2] Available from: `http://kaldi.sf.net`

**Table 2**. Development set results: vocabulary size is the intersection between LM and pronunciation dictionary vocabularies; perplexity (PPL) figures are obtained considering sentence beginning and ending markers; and for multilingual DNNs we show the order of the languages used to train the networks.

| Language | Vocab | PPL | ML-GMM | DNN | Multilingual DNN | |
|---|---|---|---|---|---|---|
| | | | WER(%) | WER(%) | Languages | WER(%) |
| CZ | 29K | 823 | 18.5 | 15.8 | — | — |
| DE | 36K | 115 | 13.9 | 11.2 | CZ →DE | 9.4 |
| FR | 16K | 341 | 25.8 | 22.6 | CZ →DE →FR | 22.6 |
| SP | 17K | 134 | 26.3 | 22.3 | CZ →DE →FR →SP | 21.2 |
| PT | 52K | 184 | 24.1 | 19.1 | CZ →DE →FR →SP →PT | 18.9 |
| RU | 24K | 634 | 32.5 | 27.5 | CZ →DE →FR →SP →PT →RU | 26.3 |
| PL | 29K | 705 | 20.0 | 17.4 | CZ →DE →FR →SP →PT →RU →PL | 15.9 |

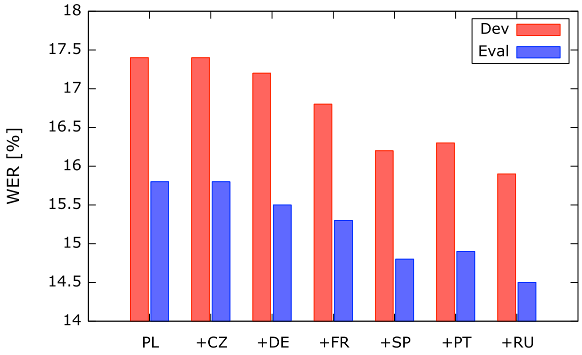

**Fig. 2**. Mono- and multi-lingual DNN results on Polish. The languages are added left-to-right starting with Czech and ending with Polish. Hence '+FR' corresponds to the schedule CZ →DE →FR →PL.

a first-order low-pass momentum (0.5). For the multilingual DNNs, an initial learning rate of 0.04 is used.

A comparison of the WERs obtained by the monolingual and multilingual DNNs for the different languages in Table 2 supports our hypotheses: the hidden layers are indeed transferable between languages, and training them with more languages, by and large, makes them better suited for the target languages. These trends are shown in greater detail for Polish (in Figure 2) and Russian (in Table 3).

It is important to note that the different systems do not control for the amount of data; a system with more languages is trained on more data and some of the performance gains may well be attributed to that. However, we also notice that just adding more data may not always improve results. For example, in Figure 2 we see worse performance by adding Portuguese, and the Czech data did not lower WER for either Polish or Russian. This may indicate a need for better cross-corpus normalization, for example, using speaker adaptive training. Conversely, this may also indicate that the sequential training protocol followed here is suboptimal. In fact, for the systems shown in Figure 2, training on Russian after Spanish and then on Polish leads to similar WER as when Portuguese is used for finetuning after Spanish. These issues are currently under investigation.

**Table 3**. Mono- and multi-lingual DNN results on Russian.

| Languages | Dev | Eval |
|---|---|---|
| RU | 27.5 | 24.3 |
| CZ →RU | 27.5 | 24.6 |
| CZ →DE →FR →SP →RU | 26.6 | 23.8 |
| CZ →DE →FR →SP →PT →RU | 26.3 | 23.6 |

## 5. DISCUSSION

We presented experiments with multilingual training of hybrid DNN-HMM systems showing that training the hidden layers using data from multiple languages leads to improved recognition accuracy. The results are very promising and point to areas of future work: for instance, determining if the number of layers in the network has an effect on these results. The notion of deep neural networks performing a cascade of feature extraction, from lower-level to higher-level features, provides both an explanation for the observed effect, as well as the inkling that the effect may be more pronounced for deeper structures. There are also practical engineering issues to consider: checking whether a simultaneous training, where the randomization of observations is done across all languages in consideration, improves on the current sequential protocol; experimenting with transformations of the feature space as well as with discriminative features, some of which may enhance or mitigate this effect; and experimenting with a broader set of languages.

## 6. ACKNOWLEDGMENTS

This research was supported by EPSRC Programme Grant grant, no. EP/I031022/1 (Natural Speech Technology). We would also like to thank Tanja Schultz and Ngoc Thang Vu for making the GlobalPhone language models available to us, and Miloš Janda for help with the baseline systems.

## 7. REFERENCES

[1] T Schultz and A Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Proc. Eurospeech*, 1997, pp. 371–374.

[2] T Schultz and A Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, 1998.

[3] T Schultz and A Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–52, 2001.

[4] SM Siniscalchi, DC Lyu, T Svendsen, and CH Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 3, pp. 875–887, 2012.

[5] W Byrne, P Beyerlein, JM Huerta, S Khudanpur, B Marthi, J Morgan, N Peterek, J Picone, D Vergyri, and W Wang, "Towards language independent acoustic modeling," in *Proc. ICASSP*. IEEE, 2000, pp. 1029–1032.

[6] KC Sim and H Li, "Robust phone set mapping using decision tree clustering for cross-lingual phone recognition," in *Proc. ICASSP*. IEEE, 2008, pp. 4309–4312.

[7] KC Sim, "Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition," in *Proc. ASRU*. IEEE, 2009, pp. 546–551.

[8] D Povey, L Burget, M Agarwal, P Akyazi, F Kai, A Ghoshal, O Glembek, N Goel, M Karafiát, A Rastrow, RC Rose, P Schwarz, and S Thomas, "The subspace Gaussian mixture model—A structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[9] L Burget, P Schwarz, M Agarwal, P Akyazi, K Feng, A Ghoshal, O Glembek, N Goel, M Karafiát, D Povey, A Rastrow, R Rose, and S Thomas, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proc. IEEE ICASSP*, 2010, pp. 4334–4337.

[10] L Lu, A Ghoshal, and S Renals, "Regularized subspace Gaussian mixture models for cross-lingual speech recognition," in *Proc. IEEE ASRU*, 2011.

[11] Y Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, January 2009.

[12] Y Bengio, A Courville, and P Vincent, "Representation learning: A review and new perspectives," `arXiv:1206.5538`.

[13] P Swietojanski, A Ghoshal, and S Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc IEEE SLT*, 2012.

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

[15] H Bourlard and N Morgan, *Connectionist Speech Recognition—A Hybrid Approach*, Kluwer Academic, 1994.

[16] H Hermansky, DPW Ellis, and S Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, 2000.

[17] A Stolcke, F Grézl, M-Y Hwang, X Lei, N Morgan, and D Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. IEEE ICASSP*, 2006.

[18] F Grézl, M Karafiát, and M Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. IEEE ASRU*, 2011.

[19] S Thomas and H Hermansky, "Cross-lingual and multistream posterior features for low resource LVCSR systems," in *Proc. INTERSPEECH*, 2010.

[20] S Thomas, S Ganapathy, and H Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. IEEE ICASSP*, 2012.

[21] S Renals, N Morgan, H Bourlard, M Cohen, and H Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 161–174, 1994.

[22] AJ Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 298–305, 1994.

[23] A Mohamed, GE Dahl, and G Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 14–22, 2012.

[24] GE Dahl, D Yu, L Deng, and A Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[25] T Schultz, "GlobalPhone: a multilingual speech and text database developed at Karlsruhe University," in *Proc. ICLSP*, 2002, pp. 345–348.

[26] 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 *Proc. IEEE ASRU*, December 2011.

[27] N T Vu, T Schlippe, F Kraus, and T Schultz, "Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit," in *Proc. INTERSPEECH*, 2010.

[28] J Bergstra, O Breuleux, F Bastien, P Lamblin, R Pascanu, G Desjardins, J Turian, D Warde-Farley, and Y Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. SciPy*, 2010.